



**Rapport de stage de 4.A :**  
**Présenté par :**  
**ABOUABDALLAH Mohamed Anwar**  
**4A. Mathématiques appliquées et Modélisation 2017-2018**

**Estimation des paramètres de diffusion d'une population dans un paysage hétérogène à partir d'équations aux dérivées partielles et de données génétiques spatialisées :**

Lieu de stage



Institut Nationale de recherche en Agronomie

Encadrement

Tuteurs : Pierre Franck : Unité « Plantes et Système de Culture Horticole »,  
Olivier Bonnefon : Unité « Biostatistique et Processus Spatiaux »

**INRA Centre de recherche PACA**  
**228 route de l'Aérodrome Domaine Saint Paul - Site Agroparc CS 40509**  
**84914 Avignon Cedex 9**  
**Tél. : 04 32 72 20 00 - Fax : 04 32 72 20 42**

## Remerciements :

**M**on stage s'inscrit dans le domaine des mathématiques appliquées pour l'écologie, un sous domaine des mathématiques que je n'ai découvert que très récemment. Ceci a fait que plusieurs membres de laboratoire de recherche m'ont aidé et permis de me familiariser avec le sujet et de réussir mon stage. Tout d'abord, je souhaite remercier tout particulièrement mes maîtres de stage, Olivier Bonnefon et Pierre Franck, pour leur disponibilité, patience et ainsi que leurs explications. Je les remercie également pour la précieuse expérience qu'ils m'ont permis d'acquérir.

Je tiens aussi à remercier Mr. Étienne Klein, directeur de l'Unité BioSP de l'I.N.R.A. d'Avignon de m'avoir accueilli au sein de son Unité et pour m'avoir ouvert l'une des portes du monde de la recherche dans un secteur aussi riche en perspectives d'avenir.

Je remercie aussi mon cobureau Soufiane kharbach, pour son accueil et ses clarifications sur les démarches statistiques, : Julien Papaïx, mon voisin du bureau, pour les différentes explications concernant mon sujet. Enfin je tiens à remercier tous les membres de mon unité pour leur accueil.

### **Membres de l'unité BIOSP (Journée Hors des murs) :**



## Table des matières

<b>1 Présentation du stage :</b>	<b>3</b>
1.1 Tableau des notations : . . . . .	3
1.2 Introduction : . . . . .	4
1.2.1 Présentation de l'entreprise : . . . . .	4
1.2.2 Projet de stage : . . . . .	5
1.2.3 Intérêts du stage : . . . . .	6
1.2.4 Objectives du stage : . . . . .	7
<b>2 Approche théorique :</b>	<b>8</b>
2.1 Quelques bases de biologies : . . . . .	8
2.1.1 Lois de Mendel : . . . . .	8
2.1.2 L'équilibre de Hardy-Weinberg : . . . . .	8
2.2 Les données génétiques : . . . . .	9
2.3 Partie statistique : . . . . .	9
2.3.1 Le modèle de capture : . . . . .	10
2.3.2 F-Modèle [9] : . . . . .	10
2.3.3 Le modèle génétique et calcul des Probabilités : . . . . .	11
2.3.4 Calcul de la vraisemblance : . . . . .	11
2.4 Partie numérique : . . . . .	12
2.4.1 Modélisation par des E.D.P. : . . . . .	12
2.4.2 Méthodes des éléments finis : . . . . .	12
2.4.3 Présentation de la méthode Quasi-Newton et utilisation : . . . . .	13
2.5 Présentation de la démarche : . . . . .	14
2.5.1 Partie I : Calcul du coefficient de diffusion à partir des fréquences corrigées . . . . .	14
2.5.2 Partie II :Calcul des coefficients de diffusion à partir des fréquences simulées : . . . . .	15
<b>3 Travail effectué :</b>	<b>16</b>
3.1 Outils utilisés : . . . . .	16
3.1.1 Pré-calculs et méthodes statistiques : . . . . .	16
3.1.2 Optimisation : . . . . .	16
3.2 Partie I : Calcul du coefficient de diffusion à partir des fréquences corrigées : . . . . .	17
3.3 Partie II : Calcul des coefficients de diffusion à partir des fréquences simulées : . . . . .	20
<b>4 Résultats obtenus :</b>	<b>22</b>
4.1 Partie I : Calcul des paramètres de diffusion à partir de données réels et corrigés . . . . .	22
4.1.1 Calcul des Fréquences génétiques : . . . . .	22
4.1.2 Comparaison des résultats obtenus à partir des données simulées/ et les données de capture : . . . . .	23
4.1.3 Interprétation : . . . . .	25
4.2 Partie II : Calcul des paramètres de diffusion à partir de données simulées . . . . .	28
<b>5 Conclusion :</b>	<b>29</b>

<b>5</b>	<b>Références :</b>	<b>30</b>
<b>5</b>	<b>Annexes :</b>	<b>30</b>

## 1 Présentation du stage :

### 1.1 Tableau des notations :

**D**urant l'écriture de ce rapport je vais devoir utiliser un grand nombre de notation similaire à celles utilisée dans le premier document de la bibliographie [11].

Le tableau suivant les résumera :

Notation	Utilité
$\Omega$	La région d'étude .
$\Omega^h$	Habitat ou verger h .
$u(t, x)$	Densité de dispersion.
$u^h(t, x)$	Densité de dispersion de l'habitat $\Omega^h$ .
$u_0(t, x)$	Densité de pré-dispersion.
$u_0^h(t, x)$	Densité de pré-dispersion de l'habitat $\Omega^h$ .
$\alpha$	Densité de pré-dispersion de l'ensemble des habitats.
$w_\infty(x)$	Densité de dispersion cumulé.
$w_\infty^h(x)$	Densité de dispersion cumulé à l'habitat $\Omega^h$ .
$X_t$	Coordonnées d'un individu au moment t.
$C_\tau$	Nombre d'individus capturés dans le piège $\theta_\tau$
$C_\tau^h$	Nombre d'individus cumulés capturés dans le piège $\theta_\tau$ .
$D(X)$	Coefficient de diffusion au point de coordonnées X.
$dt$	Pas de temps.
$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$	l'opérateur Laplacien.
$dX_t$	Déplacement élémentaire pendant un pas de temps dt.
$\beta_\tau$	Efficacité du piège $\theta_\tau$ .
$ \theta_\tau $	Surface du piège.
$\omega(x)$	Cumul de densité de population du temps initiale, $t_0$ à t, soit $\int_0^t u(x, t) dt$
$\omega^h(x)$	Cumul de densité de population du verger h de $t_0$ à t, soit $\int_0^t u^h(x, t) dt$
$v$	L'espérance de vie.
$G_{i\tau}$	Le génotype d'un individu i piégé dans le piège $\tau$ .
$P_{h\lambda_a}$	La fréquence allélique de l'allèle au locus $\lambda$ dans la sous population h.
$\tau \in [1 : H]$	Indice du piège.
$h \in [1 : H]$	Indice du Verger.
$\lambda \in [1 : \Lambda]$	Indice du Locus.
$i \in [1 : G]$	Indice du génotype.

## 1.2 Introduction :

### 1.2.1 Présentation de l'entreprise :

#### Présentation Globale :

**J**'ai effectué mon stage au sein de deux unités de l'I.N.R.A., un organisme français de recherche en agronomie fondé en 1946 ayant le statut d'Établissement public à caractère scientifique et technologique. Il est sous la double tutelle du ministère chargé de la Recherche et du ministère chargé de l'Agriculture. Cet organisme est considéré tel le premier institut de recherche agronomique en Europe et le deuxième dans le monde en nombre de publications en sciences agricoles et en sciences de la plante et de l'animal. L'I.N.R.A. mène des recherches afin d'améliorer la qualité de l'alimentation, ainsi que pour une agriculture durable et pour un environnement préservé et valorisé [2].

Cet institut est composé de vingt centres régionaux dont le centre I.N.R.A. d'Avignon, dont les recherches s'organisent autour de trois pôles de compétences : le pôle production horticole intégrée, le pôle adaptation au changement global et le pôle santé des plantes.

Pour conclure, voici une brève présentation des deux unités dans lesquelles j'ai effectué mon stage :

- **Biosp** : L'unité Biostatistique et processus spatiaux, c'est une unité qui dépend du département M.I.A.<sup>1</sup> Elle développe des travaux en statistique, en systèmes dynamiques, en écologie-épidémiologie et aux interfaces entre ces différentes disciplines avec un intérêt particulier pour les questions spatiales et spatio-temporelles [4].  
Cette unité compte parmi ses membres huit doctorants, une poste-doctorante, sept ingénieurs et treize chercheurs et est dirigée par Ètienne Klein.  
J'ai passé la totalité de mon stage au sein de cette unité.
- **PSH** : L'unité Plantes et Systèmes de Culture Horticoles, elle a pour mission de contribuer par des approches d'écophysiologie et d'agroécologie, à la mise au point de systèmes de culture des fruits et légumes et de scénarios paysagers en zone méditerranéenne afin d'améliorer la qualité des produits récoltés et le respect de l'environnement [5].

#### L'I.N.R.A. en chiffre :

**A**vec 3.981 publications référencées dans le Web of Science en 2013, la production scientifique de l'Inra a crû de plus de 69% depuis 2000, soit un taux moyen de croissance d'un peu plus de 4%.

De 2001 à 2011 [période pour laquelle ces données sont disponibles], la proportion d'articles parus dans les journaux les plus cités de chaque domaine disciplinaire est passée de 49% à 66%.

Par ces chiffres, l'I.N.R.A. s'impose comme un acteur majeur de la recherche agronomique dans le monde par la qualité de ses productions scientifiques. L'Institut est présent dans le top 1% des institutions les plus citées au monde dans 15 des 22 champs disciplinaires.

En nombre de citations référencées dans le Web of Science, l'Institut se situe en 3e position mondiale dans le domaine de l'agriculture et se place au 4e rang mondial en sciences des plantes et de l'animal. De plus, l'Inra se positionne dans les premiers organismes mondiaux en microbiologie, en écologie et en environnement.

Voici un tableau qui résume la position de l'Inra dans le Monde, Europe et en France en terme de cita-

1. Le département Mathématiques et Informatique Appliquées (M.I.A.) de l'I.N.R.A. développe la production de connaissances génériques, de mise au point de méthodes, d'outils et de savoir-faire, dans ses champs de compétences que sont les mathématiques et l'informatique appliquées aux domaines de l'alimentation, l'agriculture et l'environnement.

tion<sup>2</sup> :

Disciplines	Classement					
	D'après le nombre de citations reçues	D'après le nombre de publications*				
Sciences agronomiques	3/590	2	1	3	2	1
Sciences de l'animal & du végétal	3/1069	2	1	5	1	1
Microbiologie	24/435	7	3	13	4	3
Environnement / Ecologie	27/720	6	2	15	4	1

### 1.2.2 Projet de stage :

**M**on stage est accompli dans le contexte de l'écologie du paysage où les équations de réaction-diffusion<sup>3</sup> de réaction-diffusion [10] permettent de modéliser la dynamique spatio-temporelle d'une population en représentant l'évolution temporelle de la densité d'individus  $u(t,x)$  en chaque point de l'espace sous l'effet de la dispersion<sup>4</sup> et de la reproduction/mortalité<sup>5</sup>.

Ces équations sont des E.D.P. paraboliques de la forme :

$$\partial_t u(x,t) = \nabla(D(x)u) + f(x,u) \quad (1)$$

En particulier, dans un paysage hétérogène, un coefficient de diffusion  $D(x)$  variable dans l'espace (*déterminée par exemple par l'occupation du sol en chaque position.*) permet de modéliser des mobilités différentes des individus en fonction de l'environnement dans lequel ils se trouvent, estimer les coefficients de diffusion d'un organisme en fonction de l'hétérogénéité de l'environnement est un enjeu important en écologie du paysage notamment en termes de biologie de la conservation et de gestion des organismes nuisibles.

**D**ifférentes types d'observations sont habituellement utilisés pour suivre la dynamique spatio-temporelle d'un organisme :

- **Comptages** : Il a pour vocation à modéliser un nombre entier aléatoire évoluant dans le temps.
- **Suivi de trajectoires** : Est une méthode qui permet d'étudier les déplacements des animaux. Ceux-ci peuvent être suivis directement ou équipés d'émetteurs/récepteur GPS ou d'émetteurs VHF.
- **Captures-marquages-recaptures** : Une méthode d'inférence statistique couramment utilisée en écologie pour estimer la taille d'une population animale.

Dans un travail récemment publié [11] (Roques et al 2016), un modèle statistique a été développé pour des dispositifs où les individus collectés dans des pièges spatialisés sont génotypés pour plusieurs marqueurs génétiques par exemple locus micro-satellites.

Grâce à des simulations, il a été montré qu'il était possible d'estimer les coefficients de diffusion quand

2. Classement d'après le nombre de publications : l'ESI établissant ses classements à partir du nombre de citations, il est possible que des organismes plus productifs, mais moins cités ne soient pas comptabilisés sans possibilité de vérification.

3. **Équation de réaction-diffusion** : Ce sont des équations issues d'un modèle mathématique qui décrit l'évolution des concentrations d'une ou plusieurs substances spatialement distribuées et soumises à deux processus [3].

4. Terme de diffusion.

5. Terme de réaction.

on connaît les fréquences allèles des marqueurs des populations sources localisées dans l'espace et qu'elles sont génétiquement différenciées. Cette méthode par maximum de vraisemblance d'estimations où E.M.V.<sup>6</sup> des coefficients de diffusion s'appuie sur les proportions d'individus issus de chaque source initiale capturée dans chaque piège après diffusion. Les proportions estimées génétiquement sont comparées aux proportions qui sont calculées à partir d'un système d'équations de réaction-diffusion : une pour chacune des sources initiales.

Ainsi l'objectif de ce stage est d'évaluer les performances de la méthode et de proposer des pistes d'amélioration en vue de son application à des données de génétiques du paysage. Les analyses réalisées à ce jour suggèrent au moins trois pistes d'amélioration possibles, lesquelles pourront être toutes ou en partie intégrées à la méthode dans le cadre du stage. D'une part, il est rarement possible de connaître exactement les fréquences allèles dans les sources avant la diffusion, soit parce qu'on n'a échantillonné qu'un petit nombre d'individus avant la diffusion, soit parce que plusieurs générations se sont succédées entre l'échantillon d'apprentissage des fréquences allèles et l'échantillon d'individus piégés.

D'autre part, quand les sources sont trop faiblement différenciées, la vraisemblance d'appartenance à une source peut ne pas être assez résolutive. Utiliser des vraisemblances d'apparentement (Plein-frères, Demi-frères ou non Apparentés) entre paires d'individus pourrait être une option pour résoudre ce problème. Finalement, le modèle de système d'E.D.P. utilisé pourrait être amélioré pour prendre en compte un terme de convection.

Ce stage consistera à écrire les équations de réaction-diffusion, à les mettre en œuvre numériquement sur une mosaïque paysagère stylisée, à adapter les fonctions de vraisemblance pour corriger les faiblesses identifiées ci-dessus et enfin, à étudier les propriétés statistiques des estimateurs sur des jeux de données simulées tout en considérant des écarts aux hypothèses du modèle d'observation.

### 1.2.3 Intérêts du stage :

**C**e stage est très intéressant puisqu'il va me permettre d'avoir une première approche sur les métiers de la recherche ainsi que la compréhension du travail d'un chercheur. En abordant mes tâches, j'ai pu acquérir beaucoup de connaissance sur l'écologie et la génétique du paysage grâce aux différents documents qui avaient été mis en ma disposition et ceux que j'ai pu trouver sur internet. De plus, ce stage m'a permis d'acquérir quelques connaissances sur les principes de modélisation :

- La modélisation par des équations différentielles Partielles.
- La mise en œuvre numérique d'E.D.P.
- Les méthodes statistiques associées aux marqueurs génétiques (*écologie moléculaire*).
- Le modèle spatialisé et non-spatialisé.

De plus, ce stage m'a permis d'implémenter et/ou d'utiliser des algorithmes pour des méthodes vues en cours de statistiques, d'optimisation et analyse numérique ainsi que d'avoir des connaissances sur des méthodes que je verrais au second semestre comme les méthodes aux éléments finis et les statistiques bayésiennes.

Pour conclure cette partie, durant mon stage j'ai utilisé deux langages de programmation, le premier est

#### 6. Estimateur du Maximum de Vraisemblance :

Pour  $(X_1, \dots, X_n)$  un échantillon de variables aléatoires i.i.d. de loi  $P_\theta$ , on appelle E.M.V. la quantité :  $\widehat{\theta}_{MV}$  tel que

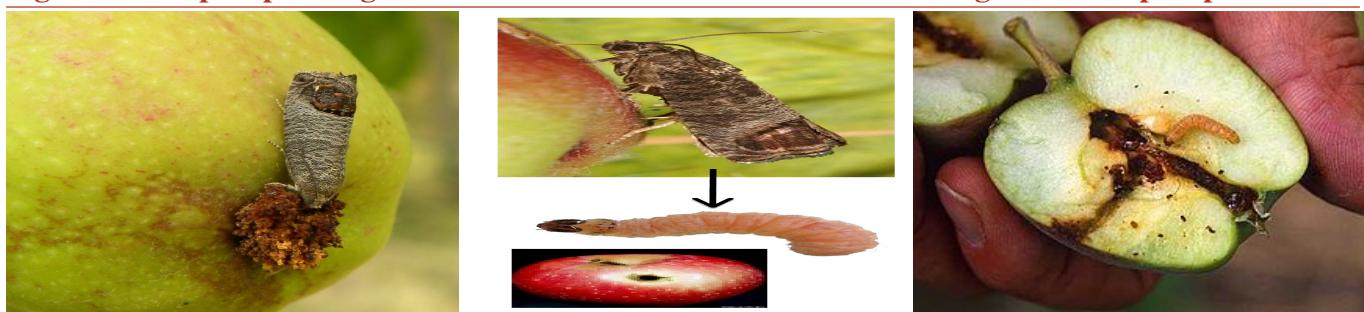
$$L(X, \widehat{\theta}_{MV}) = \max \prod_{k=1}^n f(x_i; \theta)$$

R(vue en cours de statistiques) et FreeFem++ qui est un langage de assez proche du c++. Ceci m'a permis d'assimiler et de comprendre quelques bases des applications mathématiques en écologie et en biologie.

#### 1.2.4 Objectives du stage :

**L**e carpocapse des pommes et des poires est un insecte de l'ordre des Lépidoptères, un ordre d'insectes holométaboles<sup>7</sup> [?] dont la forme adulte est communément appelée papillon, dont la larve est appelée chenille, et la nymphe chrysalide, de la famille des tortricidés, une famille qui regroupe des insectes microlépidoptères [?] dont les chenilles sont phytophages, dont la larve se développe à l'intérieur des fruits.

**Figure 1 : Carpocapse / Figure 2 : Pondaison d'un œuf dans un fruit /Figure 3 : Carpocapse chenille**



Ainsi, mon stage s'inscrit dans la continuité des travaux effectués par "**Boris Michenskié**" lors de son stage de fin d'étude sur l'analyse de dispersion du Carpocapse, tout en me basant sur les études menées par plusieurs chercheurs de l'I.N.R.A. [11] : L.Roques, E.Walker, P.Franck, S.Soubeyrand, E.K.Klein. Pour conclure, les objectives de ce stage sont :

- Appliquer l'approche théorique proposée par Lionel Roque afin d'estimer la diffusion des populations de carpocapse en fonction de l'hétérogénéité du paysage.  
Premièrement j'appliquerai cette méthode sur des données d'observations génétiques recueillies puis sur des données corrigées issues du F.Modèle qui permet de ré-échantillonner les données afin de pouvoir obtenir une distribution de l'ensemble des fréquences sur la totalité des vergers puis à recalculer la vraisemblance.
- Ensuite par l'intermédiaire de la méthode Quasi-Newton, je vais chercher les paramètres de diffusion optimaux afin de les comparer avec ceux trouvés auparavant. Cela sera effectué premièrement sur un modèle spatialisé, c'est à dire que nos probabilités dépendent des coordonnées spatiales puis non spatialisé.
- La seconde partie de mon stage consiste à simulé des densités de populations par la méthodes des éléments finis, puis les ajouter sur les fréquences alléliques afin de dégrader le modèle. Puis je vais estimer les paramètres de diffusion à partir d'un nombre de vergers fixé (2/4/8 ...) et d'individus fixés (2/4/8 ...) dans le but d'effectuer une approximation de l'impact du nombre de verger sur l'approximation de ces paramètres.

7. **Holométaboles**Qualifie les insectes chez qui le passage de l'état de larve à l'état adulte se fait par la transition d'un état de nymphe.

## 2 Approche théorique :

### 2.1 Quelques bases de biologies :

**U**ne grande partie des modèles de biostatistique repose sur les lois de Mendel et l'équilibre de Hardy-Weinberg que je vais présenter ci-dessous.

#### 2.1.1 Lois de Mendel :

Les lois de Mendel sont trois lois concernant les principes de l'hérédité biologique, énoncées par le moine et botaniste tchèque Gregor Mendel.

Dans le cadre de mon stage, je ne vais avoir recours qu'à la première et la troisième lois :

- **Loi d'uniformité des hybrides de première génération :** Le croisement deux races pures distinctes par un seul caractère implique que tous les descendants de la première génération sont identiques.
- **Ségrégation indépendante des caractères héréditaires multiples :** Cette règle ne s'applique que si les gènes responsables des caractéristiques se situent sur différents chromosomes ou s'ils sont éloignés sur le même chromosome. C'est le partage d'allèles dans des gamètes différents.

#### 2.1.2 L'équilibre de Hardy-Weinberg :

Dans une population théorique idéale, les fréquences des allèles et des génotypes au cours des générations suivent une loi simple appelée loi de Hardy-Weinberg qui constitue le modèle de référence en génétique des populations. Cette loi postule qu'au sein d'une population (idéale)<sup>8</sup>, il y a équilibre des fréquences allélique et génotypique d'une génération à l'autre [1]. À partir de ce modèle je vais calculer les probabilités génétiques à partir des fréquences alléliques par la loi de Hardy-Weinberg dont le principe est le suivant :

Soit  $A_1$  et  $A_2$  deux allèles d'un même locus

- On pose  $p$  la fréquence de l'allèle  $A_1$  avec  $0 < p < 1$ .
- Et  $q$  est la fréquence de l'allèle  $A_2$  avec  $0 < q < 1$ .

À partir de modèle et de la loi des probabilités totales, le calcul de nos fréquences génétiques s'effectuera de cette matière :

- La fréquence du génotype  $A_1A_1 \Rightarrow$  Homozygote est  $p^2$
- La fréquence du génotype  $A_1A_2 \Rightarrow$  Hétérozygote est  $2pq$
- La fréquence du génotype  $A_2A_2 \Rightarrow$  Homozygote est  $q^2$

Dans le cadre de mon projet, j'aurais entre 4 et 51 allèles par locus ainsi la formule sera beaucoup plus complexe. Dans la sous-section suivante je vais présenter les données que je vais utiliser durant mon stage.

---

8. **Population théorique idéale :** Se définit par les caractéristiques suivantes : Population close génétiquement à effectif infini où les croisements sont entièrement aléatoires d'organismes diploïdes à reproduction sexuée et à générations non chevauchantes. Tous les individus, quel que soit leur génotype, ont la même capacité à se reproduire et à engendrer une descendance viable avec une absence de mutation et de distorsion de ségrégation meiotique

## 2.2 Les données génétiques :

**L**es carpocapses étudié ont été collectés sur un ensemble de vergers de la basse vallée de la Durance, dont le suivi est assuré par l'INRA depuis plusieurs années. Dans ce travail, les individus collectés ont été analysés sur 20 marqueurs microsatellites. Les vergers sont considérés comme des populations sur lesquelles les fréquences alléliques ont été calculées.

Voici un résumé des données utilisées :

### → Échantillonnage :

- **Individus** : J'ai à ma disposition 995 individus génotypés durant l'année 2006 notre t=0 et 643 pris en 2007 notre t=1.
- **Vergers** : L'étude a été faite sur cinquante et un vergers parmi trois mille possibles ⇒ Notre échantillon ne représente que 1.66% de la population totale.

### → Données génétiques :

- **Locus** : J'étudie les fréquences alléliques situées sur vingt Locus/Loci, c'est à dire vingt emplacements physiques précis et invariables sur un chromosome.
- **Allèles** : Je mobilise entre quatre et cinquante allèles<sup>9</sup> par Locus.

Ci dessous un petit exemple des données que je vais utiliser durant mon projet, il s'agit d'un data.frame R contenant les noms des locus et les allèles.

### *Exemple de données utilisées :*

Ind	Parcelle	Cp1.180	Cp3.169	Cp3.179	Cp1.60	Cp1.61	Cp5.24	Cp4.5	Cp6.46	...
4	3	208 208	196 200	164 164	197 201	212 208	210 214	191 191	216	...
8	144	208 208	194 194	164 164	201 203	206 208	210 210	191 193	203	...
11	144	208 208	194 194	158 164	197 197	203 202	210 214	191 199	216	...
15	144	203 209	194 214	164 178	199 199	203 202	210 210	191 191	216	...
16	144	203 203	194 222	164 178	195 195	203 208	210 210	191 199	214	...
17	144	203 203	198 225	164 190	195 195	208 202	210 210	191 191	216	...
18	144	203 209	194 214	164 178	197 199	203 202	210 210	191 191	212	...
19	144	203 203	194 214	158 164	195 201	206 206	210 214	191 191	216	...
20	144	208 209	194 194	164 166	195 197	203 202	210 214	191 191	216	...
25	71	208 208	194 194	164 178	195 195	204 206	210 210	191 191	216	...
26	71	203 209	194 214	158 180	195 201	208 206	210 214	191 193	216	...
27	71	203 203	194 238	164 164	195 201	202 204	210 214	191 193	216	...
28	71	203 203	194 194	164 166	195 195	212 216	210 210	191 191	216	...

### Légende :

Dans cet extrait, il s'agit de la partie représentant, l'identifiant de l'allèle, sa passerelle<sup>10</sup> ainsi que sa position dans les différents locus : Cp3.180, Cp1.60 ...

## 2.3 Partie statistique :

**C**e sous-section se composera de quatre partie et traitera les différentes méthodes statistiques que j'ai eu à manipuler durant mon stage.

Tout d'abord je vais commencer par expliquer le modèle de capture, d'où étaient issues les données dites de capture, ensuite je vais expliquer brièvement le F-Modèle d'où sont extraites les données que j'ai principalement utilisées. Ensuite je passerais à l'explication des du modèle génétique, c'est à dire du passage des fréquences alléliques aux probabilités génétiques. Enfin, je conclurai cette partie par le caalcul de la vraisemblance.

9. Allèles : Versions variables d'un même gène ou d'un même locus génétique

10. C'est à dire le nom qu'on lui associe lors de notre implémentation

### 2.3.1 Le modèle de capture :

**P**our ce projet les individus étudiés ont été capturé par l'intermédiaires de pièges de petites tailles et d'après le document : "Using genetic data to estimate rates in heterogenous landscapes" [11], le cumule peut être considérer comme constant.

Ce qui permet d'obtenir l'équation suivante :

$$C_\tau = \beta_\tau \int_{\theta_\tau} \omega(x) d(x) = \beta_\tau |\theta_\tau| \omega(x) \quad (2)$$

De cette équation on peut trouver que le nombre d'individus arrivant de la source h capturés par un piége par piége est donné par :

$$C_\tau^h = \beta_\tau \int_{\theta_\tau} \omega^h(x) d(x) = \beta_\tau |\theta_\tau| \omega^h(x) \quad (3)$$

### 2.3.2 F-Modèle [9] :

**L**es données corrigées proviennent du F-Modèle, un modèle mis en œuvre par des chercheurs de l'inra pour l'estimation des fréquences alléliques de la population consensus. Il s'agit d'un modèle multinomial-Dirichlet, un type de modèle très utilisé en statistique bayésienne. Dans le contexte de modèle de méta population.

Dans le modèle, les Fst spécifiques à chaque sous-population, c'est-à-dire les probabilités que deux gènes choisis au hasard dans la sous-population et dans la population consensus aient un ancêtre commun dans l'union de ces populations, ce qui permet de considérer le cas de tailles de population et de taux de migration différents selon les populations .

#### Modèle :

Soit  $v = 1,..,V$  les vergers ( $V=51$ ),  $l = 1,..,L$  les locus (avec  $L=20$ ),  $a_l = 1,..,A_l$  les allèles au locus l.

Soit  $N_{v,l}$  le nombre de copies d'allèles dans le verger v au locus l. Les fréquences alléliques consensus sont notées  $FA_{cons,l,a}$ , les fréquences alléliques de chaque verger  $FA_{v,l,a}$ , et le nombre de copies d'allèles calculé sur les génotypes observés  $Ncopies_{v,l,a}$ .

On pose :

$$Ncopies_{v,l,a} \sim Multinomial(FA_{v,l,a}, N_{v,l})$$

$$FA_{v,l,a} \sim Dirichlet(FA_{cons,l,a} \theta_v)$$

$$FA_{cons,l,a} \sim Dirichlet(\alpha_l)$$

$$\theta_v = \frac{1}{Fst_v} - 1$$

Les lois a priori sont définies comme :

$$Fst_v \sim Normale(\mu, \tau)$$

Les Fst suivent une loi normale tronquée sur  $[0, 0.1]$  de paramètres  $\mu$  et  $\tau$ , avec  $\mu$  qui suit une normale de paramètres 0 et 0.1, et  $\tau$ . Dans un premier temps les  $\alpha_l$  sont fixés à 1. La vraisemblance est de la forme :

$$L(FAcons_{l,a}, Fst_v) = \prod_{v=1}^V \prod_{l=1}^L P(FAcons_{v,l,a} | FA_{v,l,a}, Fst_v)$$

Et la loi a posteriori s'écrit :

$$\pi(FAcons, Fst | Ncopies) \propto L(FAcons, Fst) \pi(FAcons) \pi(Fst) \quad (4)$$

Ainsi, à partir de cette loi là posteriori, 100 échantillons ont été créés. Dans le cadre de mon stage, je vais calculer et utiliser les fréquences alléliques moyennes de ces cent échantillons.

### 2.3.3 Le modèle génétique et calcul des Probabilités :

Le but de ce stage est d'estimer les paramètres du coefficient de dispersion à l'aide du maximum de vraisemblance, pour cela, le coefficient de dispersion a été calculé grâce à des simulations d'E.D.P. sur les densités des sous-populations et des densités cumulées.

Avant de calculer la vraisemblance je vais calculer les probabilités génétiques grâce aux fréquences alléliques en suivant les lois de Hardy-Weinberg ce qui me permet d'avoir la formule suivante :

$$P(G_{i\tau} | \Omega^h) = 2^k \prod_{\lambda=1}^{\Lambda} P_{h\lambda_{a1}} P_{h\lambda_{a2}} \quad (5)$$

Ainsi par l'intermédiaire de cette formule, on va calculé la table des probabilités génétiques.

### 2.3.4 Calcul de la vraisemblance :

On va étudier nos données sous deux cas :

- **Cas spatial** : c'est à dire un cas où la dispersion sera prise en compte. La probabilité qu'un individu provienne d'un verger donné est égale au taux d'individus capturés provenant de cet habitat.

Ainsi la probabilité spatiale est donnée par :  $\Pi_{h\tau} = \frac{C_\tau^h}{C_\tau}$

Du coup la vraisemblance sera calculée selon cette formule :

$$L(E) = \prod_{\tau=1}^H \prod_{i=1}^G \sum_{h=1}^H \left[ \frac{C_\tau^h}{C_\tau} \prod_{\lambda=1}^{\Lambda} P_{h\lambda_{a1}} P_{h\lambda_{a2}} \right] \quad (6)$$

- **Cas non spatial** : c'est à dire un cas où la dispersion ne sera pas prise en compte. Ainsi la vraisemblance sera indépendante des pièges  $\tau$  et contrairement au premier cas elle ne dépendra que de sources  $A$  et des densités  $\alpha$ . D'où afin de calculer la probabilité on utilisera cette formule :

$$\Pi_{h\tau} = \frac{\alpha_h * A_h}{\sum_{h'=1}^H [\alpha_{h'} * A_{h'}]}$$

La vraisemblance sera calculée par le biais de formule :

$$L(E) = \prod_{\tau=1}^H \prod_{i=1}^G \sum_{h=1}^H \left[ \Pi_{h\tau} \prod_{\lambda=1}^{\Lambda} P_{h\lambda_{a1}} P_{h\lambda_{a2}} \right] \quad (7)$$

## 2.4 Partie numérique :

Cette sous-section se composera de trois parties et traitera les différentes méthodes numérique que j'ai mobiliser durant mon stage.

Tout d'abord je vais commencer par expliquer l'équation différentielle qui rentre en jeu dans le calcul de la vraisemblance et dans le modèle spatialisé. Ensuite, j'exposerais les méthodes des éléments finis qui permettent de résoudre cette équation. Enfin, je vais expliciter les méthodes de Newton utilisé pour le calcul des paramètres de diffusion.

### 2.4.1 Modélisation par des E.D.P. :

- Comme expliqué auparavant, la probabilité spatiale est donnée par cette formule  $\Pi_{h\tau} = \frac{C_\tau^h}{C_\tau}$ .

- $C_\tau = \beta_\tau \int_{\theta_\tau} \omega^h(x) d(x)$  avec  $\omega(x) = \int_0^t u(x, t) dt$
- $C_\tau^h = \beta_\tau \int_{\theta_\tau} \omega^h(x) d(x) = \beta_\tau |\theta_\tau| \omega^h(x)$  avec  $\omega^h(x) = \int_0^t u^h(x, t) dt$
- Les densités de chaque sous-population d'origine  $\Omega - h$  sont issus de équation différentielle de Fokker-Planck :

$$\begin{cases} \frac{\partial u^h}{\partial t} = \Delta(D(x)u^h) - \frac{u^h}{v}, & t > 0, \forall x \in \Omega \\ u^h(0, x) = u_0^h(x), & \forall x \in \Omega \end{cases}$$

Ainsi la résolution de l'équation différentielle permettrait de déterminer les probabilité spatiales. Cette résolution s'effectuera par la méthode des éléments finis que je présenterais ci-dessous.

### 2.4.2 Méthodes des éléments finis :

Les méthodes aux éléments finis M.E.F ou F.E.M. sont des méthodes mathématiques utilisées pour résoudre numériquement des équations aux dérivées partielles.

Dans le cadre de mon stage, le calcul de vraisemblance nécessite les valeurs d'abondances des individus provenant de chaque verger par rapport à l'abondance totale. Ces densités de populations peuvent être calculées à l'aide du modèle de l'équation de Franck-Planck :

$$\begin{cases} \frac{\partial u^h}{\partial t} = \Delta(D(x)u^h) - \frac{u^h}{v}, & t > 0, \forall x \in \Omega \\ u^h(0, x) = u_0^h(x), & \forall x \in \Omega \end{cases}$$

Ce qui implique que pour simuler l'évolution spatio-temporelle de la densité de population je me base sur cette équation (*lors de ce stage la mortalité n'est pas prise en compte car elle n'influera pas sur les paramètres optimaux*).

La simulation consiste à discréteriser l'espace et le temps de l'équation 1. Plus précisément, je cherche une approximation de la solution sous la forme :

$$U_h = \sum \alpha_i \Phi_i$$

Avec comme inconnue  $\alpha = (\alpha_1 \dots \alpha_N)$  et  $\forall \Phi \in$  base de l'élément fini.

Ceci impose l'utilisation d'un schéma d'Euler implicite et un pas de temps  $t$ , ainsi que la formulation faible M.E.F. en espace<sup>11</sup> conduisant à un système de la forme :

$$\begin{cases} M = m_{ij} = \int_{\Omega} \Phi_i \Phi_j + \vec{\nabla} D \cdot \vec{\nabla} \Phi_j \Phi_i + \vec{\nabla} \Phi_j \cdot \vec{\nabla} \Phi_i D \\ b = \int_{\Omega} (U_{0h} + \frac{U_{h1}^{tm1}}{dt}) \Phi_j \end{cases}$$

### 2.4.3 Présentation de la méthode Quasi-Newton et utilisation :

**D**ans le cadre de mon stage je suis ramené à utiliser les méthodes de Newton afin de trouver les Zéros de la fonction dérivée de la log-vraisemblance qui représente les points d'inflexion<sup>12</sup> locaux de notre log-vraisemblance et donc nos maximums locaux de vraisemblance qui ne sont autre que nos paramètres de diffusion.

La méthode utilisée est la méthode Quasi-Newton [13] [8] qui n'est pas vraiment une méthode d'optimisation mais plutôt une méthode de recherche de zéros d'une fonction :

$$F : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

selon  $F(x) = 0$

Cette méthode est une imitation de l'algorithme de Newton, en tentant de calculer une approximation de  $\nabla^2 J$  et de son inverse. À l'itération  $k$ , on cherche à construire une approximation  $S^k$ , matrice symétrique définie positive, de  $[\nabla^2 J(x^k)]^{-1}$ , et  $\rho_k$  un paramètre positif donné par un algorithme de minimisation dans  $\mathbb{R}$ , le long de la direction de descente<sup>13</sup>  $d^k = -S^k \nabla J(x^k)$  tels que l'opération classique :

$$x^{k+1} = x^k - [\nabla^2 J(x^k)]^{-1} \nabla J(x^k)$$

Sera remplacée par l'opération plus simple et moins coûteuse :

$$x^{k+1} = x^k - \rho_k S^k \nabla J(x^k)$$

**C**omme expliqué auparavant, cette méthode sera appliquée afin de pouvoir trouver le zéros de la dérivée de la log-vraisemblance au maximum de la vraisemblance<sup>14</sup>. Ce maximum qui correspond à notre estimateur du Maximum de la Vraisemblance  $\widehat{\theta}_{MV}$ .

11. **Formulation faible M.E.F. :** Méthode ayant pour dessein la simplification de l'obtention des fonctions  $u^h$ .

12. **Point d'inflexion :** Un point où s'opère un changement de concavité d'une courbe plane.

13. Soit  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  et  $v \in \mathbb{R}^n$ . Le vecteur  $d \in \mathbb{R}^n$  est appelé **direction de descente pour  $J$  à partir du vecteur  $v$**  si la fonction  $t \in \mathbb{R} \mapsto J(v + td) \in \mathbb{R}$  est décroissante en  $t = 0$ , c'est-à-dire s'il existe  $\eta > 0$ , tel que  $\forall t \in [0, \eta]$

$$J(v + td) < J(v)$$

14. Car la fonction  $x \mapsto \log(x)$  est croissante.

Cette méthode est implémentée de cette manière :

### Algorithme DFP (Davidson-Fletcher-Powell)

1. Choisir  $S^0$  matrice sdp
2. On pose  $k = 0$
3. Choisir  $x^0 \in \mathbb{R}^n$  et  $\varepsilon > 0$
4. **Tant que**  $\|x^{k+1} - x^k\| \geq \varepsilon$  et  $k \leq k_{\max}$  **faire**
  - Calculer  $\rho_k$  par une méthode de recherche linéaire le long du vecteur  $d^k = -S^k \nabla J(x^k)$
  - Calculer  $x^{k+1} = x^k + \rho_k d^k$
  - Poser  $\delta^k = x^{k+1} - x^k = \rho_k d^k$
  - Calculer  $\gamma^k = \nabla J(x^{k+1}) - \nabla J(x^k)$
  - Calculer

$$S^{k+1} = S^k + \frac{\delta^k (\delta^k)^T}{(\delta^k)^T \gamma^k} - \frac{S^k \gamma^k (S^k \gamma^k)^T}{(\gamma^k)^T (S^k \gamma^k)}$$

Dans le cadre de ce T.P., on souhaite minimiser les paramètres du coefficient de diffusion  $\Rightarrow$  On va devoir trouver  $(\theta^*, \lambda^*)$  tel que  $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*, \theta_5^*)$  sera notre un point-selle de la fonction vraisemblance  $L$ . Cela fait que  $\theta^*$  sera un point de minimum de  $L$  sur son domaine de définition et donc  $(\theta^*, \lambda^*)$  devra simplement satisfaire les conditions du théorème des extrémas liés<sup>15</sup>, condition connue aussi sous l'appellation : "**Condition de qualification des contraintes**". Cela fait que nos solution doivent vérifier le système de conditions nécessaires de Lagrange :

$$\begin{cases} \nabla J(u^*) + \sum_{i=1}^p \lambda_i \nabla h_i(u^*) = 0 \\ h_i(u^*) = 0 \end{cases}$$

## 2.5 Présentation de la démarche :

### 2.5.1 Partie I : Calcul du coefficient de diffusion à partir des fréquences corrigées

**M**on objectif est d'améliorer l'estimation des paramètres du coefficient de diffusion des carpo-capses à partir des données récoltées en 2006 et 2007. Pour y parvenir, j'ai du mobiliser des méthodes statistiques et numériques qui m'ont permis en un premier temps de trouver les probabilités génotypiques à partir desquels on a effectué les calculs de la vraisemblance.

Ensuite on va mobiliser les méthodes de Newton afin de trouver les maximums locales de la vraisemblance qui seront nos coefficients de diffusion locaux.

15. **Théorème des extrémas liés** : Soit  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  différentiable et  $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$  de classe  $\mathcal{C}^1$ . On note

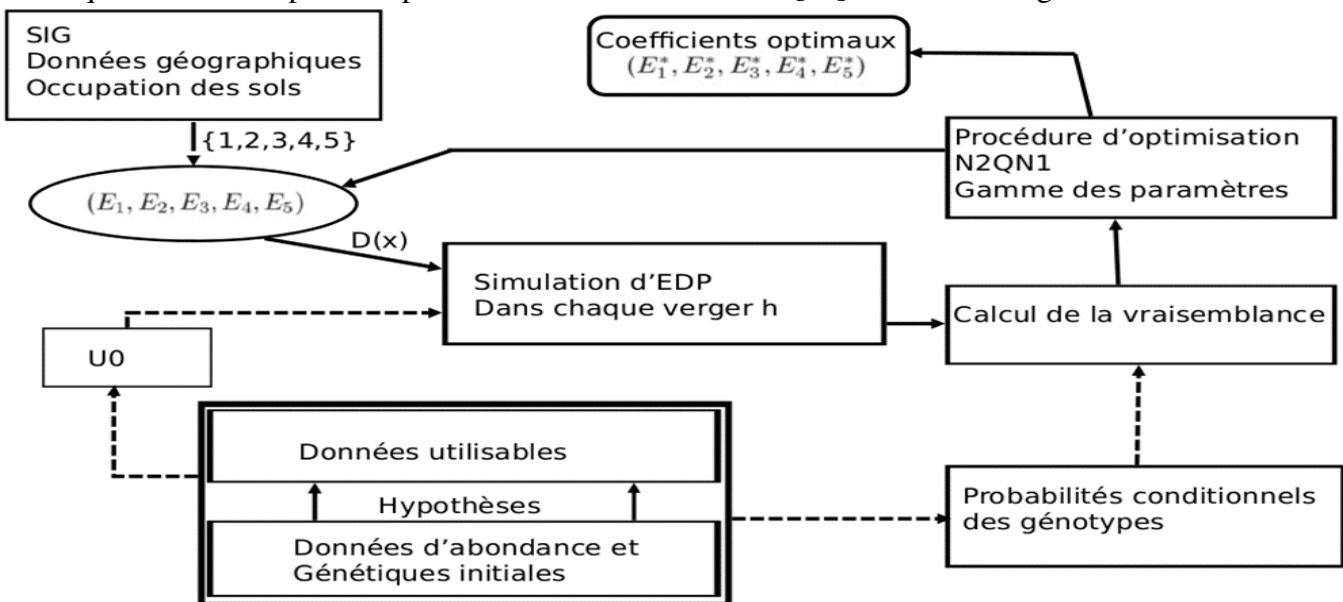
$$U = \{v \in \mathbb{R}^n, h_1(v) = \dots = h_p(v) = 0\}$$

On suppose que  $J$  admet un minimum local (ou  $-J$  admet un maximum local)  $u^*$  sur l'ensemble  $U$ , et que la famille  $\{\nabla h_1(u^*), \dots, \nabla h_p(u^*)\}$  est libre.

Alors il existe un  $p$ -uplet  $\{\lambda_1, \dots, \lambda_p\}$  de  $\mathbb{R}^p$  tel que

$$\nabla J(u^*) + \sum_{i=1}^p \lambda_i \nabla h_i(u^*) = 0$$

Voici un graphique qui résume la démarche nécessaire pour l'obtention des résultats. Le schéma est le même que celui utilisé par mon prénécessaire Boris Michenski [12] lors de son stage de fin d'étude :



### 2.5.2 Partie II :Calcul des coefficients de diffusion à partir des fréquences simulées :

Comme présenté lors de l'introduction, je vais commencer par une simulation des densités des vergers générée par l'intermédiaire de FreeFem ++ ensuite je vais les rajouter au fréquences alléliques avant d'effectuer le même plan d'optimisation. Voici un tableau expliquant la démarche effectuée :

Générer les fréquences simulé pour un coefficient de diffusion connu :

- On choisit un nombre de verger n avec les fréquences de 2006
- On simule à t=0 et t=1
- On calcule les fréquences alléliques à t=1.
- Pour chaque verger on génère N génotype à partir de nos fréquences alléliques.
- On effectue les précalculs de la table de probabilité génétiques.

On retrouve le coefficient de diffusion par optimisation :

- On simule les 10 populations
- Calcul de la vraisemblance à partir de N génotype
- On applique la B.F.G.S.
- On sort le coefficient de diffusion simulé

Approximation de l'erreur :

- Calcul de l'erreur par rapport au nombre de vergers.
- On trace le graphique du pourcentage de l'erreur.

### 3 Travail effectué :

#### 3.1 Outils utilisés :

##### 3.1.1 Pré-calculs et méthodes statistiques :

**P**our les pré-calculs et les méthodes statistiques, j'ai principalement utilisé libre office 5, un logiciel open source très proche d'Excel, il sera utilisé pour la lecture et la compréhension des données. Ensuite, on va utiliser R.Studio, un environnement de développement multi-plateforme open source pour R, un langage de programmation utilisé pour le traitement de données et l'analyse statistique. Cet environnement a été utilisé dans un premier temps pour le calcul des probabilités génétiques à partir de deux jeux de fréquences (probabilités) allèles.

Le premier jeu de données a été récolté et calculé à partir du modèle de capture, tandis que le second a été simulée par Emily Walker<sup>16</sup> et Julien Papaix<sup>17</sup> grâce à une méthode M.C.M.C.<sup>18</sup>(Voir F.Model).

Ensuite, on sera ramené à réutiliser ce logiciel dans la seconde partie de ce stage lors de la simulation des tables de fréquences allèles qui permettraient de faire une estimation des paramètres du coefficient de diffusion à partir des vergers.

##### 3.1.2 Optimisation :

###### Le calcul de la vraisemblance :

**P**our le calcul de la vraisemblance, je vais utiliser FreeFem++ [7], logiciel opensource essentiellement déployé pour résoudre numériquement des équations différentielles par éléments finis.

Ce logiciel a été créé en 1987 par Pascal Pironneau. Il a initialement été codé en Pascal mais sa dernière version possède son propre langage principalement inspiré du c++.

###### N2QN1

**P**our l'optimisation, on va utiliser N2QN1, une library qui contient des routines qui implémentent la méthode d'optimisation sous contrainte du type BFGS développée en fortran 77 puis par Claude Lemaréchal et Eliane Panier et qui est disponible sous licence libre de l'I.N.R.I.A. [6]. Il s'agit d'un solveur<sup>19</sup> qui permet réaliser les pas d'optimisations.

Exemple d'appel pour cette librairie :

```
subroutine n2qn1 (simul , n , x , f , g , dxmin , df1 , epsabs , imp , io , mode , iter , nsim
, binf , bsup , iz , rz , izs , rzs , dzs )
```

###### Cluster de Biosp :

**L**a taille des données ainsi que la complexité informatique des algorithmes m'ont poussé à soumettre mes programme au cluster de calculs Biosp, un outil qui a pour motif favoriser le calcul scientifique à

16. Ingénierie de recherche I.N.R.A.

17. Chercheur à l'I.N.R.A.

18. **Les méthodes de Monte-Carlo par chaînes de Markov, ou méthodes (M.C.M.C.)** : Sont une classe de méthodes d'échantillonnage à partir de distributions de probabilité. Ces méthodes de Monte-Carlo se basent sur le parcours de chaînes de Markov qui ont pour lois stationnaires les distributions à échantillonner.

19. Un programme logiciel permettant de calculer et fournir le résultat d'un problème après sa transcription informatique

haute performance. Il est constitué de 18 noeuds en plus de 196 cœurs pour une R.A.M. de 500 G.O.

### 3.2 Partie I : Calcul du coefficient de diffusion à partir des fréquences corrigées :

**A**vant d'effectuer le calcul de vraisemblance, je vais tout d'abord faire le calcul des probabilités génétiques partir des fréquences allèles de cette manière<sup>20</sup> :

Formule utilisée pour le calcul des probabilités génétiques :

$$P(G_{i\tau} | \Omega^h) = \prod_{\lambda=1}^{\Lambda} P_{h\lambda_{a1}} P_{h\lambda_{a2}}$$

La différence par rapport à la partie théorique est due au fait que le  $2^k$  ne changera pas le maximum de vraisemblance mais influera fortement sur le temps de calcul.

Il y a deux dates de prélèvement, l'une en 2006, l'autre en 2007. Les données correspondent à des individus génotypés. D'une part, les fréquences allèles sont calculées à partir des données de 2006, et d'autre part, pour chaque génotype de 2007, on pré-calcule la probabilité génotypique de provenir d'un des vergers de 2006. Ainsi, pour les données de capture, on sélectionne les fréquences allèles ainsi que les numéros de correspondances locus-allèle de cette manière :

ref	VNB	VNB	VNB	Sexe	début d'hergènes	Parcels	Rang	Altitude	Traitement	Type	Spécie	Altura_ECCO	Résistant	GPNs	
4	4	053247	630624	M	02/05/2007	3	4	2	0	052765	6306217	0	Ponier	0	0
8	8	053204	630458	M	15/06/2007	144	6	7	14	053360	6304581	0	Ponier	522	1
11	11	053140	630451	F	05/06/2007	144	5	1	14	053360	6304511	0	Ponier	267	0
15	15	053174	6304537	F	24/05/2007	144	4	2	14	053360	6304531	0	Ponier	166	0
16	16	053174	6304537	F	17/05/2007	144	4	2	14	053360	6304531	0	Ponier	640	1
17	17	053174	6304537	F	06/05/2007	144	4	2	14	053360	6304531	0	Ponier	NA	NA
18	18	053174	6304537	M	26/05/2007	144	4	2	14	053360	6304531	0	Ponier	78	0
19	19	053240	630450	F	06/05/2007	144	8	6	14	053360	6304501	0	Ponier	NA	NA
20	20	053105	630455	F	07/05/2007	144	6	10	14	053360	6304551	0	Ponier	161	0
25	25	054493	6304878	M	02/05/2007	71	1	4	9	054471	6304846	0	Ponier	121	0
26	26	054493	6304878	M	16/05/2007	71	1	4	9	054471	6304846	0	Ponier	1356	1
27	27	054510	6304870	F	30/04/2007	71	6	9	9	054471	6304846	0	Ponier	144	0
28	28	054510	6304870	F	09/05/2007	71	6	9	9	054471	6304846	0	Ponier	NA	NA
29	29	054510	6304870	M	30/04/2007	71	6	9	9	054471	6304846	0	Ponier	62	0
31	31	054417	6304814	F	02/05/2007	71	7	2	9	054471	6304846	0	Ponier	206	0
33	33	054417	6304814	M	11/05/2007	71	7	2	9	054471	6304846	0	Ponier	1329	1
34	34	054474	6304870	F	11/05/2007	71	1	3	9	054471	6304846	0	Ponier	114	1
36	36	054446	6304854	F	23/05/2007	71	1	1	9	054471	6304846	0	Ponier	307	1
39	39	054427	6304820	F	06/05/2007	71	7	3	9	054471	6304846	0	Ponier	NA	NA
40	40	054427	6304820	M	14/05/2007	71	7	3	9	054471	6304846	0	Ponier	1749	1
41	41	053069	6305207	F	24/05/2007	137	3	3	7	053360	6305206	0	Ponier	140	0
42	42	054493	6304861	M	25/05/2007	71	1	2	9	054471	6304846	0	Ponier	354	1

	c.0..191..193..195..199.
1	191
2	193
3	195
4	199

V1	V2	V3	V4	V5	V6
6 1.843480e-03	0.16912936	0.6833394	0.047621240	0.09482218	3.244389e-03
7 2.038252e-03	0.18779508	0.6712089	0.048806607	0.08752767	2.624108e-03
8 1.18491e-02	0.16573292	0.6922234	0.065585796	0.06465721	6.157760e-04
9 2.156544e-04	0.30335851	0.5502646	0.062241816	0.08391227	7.099782e-06
10 3.455719e-03	0.15849965	0.7000397	0.051633615	0.08363734	2.735952e-03
11 2.310992e-03	0.17137940	0.6782792	0.058404829	0.08803629	1.589240e-03
12 4.377013e-03	0.15822159	0.6841695	0.058949038	0.09391192	3.709439e-04
13 5.185297e-08	0.14160852	0.6689930	0.0732111191	0.11219666	4.000540e-03
14 2.188410e-08	0.27166432	0.6077992	0.056291804	0.06425161	6.1030181e-03
15 3.920627e-03	0.16592902	0.6773099	0.056571246	0.09490261	1.366544e-03
16 9.630839e-04	0.12028172	0.7598821	0.061919583	0.05553613	1.417410e-03
17 9.806848e-04	0.18530748	0.6523511	0.034126466	0.1254938	1.740411e-03
18 4.321637e-05	0.13050531	0.7376484	0.033205795	0.098594946	2.357671e-06
19 4.252986e-05	0.18257222	0.6991794	0.021591092	0.09661452	2.008923e-07
20 4.827430e-03	0.21265649	0.6609303	0.047063857	0.07335605	1.165847e-03
21 9.734740e-03	0.15263346	0.7173061	0.033583496	0.08697939	3.284051e-05
22 5.266966e-04	0.10090550	0.7099779	0.074040776	0.11454818	9.489188e-07
23 5.035949e-04	0.15501745	0.7336353	0.063900932	0.04649270	5.340012e-08
24 1.148529e-08	0.08321999	0.6057697	0.002446879	0.10455931	4.082063e-06
25 4.709685e-04	0.15967670	0.5959046	0.081318056	0.15510575	8.423906e-03
26 4.226504e-04	0.22175784	0.6300466	0.041402344	0.10070952	5.661200e-03
27 5.645852e-04	0.15822025	0.6840505	0.053163707	0.10262078	1.380156e-03
28 5.917645e-03	0.17959623	0.6846781	0.048912195	0.080121340	5.772210e-04

Showing 5 to 28 of 51 entries

- Je vais d'abord commencer par la lecture de nos tableau de données puis leur reconversion en data.frame.
  - Ensuite, on va nettoyer nos data.frame des erreurs de génotypage par l'intermédiaire de la fonction zerogohome avant de pouvoir extraire les allèles qui ne contiennent pas d'erreur.
  - Ensuite, je vais extraire les noms des allèles et des locus et les mettre dans dans des fichiers format csv.
20. Les détails sur les algorithmes utilisés dans cette partie seront donnés dans l'annexe I générée en Rmarkdown.

- Puis je vais trier les numéros des allèles par ordre croissant (*Optionnel*).
- À partir de cela je vais générer deux tables des probabilités génétiques, une première à l'aide des données de capture et des données corrigées moyenne<sup>21</sup>.

Aussitôt les tables de probabilités génétiques calculées je vais calculer les probabilités génotypiques spatiales puis non spatiales par l'intermédiaire de FreeFem++ afin de passer de cette formule  $P(g|\Omega^v) = \prod_{\lambda=1}^{\Lambda} (p_{v,\lambda_{a1}}, p_{v,\lambda_{a2}})$  à deux nouvelles formules :

- **Cas spatial :**

$$P(g_\tau|D) = \sum_{h=1}^H \frac{u_{h,\tau}}{u_\tau} P_{g,h}$$

L'implémentation de la log-vraisemblance se fera en trois parties, une première dans le premier programme qui sera notre simulateur et permettra de générer nos probabilités spatiales (Voir la probabilité spatiale), puis on va initialiser nos probabilités génotypiques dans un programme pre-computelikelihood avant de calculer la log-vraisemblance dans Computelikelihood. Voici quelques extraits de nos programmes :

#### Initialisation des tableaux dans FF++ :

```

1 for (int indiv=0; indiv<NGENOTYPES; indiv++){
2   fileVergerOfGenotype >> vergerOfGenotype [indiv];
3   auxx=0;
4   for (int verger=0; verger<NVERGERS; verger++){
5     proba >> tableProba(indiv, verger);
6     auxx+=tableProba(indiv, verger);
7   if (!(auxx>0)){
8     cout<<"bug genetic proba null with indiv "<<indiv<<endl;
9   exit(0);}}
```

#### Calcul de la vraisemblance dans FF++ :

```

1 for (int numG=0; numG<NGENOTYPES; numG++){
2   tau=vergerOfGenotype [numG]; // index du verger associe au genotype numG
3   computeUtau (tau);
4   uTotal [tau-1]=uTau;
5   if (uTau<=0){
6     cout<<"uTau is null ="<<uTau<<endl;
7   exit(1);
8   sigmaUtaui (NVERGERS*numG+ (tau-1))=0;
9   for (int numV=0; numV<NVERGERS; numV++){
10    sigmaUtaui (NVERGERS*numG+ (tau-1))+=(Ucenter [numV])*tableProba (numG, numV);
11   if (sigmaUtaui (NVERGERS*numG+ (tau-1))==0){
12     cout<<"erreur , sigmautau is null numG ="<<numG<<endl;
13   exit(1);
14   likelihood+=log (sigmaUtaui (NVERGERS*numG+ (tau-1))/uTau);
15 }}
```

21. Voir Annexe 1 pour plus de précision

- **Cas non spatial :**

$$P(g_\tau|D) = \sum_{h=1}^H \frac{\alpha_h * A_h}{\sum_{h'=1}^H [\alpha_{h'} * A_{h'}]} P_{g,h}$$

Dans ce cas(Voir la probabilité spatiale) je ne vais pas utiliser les probabilités spatiales et donc l'utilisation du simulateur n'est pas nécessaire. Ensuite, je vais utiliser un script shell qui permet à partir de la librairie N2QN1 présentée auparavant, d'appliquer l'algorithme de Quasi-Newton afin de calculer nos paramètres de diffusion qui seront les point d'inflexion des dérivées des nos fonctions log-vraisemblances et jouent donc les rôles des maximum de vraisemblances  $D = (\widehat{\theta}_{MV,i})_{i \in 1,5}$ .

### 3.3 Partie II : Calcul des coefficients de diffusion à partir des fréquences simulées :

**L**a seconde partie de mon stage se composera en deux sous parties :

- **Partie I : Simulation des fréquences génétiques et allélique :** Cette sous-partie se fera en deux temps :

- Dans un premier temps je vais générer les densités pour le calcul des probabilités spatiales. C'est à dire qu'on va générer les  $C_\tau^h$  en modifiant le code FreeFem++ afin de faire en sorte de pouvoir créer un fichier "**outputForR.txt**" contenant les densités des vergers simulés. Cela permet d'obtenir un tableau de cette forme :

*Exemple de tableau de densités pour 16 Vergers :*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
1	2.00442e-07	7.48691e-08	6.54880e-08	8.21275e-08	5.29071e-08	3.61659e-08	1.01651e-07	1.00608e-07	1.15383e-07	7.57307e-08	9.22240e-08	6.02639e-08
2	2.05787e-06	6.01313e-06	2.57560e-06	2.82894e-06	2.42455e-06	1.48961e-06	2.67363e-06	2.81568e-06	2.20056e-06	1.33996e-06	3.20298e-06	2.39295e-06
3	6.30121e-06	8.81465e-06	2.11963e-05	1.06602e-05	7.49049e-06	5.57207e-06	6.92179e-06	8.35632e-06	6.07762e-06	4.32364e-06	7.82386e-06	5.91956e-06
4	1.96578e-07	2.42600e-07	2.65272e-07	5.33409e-07	1.18269e-07	2.13199e-07	2.85017e-07	1.86261e-07	1.28258e-07	2.42804e-07	1.56339e-07	
5	9.55715e-07	1.57675e-06	1.42578e-06	4.13831e-06	1.35681e-06	1.15111e-06	1.19659e-06	9.98884e-07	6.46532e-07	1.28850e-06	1.42653e-06	
6	4.66595e-08	6.91167e-08	7.50372e-08	6.40600e-08	9.74694e-08	2.94936e-07	5.35789e-08	5.76624e-08	4.72842e-08	3.28163e-08	5.89583e-08	6.21559e-08
7	1.42343e-07	1.30308e-07	9.79096e-08	1.21080e-07	8.67336e-08	5.64766e-08	2.83983e-07	1.54183e-07	1.65981e-07	8.20116e-08	1.80684e-07	1.04455e-07
8	1.13330e-06	1.14622e-06	9.87184e-07	1.36301e-06	7.49052e-07	5.05285e-07	1.26220e-06	2.41521e-06	1.07940e-06	6.75917e-07	1.41104e-06	7.80219e-07
9	5.18623e-07	3.37010e-07	2.70541e-07	3.31996e-07	2.37654e-07	1.57667e-07	5.13745e-07	4.09933e-07	8.60295e-07	2.90046e-07	4.34083e-07	2.91771e-07
10	2.69485e-07	1.74742e-07	1.64598e-07	1.95122e-07	1.32175e-07	9.39283e-08	2.14163e-07	2.15918e-07	2.38882e-07	6.52347e-07	2.01091e-07	1.49520e-07
11	4.20177e-07	5.07473e-07	3.62440e-07	4.54322e-07	3.12457e-07	2.02009e-07	5.94553e-07	5.73702e-07	4.47259e-07	2.50873e-07	9.27737e-07	5.31538e-07
12	8.82305e-07	1.24864e-06	8.95524e-07	9.46811e-07	1.14827e-06	6.95509e-07	1.11979e-06	1.00158e-06	9.90807e-07	5.94288e-07	1.15929e-06	3.06310e-06
13	1.90947e-06	2.88821e-06	2.72861e-06	2.46358e-06	4.27509e-06	4.33372e-06	2.26878e-06	2.31614e-06	2.00118e-06	1.31934e-06	2.46507e-06	3.03371e-06
14	3.33775e-05	1.99436e-05	2.18013e-05	1.84990e-05	2.82268e-05	5.07663e-05	1.53730e-05	1.65965e-05	1.35457e-05	9.38069e-06	1.69533e-05	1.76579e-05
15	2.08797e-07	3.05436e-07	3.29500e-07	2.83695e-07	4.24935e-07	7.62804e-07	2.39067e-07	2.56425e-07	2.11603e-07	1.47794e-07	2.62182e-07	2.80229e-07
16	8.09624e-06	1.22910e-05	1.16920e-05	1.05117e-05	1.84173e-05	1.87391e-05	9.60973e-06	9.84695e-06	8.46721e-06	5.59193e-06	1.04618e-05	1.27140e-05

- Ensuite, pour chaque locus, on va initialiser nos fréquences allélique à t=0, puis à l'aide de notre table de densité, on va simuler les fréquences à la date t=1 par l'intermédiaire de cet algorithme qui prend entrée le chemin de la table des densités, celle des fréquences à la date t=0 et le numéro du locus et qui retourne la table des fréquences allèles simulées à la date t=1.

Ensuite on crée nos fichiers de tables de fréquences allèles pour (2v,4v,8v,16v) et pour 20 locus soit 80 tables.

Voici l'algorithme qui permet de générer nos tables simulés :

```

1  InitialisationFsimule <- function (chemin , chemin2 , number) {
2      trouve<-paste(chemin , number , ".csv" , sep="")
3      a<-read.csv(trouve , sep=" ")
4      F0f=a[1 ,]
5      n=nverger
6      for (k in 2:n){
7          F0=as.numeric(a[k ,])
8          F0f<-rbind(F0f,F0)
9      }
10     n1 = dim(F0f)[1]
11     n2=dim(F0f)[2]
12     F1f<-c()
13     f2f<-list()
14     U<-calculdesUgene(chemin2 , n)
15     F1f2<-matrix(0 , nrow=n1 , ncol=n2)
16     for (i in 1:n1){
17         for (j in 1:n1){ #car u est une matrice de taille n1=8*n1=8
18             F1f<-F0f[j ,]* (U[i , j]/U[i , n1+1])}

```

```

19      for (k in 1:n2){
20        F1f2[i,k]<- unlist(F1f[k])
21      }
22    }
23  return(F1f2)
}

```

Ainsi, à partir des densité qu'on a générée par l'intermédiaire de FreeFem++, les fréquences alléliques

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	192	194	196	198	200	206	208	210	212	214	215	216	218	220	222	224	226	228	230	232
3	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
4	0.03386940	0.02357832	0.00598029	0.02859882	0.03335675	0.01180747	0.02413446	0.01208276	0.003986127	0.000247396	0.001723507	1.250294e-04	0.002264914	0.001093291	0.0201537	0.0165474				
5	0.005365476	0.024112794	0.006119478	0.002721895	0.02413446	0.01208276	0.003986127	0.000247396	0.001708556	1.279349e-04	0.00217038	0.001124923	0.0201537	0.0165474						
6	0.003059760	0.02194688	0.005566889	0.02476113	0.0105127	0.01099104	0.0035626187	0.0002162940	1.635825e-04	0.000217038	0.0010239441	0.0201537	0.0165474							
7	0.003886115	0.02751774	0.006879393	0.010310479	0.03863993	0.013779381	0.0049456262	0.002735855	0.000217038	1.459123e-04	0.000264661	0.001282990	0.0201537	0.0165474						
8	0.002546483	0.01626677	0.0004653048	0.002607483	0.0258429	0.009147322	0.000201791	0.0001618593	0.0001535122	9.685952e-05	0.0001754228	0.00081674	0.0201537	0.0165474						
9	0.002140121	0.01535181	0.0003895717	0.01731894	0.02171853	0.007687614	0.00015358103	0.0001528386	0.0001125429	8.40264e-05	0.0001474291	0.007157681	0.0201537	0.0165474						
10	0.003791140	0.02719512	0.006897566	0.036867983	0.0347351	0.01361894	0.000492391	0.000270475	0.00011993652	1.442016e-04	0.00021648	0.001287950	0.0201537	0.0165474						
11	0.006748730	0.04541980	0.002278577	0.05461414	0.05840792	0.02424279	0.000799863	0.000491865	0.003548983	2.556974e-04	0.000468073	0.002571290	0.0201537	0.0165474						



V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	
1	0.0036940	0.025702	0.005980229	0.0285988	0.03335675	0.01180747	0.000247396	0.0001723507	1.260294e-04	0.000264914	0.001093291	0.0201537	0.0165474							
2	0.0036940	0.024112794	0.006119478	0.02413446	0.01208276	0.003986127	0.000247396	0.0001723507	1.279349e-04	0.000217038	0.001124923	0.0201537	0.0165474							
3	0.0036940	0.02194688	0.005566889	0.02476113	0.0105127	0.01099104	0.0035626187	0.0002162940	1.635825e-04	0.000217038	0.0010239441	0.0201537	0.0165474							
4	0.0036940	0.02751774	0.006879393	0.010310479	0.03863993	0.013779381	0.0049456262	0.002735855	0.000217038	1.459123e-04	0.000264661	0.001282990	0.0201537	0.0165474						
5	0.0036940	0.01626677	0.0004653048	0.002607483	0.0258429	0.009147322	0.000201791	0.0001618593	0.0001535122	9.685952e-05	0.0001754228	0.00081674	0.0201537	0.0165474						
6	0.002140121	0.01535181	0.0003895717	0.01731894	0.02171853	0.007687614	0.00015358103	0.0001528386	0.0001125429	8.40264e-05	0.0001474291	0.007157681	0.0201537	0.0165474						
7	0.003791140	0.02719512	0.006897566	0.036867983	0.0347351	0.01361894	0.000492391	0.000270475	0.00011993652	1.442016e-04	0.00021648	0.001287950	0.0201537	0.0165474						
8	0.006748730	0.04541980	0.002278577	0.05461414	0.05840792	0.02424279	0.000799863	0.000491865	0.003548983	2.556974e-04	0.000468073	0.002571290	0.0201537	0.0165474						



V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
1	2.00442e-07	7.48691e-08	6.54880e-08	8.21275e-08	5.29071e-08	3.61569e-08	1.01651e-07	1.00608e-07											
2	2.09787e-06	6.01313e-06	2.57560e-06	2.82894e-06	2.42455e-06	1.48961e-06	2.67363e-06	2.81568e-06											
3	6.30121e-06	8.81465e-06	2.11963e-05	1.06602e-05	7.49049e-06	5.57207e-06	9.62179e-06	8.35632e-06											
4	1.96578e-06	2.42600e-07	2.65272e-07	5.33409e-07	1.69407e-07	1.18269e-07	2.13199e-07	2.85017e-07											
5	9.56716e-07	1.57675e-06	1.42578e-06	1.28596e-06	4.13831e-06	1.35681e-06	1.15311e-06	1.19699e-06											
6	4.66593e-08	6.91167e-08	7.50372e-08	6.40600e-08	9.74694e-08	2.94936e-07	5.35736e-08	5.76624e-08											
7	1.42343e-07	1.30308e-07	9.79096e-08	1.21080e-07	8.67336e-08	5.64766e-08	2.83983e-07	1.54183e-07											
8	1.13330e-06	1.14622e-06	9.87184e-07	1.36301e-06	7.49052e-07	5.05285e-07	1.26220e-06	2.41521e-06											



V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
1	0.002442e-07	0.00748691e-08	0.00654880e-08	0.00821275e-08	0.00529071e-08	0.00361569e-08	0.0101651e-07	0.0100608e-07											
2	0.00209787e-06	0.00601313e-06	0.00257560e-06	0.00282894e-06	0.00242455e-06	0.00148961e-06	0.00267363e-06	0.00281568e-06											
3	0.00630121e-06	0.00881465e-06	0.00211963e-05	0.00106602e-05	0.00749049e-06	0.00557207e-06	0.00962179e-06	0.00835632e-06											
4	0.00196578e-06	0.00242600e-07	0.00265272e-07	0.00533409e-07	0.00169407e-07	0.00118269e-07	0.00213199e-07	0.00285017e-07											
5	0.00956716e-07	0.00157675e-06	0.00142578e-06	0.00128596e-06	0.00413831e-06	0.00135681e-06	0.00115311e-06	0.00119699e-06											
6	0.00466593e-08	0.00691167e-08	0.00750372e-08	0.00640600e-08	0.00974694e-08	0.00294936e-07	0.00535736e-08	0.00576624e-08											
7	0.00142343e-07	0.00130308e-07	0.00979096e-08	0.00121080e-07	0.00867336e-08	0.00564766e-08	0.00283983e-07	0.00154183e-07											
8	0.00113330e-06	0.00114622e-06	0.00987184e-07	0.00136301e-06	0.00749052e-07	0.00505285e-07	0.00126220e-06	0.00241521e-06											

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
1	1.002442e-07	0.00748691e-08	0.00654880e-08	0.00821275e-08	0.00529071e-08	0.00361569e-08	0.0101651e-07	0.0100608e-07											
2	1.00209787e-06	0.00601313e-06	0.00257560e-06	0.00282894e-06	0.00242455e-06	0.00148961e-06	0.00267363e-06	0.00281568e-06											
3	1.00630121e-06	0.00881465e-06	0.00211963e-05	0.00106602e-05	0.00749049e-06	0.00557207e-06	0.00962179e-06	0.00835632e-06											
4	1.00196578e-06	0.00242600e-07	0.00265272e-07	0.00533409e-07	0.00169407e-07	0.00118269e-07	0.00213199e-07	0.00285017e-07											
5	1.00956716e-07	0.00157675e-06	0.00142578e-06	0.00128596e-06	0.00413831e-06	0.00135681e-06	0.												

## 4 Résultats obtenus :

#### **4.1 Partie I : Calcul des paramètres de diffusion à partir de données réels et corrigés**

#### 4.1.1 Calcul des Fréquences génétiques :

À partir du script présenté dans l'annexe 1, on a calculé le tableau de fréquence génétiques à partir des formules de probabilités conditionnels. Ces calculs ont été fait à partir de deux types de données, premièrement, par l'intermédiaire des données récoltés puis par des données qui ont été simulées à partir de la méthode de Monte-Carlo par chaîne de Markov présenté dans la section F-Modèle.

*Tableau généré par les vraies données :*

*Tableau généré par des données corrigées :*

	V1	V2	V3	V4	V5	V6	V7
1	1.185926e-46	6.041144e-41	5.875618e-40	6.706207e-40	2.082045e-47	1.212887e-37	7.513324e-39
2	2.765098e-41	2.077058e-39	1.637576e-39	1.470729e-38	7.062759e-43	1.065735e-37	3.130038e-38
3	2.214564e-44	1.400628e-39	3.844777e-41	1.199109e-38	1.104140e-46	1.130620e-36	6.707222e-38
4	3.717747e-40	8.945183e-37	2.508228e-36	1.514189e-35	2.795032e-42	1.214700e-35	9.764186e-37
5	1.016384e-66	1.144611e-63	2.522762e-53	1.440752e-48	3.923006e-67	3.317548e-52	3.198936e-52
6	1.510702e-53	4.205249e-51	4.480011e-51	2.315297e-46	6.077938e-54	2.855826e-46	1.488864e-46
7	8.457861e-57	1.177977e-46	2.434614e-44	5.749991e-41	2.449395e-49	2.513401e-39	2.083058e-41
8	8.520565e-52	2.693556e-45	6.683590e-43	6.372251e-43	2.024153e-51	4.096001e-42	3.886128e-42
9	4.218711e-38	1.328032e-37	2.262782e-36	1.038227e-37	9.500099e-35	1.397854e-36	4.798462e-36
10	2.169734e-48	4.724379e-40	4.585752e-42	2.045261e-38	5.574898e-48	9.389321e-37	5.658421e-37
11	4.794565e-43	2.359229e-46	1.713126e-48	8.835759e-38	2.581860e-45	1.838770e-37	1.702445e-38
12	7.872882e-39	2.294922e-35	8.873464e-36	6.447692e-35	1.336823e-41	1.126054e-33	1.233115e-34
13	4.269510e-47	1.215590e-48	7.774493e-41	4.138714e-39	1.390777e-44	6.156052e-42	1.998858e-42
14	1.978842e-49	4.074559e-43	3.524910e-45	5.997458e-39	1.828499e-47	1.172677e-39	2.474928e-38
15	5.453843e-39	3.821412e-37	6.799607e-35	1.184887e-33	2.058446e-37	8.459079e-34	4.046629e-34
16	1.533355e-48	4.469130e-41	4.402457e-42	9.018151e-39	9.066556e-50	1.257084e-38	5.246592e-38
17	9.935987e-46	2.533382e-47	6.014797e-47	9.824600e-42	1.292735e-62	6.429714e-45	5.756812e-42
18	1.075581e-49	1.235393e-43	7.383238e-46	5.043416e-41	7.138336e-48	1.163443e-40	1.623935e-40
19	3.928261e-41	2.712048e-39	3.427217e-39	4.054664e-38	1.037114e-46	1.472860e-37	7.197519e-37
20	7.069340e-39	7.452035e-37	1.402252e-39	6.895216e-35	3.624248e-38	3.072866e-36	2.012924e-35

## Présentation des tableaux :

**I**l y a deux dates de prélèvement des données, premièrement, 2006 qui correspond à  $t = 0 = t_0$ , puis, 2007 qui correspond à  $t = 1 = t_1$ . D'une part, les fréquences alléliques sont calculées à partir des données de 2006, et d'autre part, pour chaque génotype de 2007, on pré-calcule la probabilité génotypique de provenir d'un des vergers de 2006.

En ce qui concerne les données de capture, on constate que des génotypes 2007 ont des locus non rencontrés dans les génotypes 2006, cela conduit à une vraisemblance nulle. Pour sortir de cette impasse (*optimiser une fonction nulle*), on ajoute un 52<sup>ème</sup> verger/source fictif constitué d'une population avec un génotype moyen. Ce 52<sup>ème</sup> verger sera aussi rajouté pour les données corrigées afin de pouvoir faire la comparaison de ces tableaux.

## Comparaison des résultats :

**À**partir de cet extrait, on peut constater que le tableau calculé à partir des données corrigé semble être meilleur que celui obtenu à partir des données de capture. Cela se montre à plusieurs niveaux :

- ★ Premièrement, on ne peut observer que les fréquences maximales ne se trouve que dans le 52<sup>ème</sup> verger pour le tableau généré à partir des données de capture, ce qui permet de dire que l'information n'est portée que par le verger fictif, contrairement au second tableau où l'information est portée par plusieurs vergers.
- ★ Deuxièmement, le premier tableau contient essentiellement des zéros ce qui permet de dire que le calcul de la table de probabilité génétique à partir des données de capture ne permet de détecter les allèles rares. Le second tableau, quand à lui, permet d'avoir des faibles fréquences même pour les allèles rares.

Cela peut s'expliquer par le fait que l'on ne dispose que des données récoltés sur 51 vergers parmi 3000 soit 0.51%, ces résultats montrent aussi la pertinence du F-Modèle qui tient sa force du fait qu'il est basé sur un approche Bayésienne dont le but est de donner les mêmes résultats que les statistique fréquentiste<sup>22</sup> par des procédés moins coûteux en calcul.

### 4.1.2 Comparaison des résultats obtenus à partir des données simulées/ et les données de capture :

Tout d'abord, le code que j'ai utiliser jusqu'au début de décembre ne tenait pas compte des données d'abondance, ce qui a fait qu'on a obtenu des résultats de la forme :

- ★ Données de capture :

➔ **Modèle spatialisé** : On obtient un log-vraisemblance de **2116323.7822** et des paramètres de diffusion :  $(10^{10}, 10^{10}, 10^{10}, 10^{10}, 10^{10})$  avec des gradients très proches de zéro et qui sont égaux à  $(-1.49 \times 10^{-8}, -2.02 \times 10^{-8}, -5.67 \times 10^{-9}, -8.34 \times 10^{-9}, -1.45 \times 10^{-9})$ .

Ceci pousse à dire que nos paramètres de diffusion vérifient la condition du premier ordre c'est à dire le théorème de K.K.T. :

$$KKT \left\{ \begin{array}{l} \nabla J(u^*) + \sum i = 1^q \lambda_i^* \nabla g_i(u^*) = 0 \\ \lambda_i^* g_i(u^*) = 0 \quad \forall i \in \{1, \dots, q\} \end{array} \right.$$

➔ **Modèle non spatialisé** : On obtient une log-vraisemblance de **2088413.9312** soit une différence 1.31% et une log-vraisemblance plus petite. Ce résultat permet de dire que les données sont mal spatialisé dans l'espace.

22. ou inférentielle

★ Données de capture corrigées :

- ➡ **Modèle spatialisé** : On obtient un vraisemblance de **2819308.4720 2814962.6577** et des paramètres de diffusion :  $(10^{10}, 10^{10}, 10^{10}, 10^{10}, 10^{10})$  avec des gradients très proches de zéro et égaux à :  $(-1.33 \times 10^{-8}, -1.89 \times 10^{-8}, -7.17 \times 10^{-9}, -8.41 \times 10^{-9}, -1.14 \times 10^{-9})$  qui vérifient la condition du premier ordre.
- ➡ **Modèle non spatialisé** : On obtient une vraisemblance de **2814962.6577** soit une différence 0.15% et une vraisemblance plus grande. Ce résultat permet de dire que ces données sont mieux spatialisés dans l'espace.

On peut interpréter ces résultats en disant que notre zone d'étude de  $10km^2$  paraît insuffisante, car un individu/une population est apte à traverser toute notre zone d'un côté à l'autre. Et ce résultat reste valable à la fois pour les données de capture et corrigées.

Cependant, avec mon tuteur effectue "Olivier Bonnefon", on a remarqué que le calcul de vraisemblance ne tient pas en compte les données d'abondance. Ainsi le calcul de la vraisemblance se changera de cette manière :

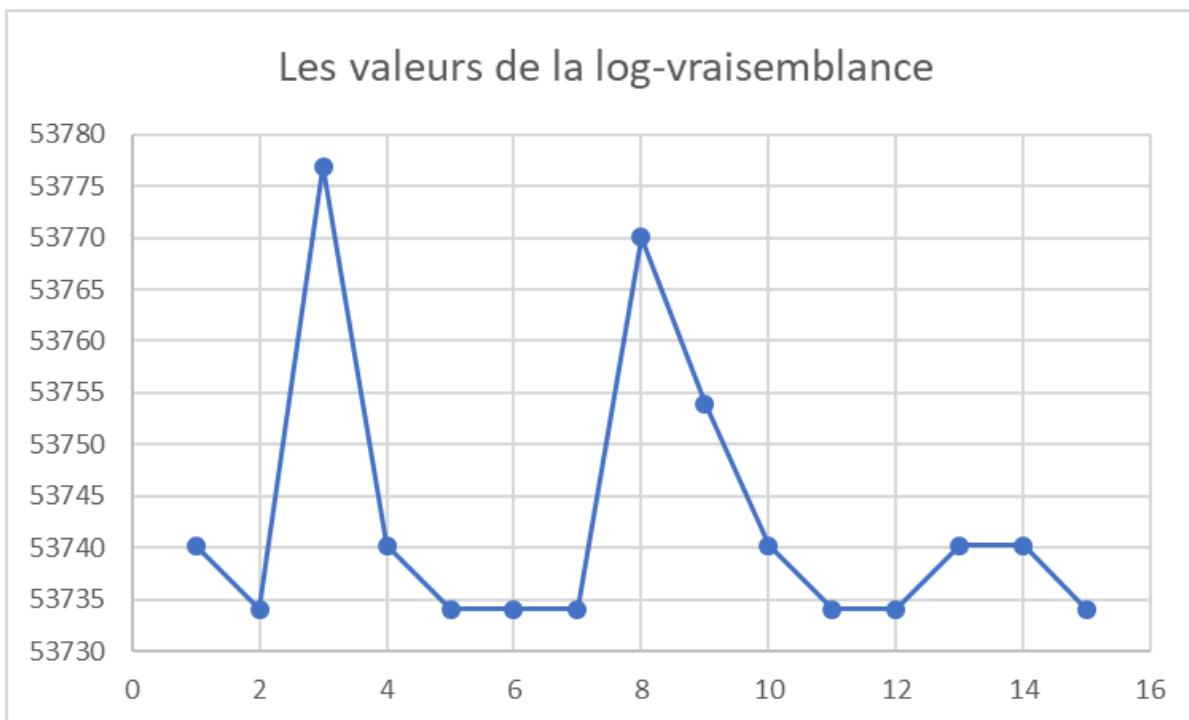
$$L(\Theta) = \prod_{\tau=1}^H \prod_{i=1}^N \sum_{h=1}^H \frac{u_{h,\tau}}{u_\tau} p_{h,i}$$

↔

$$L(\Theta) = \prod_{g=1}^G \sum_{h=1}^H \frac{u_{h,\tau}}{u_\tau} P_{g,h} \quad \text{où } \tau \text{ désigne le piège associé au génotype } g$$

La prise en compte des données d'abondance m'a permis de montrer premièrement que les données corrigées sont mieux spatialisées.

Ensuite, sur les différents calculs de la log-vraisemblance on peut observer des valeurs qui ne varient pas beaucoup.



À partir de cela, on peut dire en un premier temps que la log-vraisemblance passe de **2116323.7822** à une valeur moyenne de **53742,71244**. De plus, sur dix paramètres générés aléatoirement par l'intermédiaire de ce code :

```

1 #! /usr/bin/python
2 from random import uniform
3
4 nParam=5
5 fichier = open( 'optimParam.txt' , 'w' )
6 fichier.write( str(nParam)+'\n' )
7 fichier.write( str(1e-9)+'\n' )
8 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
9 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
10 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
11 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
12 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
13 fichier.write( str(100)+'\t'+str(10000000000)+'\n' )
14 for param in range(nParam):
15     a=uniform(5, 10)
16     b=pow(10,a)
17     fichier.write( str(b)+'\n' )
18
fichier.close()

```

createParam.py

En ce qui concerne les paramètres de diffusion, on obtient des résultats plus pertinents.

Voici un tableau présentant les différents paramètre d'optimisation sur différents :

	A	B	C	D	E	F
1	-LOG(D1)	-LOG(D2)	-LOG(D3)	-LOG(D4)	-LOG(D5)	-Log(L(D))
2	2	4,13511	8,14448	3,05146	2	53740,2673674267
3	2	2	2	8,23785	4,81075	53734,057516633
4	9,58175	4,51312	2	6,72061	2	53776,9228222814
5	2	4,13511	8,14449	3,05097	2	53740,2673674265
6	2	2	2	8,23789	4,8111	53734,0576382987
7	2	2	2	8,23815	4,80814	53734,0576429568
8	2	2	2	8,23791	4,80759	53734,0576386401
9	9,1267	2	4,13241	2	4,21598	53770,0585959837
10	8,49186	2	2	2	3,72971	53753,9689358515
11	2	4,16744	8,14424	2	2	53740,2660570053
12	2	2	2	8,23779	4,8158	53734,0576392757
13	2	2	2	8,2379	4,80942	53734,0576383882
14	2	4,16765	8,14425	2	2	53740,2660573923
15	2	4,1642	8,14396	2	2	53740,266052034
16	2	2	2	8,2379	4,81116	53734,0576382972

Comme pour les premiers calcul nos résultats vérifie les condition de **Système de condition nécessaire de Lagrange** car les gradients sont de l'ordre de  $10^{-10}$ , je prend comme exemple la première valeur : **(8.74e-10,5.51e-10,-9.07e-10,9.24e-10,1.24e-10)**.

#### 4.1.3 Interprétation :

Premièrement, il faudra expliquer ce que signifie les cinq paramètres de diffusion. Ils correspondent aux zones de bois, cultures annuels, vergers, zones urbaines et les prairies.

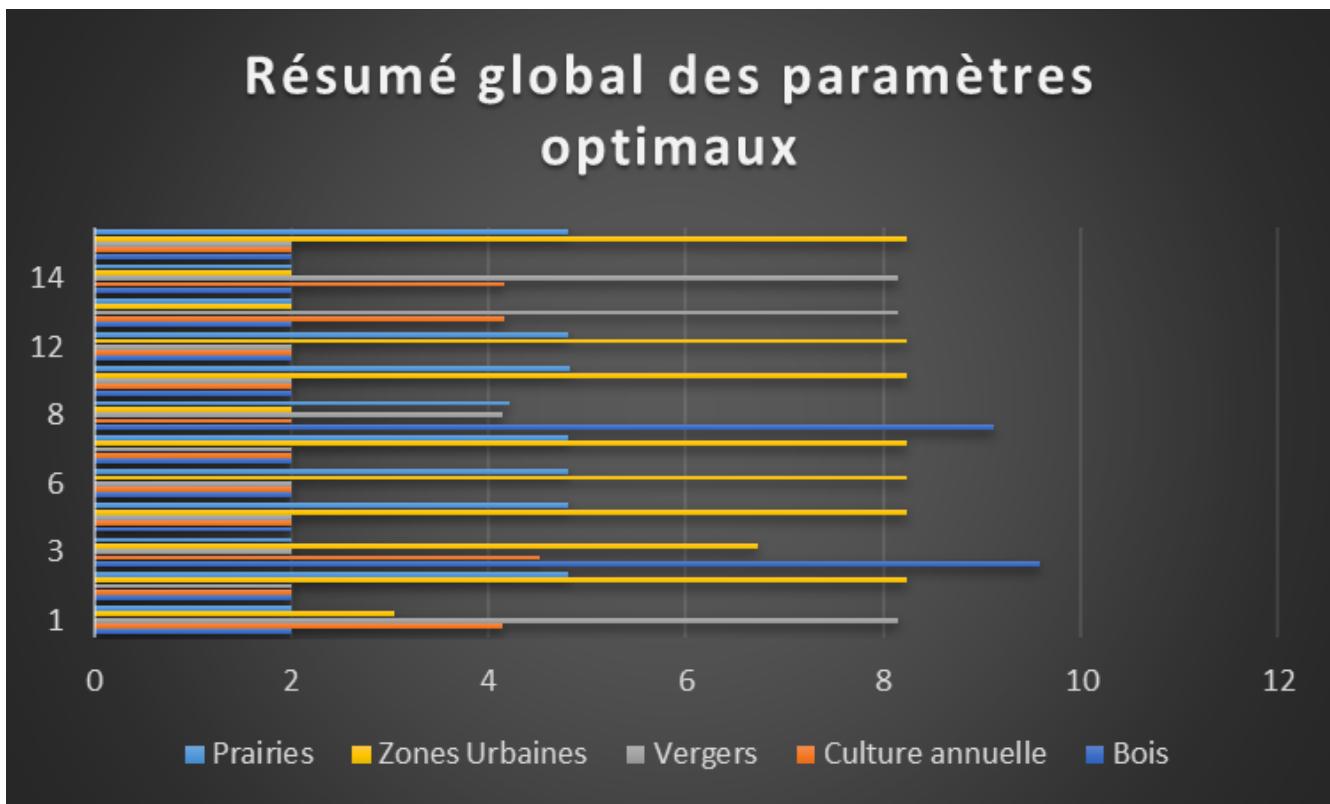
### Interprétation Mathématiques :

À partir de cette table on peut constater qu'il y a trois bassins de valeurs<sup>23</sup>

On peut les décomposer en deux familles, la première étant coloriée en jaune, présente une valeur maximale de la log-vraisemblance tourne autour de 53734, on peut voir que les paramètres de diffusions sont identiques.

Ensute on peut une seconde famille, qui présente les valeurs coloriées en nuances de bleu. Bien que la valeur maximale de la log-vraisemblance tourne autour de 53740, on peut la décomposer en deux sous famille dans la divergence se voit dans la zone urbaine.

### Interprétation biologique :



À partir de cela on peut dire que l'insecte étudié, le carpopophage, on peut dire que pour les différentes familles de valeurs obtenues :

- **Zones de bois :** Pour les différentes sous-familles de valeurs, on peut dire que la log-trajectoire =2 ce qui signifie que paramètre de diffusion est de 10 m qui est la valeur la plus basse possible. Ceci me pousse à dire que les carpopophage ne diffusent pas dans le bois, on peut ainsi interpréter cela par le fait que le bois constitue un habitat pour cet insecte où par le l'hostilité de cette zone : existence de prédateur, médicaments ...
- **Zones de cultures annuelles et prairies :** Dans ce cas, on constate deux types de valeurs de la log-trajectoire possibles, la première est de 2 pour le premier bassin de valeur( *Résultats en jaune*), puis tourne autour de 4.15 pour les deuxième et troisième bassins de valeurs et inversement pour

23. Voir couleurs.

les prairies. On peut dire que le coefficient de diffusion varie entre 10 m et 100 m. ce qui représente une petite zone d'étude, cela peut s'expliquer de la même manière que la première zone.

- **Vergers** : On s'aperçoit de l'existence de deux type de valeurs de la log-trajectoire possibles, la première est de 2 pour le premier bassin de valeur( *Résultats en jaune*), puis tourne autour de 8.14 pour les deuxième et troisième bassins de valeurs. Cela permet de dire que le coefficient de diffusion fluctue entre 10 m et 10 k.m.. On peut interpréter cela par le fait que dans notre zone d'étude, certains vergers ont été traités tandis que d'autre non, ce qui fait que dans certains endroit la diffusion est très grande.
- **Zones Urbaines** : On remarque que pour cette zone, il existe bassins de valeurs possibles qui correspondent à nos trois bassins de base :
  - ★ La log-trajectoire est de 8.1 pour le premier bassin de valeurs, ce qui permet de dire que dans cette partie de la zone d'étude la diffusion est de 10 K.M., une forte diffusion qui peut être interpréter par le fait que les insectes évitent cette zone.
  - ★ La log-trajectoire est de 2 pour le second bassin de valeurs tandis qu'elle est de 3.05 pour le troisième, on peut l'expliquer par le fait que certains particuliers possèdes des pommiers ou poiriers dans leurs jardin et que ces arbres jouent le rôle d'habitat pour le carpocapse à cause du manque de traitement.

## 4.2 Partie II : Calcul des paramètres de diffusion à partir de données simulées

**C**omme j'ai précisé durant la section précédente, j'ai effectué la simulation des fréquences génétiques et allélique à partir des densités. Ces simulations avaient pour dessein l'étude de l'influence nombre de vergers/individus sur les paramètres optimaux obtenus. Tout d'abord, j'ai choisi un coefficient de diffusion à  $D_s = 10^7$  et on essaye de le retrouver où plutôt d'approximer l'erreur des paramètres de diffusion obtenus. Malheureusement, les calculs des paramètres n'ont pas été effectué comme expliqué auparavant. En ce qui concerne les résultats obtenus, il s'agit des table de probabilité génétiques pour un nombre de vergers (2,4,8,16) et d'individus (2,4,8,16). On obtient des tableau avec des fréquences génétiques de moins en moins grandes.

### Exemple de table pour 8 vergers et 8 individus :

Données Corrigées :

	V1	V2	V3	V4	V5	V6	V7	V8
1	1.185926e-46	6.041144e-41	5.875618e-40	6.706207e-40	2.082045e-47	1.212887e-37	7.513324e-39	9.893769e-53
2	2.765098e-41	2.077058e-39	1.637576e-39	1.470729e-38	7.062759e-43	1.065735e-37	3.130038e-38	2.350213e-44
3	2.214564e-44	1.400628e-39	3.844777e-41	1.199109e-38	1.101410e-46	1.130620e-36	6.707222e-38	8.372792e-52
4	3.717747e-40	8.945183e-37	2.508228e-36	1.514189e-35	2.795032e-42	1.214700e-35	9.764186e-37	1.407498e-42
5	1.016384e-66	1.144611e-63	2.522762e-53	1.440752e-48	3.923006e-67	3.317548e-52	3.198936e-52	9.864978e-74
6	1.510702e-53	4.205249e-51	4.480011e-51	2.315297e-46	6.077938e-54	2.855826e-46	1.488864e-46	2.892356e-61
7	8.457861e-57	1.177977e-46	2.434614e-44	5.749991e-41	2.449395e-49	2.513401e-39	2.083058e-41	7.622066e-56
8	8.520565e-52	2.693556e-45	6.683590e-43	6.372251e-43	2.024153e-51	4.096001e-42	3.886128e-42	6.036797e-56

Données Corrigées avec notre bruitage :

	V1	V2	V3	V4	V5	V6	V7	V8
1	5.654282e-30	8.171900e-30	1.797656e-30	6.698470e-29	9.522730e-32	5.898252e-33	5.546597e-29	5.639774e-25
2	1.058037e-36	1.564743e-36	3.131305e-37	1.462846e-35	1.380491e-38	7.186098e-40	1.197093e-35	2.166782e-31
3	2.558435e-31	3.697600e-31	8.133987e-32	3.030906e-30	4.308820e-33	2.668826e-34	2.509710e-30	2.551870e-26
4	3.500960e-33	5.059793e-33	1.113054e-33	4.147489e-32	5.896186e-35	3.652019e-36	3.434285e-32	3.491977e-28
5	7.247292e-38	1.047421e-37	2.304119e-38	8.585663e-37	1.220562e-39	7.559997e-41	7.109268e-37	7.228696e-33
6	7.883622e-43	1.193066e-42	2.171931e-43	1.272107e-41	7.969072e-45	3.486301e-46	1.028800e-41	3.314901e-37
7	2.208594e-34	3.191989e-34	7.021743e-35	2.616459e-33	3.719630e-36	2.303890e-37	2.166531e-33	2.202927e-29
8	4.284922e-34	6.337020e-34	1.268141e-34	5.924352e-33	5.590825e-36	2.910284e-37	4.848084e-33	8.775207e-29

### Petite interprétation :

On peut observer l'influence du nombre de vergers et d'individus sur les résultats. Tout d'abord, le fait qu'on aura peu de données conduira à l'obtention d'une vraisemblance de plus loin de zéro et donc une valeur absolue de la log-vraisemblance de plus en plus petite et donc à des coefficients de diffusion  $D_{retrouv}$  assez loin de notre  $D_s$ .

Ainsi à partir de ces résultats on peut affirmer que le taux d'erreur tend vers zéros quand le nombre d'individus et de vergers est grand :

$$\lim_{\substack{nvergers \rightarrow 51 \\ nindiv \rightarrow 610}} \|D_{retrouv} - D_s\| = 0$$

Les calculs aurait put déterminer le taux d'erreur de nos résultats.

## 5 Conclusion :

- P**our conclure ce rapport, il serait judicieux d'effectuer en premier un bilan global du stage que j'ai effectué :
- ★ Premièrement, sur un plan personnel ce stage m'a permis de prendre contact avec le milieu de la recherche que j'ai pendant longtemps souhaité côtoyer, ce qui m'a conduit à découvrir le métier de chercheur, comment s'organise un laboratoire de recherche de l'envergure de l'I.N.R.A..
  - ★ Ensuite, sur un point de vu pratique, ce stage m'a permis d'élargir mes connaissances sur les langages R et FreeFem, mais aussi à me familiariser avec le système d'exploitation Linux. En plus de cela, j'ai acquis un bon nombre de connaissances sur la mise en œuvre numérique d'E.D.P., les méthodes statistiques appliquées à la biologie, la différence entre un modèle spatialisé et non spatialisé ainsi que quelques bases de la biologie.
- E**nfin, en ce qui concerne l'influence de ce stage sur mes perspectives d'avenir :
- ★ Cet expériences dans un laboratoire de recherche m'a permis de découvrir les différentes étapes de déroulement d'un projet de recherche scientifique, qui à mon opinion s'articule parfaitement avec la gestion de projet en entreprise ce qui constitue pour moi une bonne initiation dans le monde professionnel.
  - ★ Lors de mon séjour à l'unité Biosp, j'ai eu la chance de côtoyer plusieurs doctorants issus de diverse formation et spécialisés dans différents domaines mathématiques et écologiques lors des séminaires des doctorants ce qui m'a permis d'avoir une idée un peu plus vaste sur les différents domaines d'application de mon futur diplôme.

## Références

- [1] Cours génétique des populations de l'université en ligne.
- [2] <https://fr.wikipedia.org/wiki>.
- [3] <https://fr.wikipedia.org/wiki/syst>
- [4] <https://informatique-mia.inra.fr/biosp/accueil>.
- [5] <https://www6.paca.inra.fr/psh>.
- [6] N2qn1.
- [7] F. Hecht. New development in freefem++. *J. Numer. Math.*, 20(3-4) :251–265, 2012.
- [8] J-Chr.Culioli. *Introduction à l'Optimisation*.
- [9] P.Franck Olivier Bonnefon E.K.Klein J.Papaix, E.Walker. Estimation bayésienne de fréquences alléliques dans un modèle de métapopulation.
- [10] L.Roques. *Modèles de réaction-diffusion pour l'écologie spatiale*.
- [11] P.Franck S.Soubeyrand E.K.Klein L.Roques, E.Walker. Using genetic data to estimate diffusion rates in heterogeneous landscapes.
- [12] Boris Michenski. Rapport de stage de fin d'étude.
- [13] Mme Oudin. Cours d'optimisation continue.



*Mots clés :*

Dynamique des populations ; génétique des populations, équations aux dérivées partielles, analyse numérique (méthode des éléments finis) ; estimation de paramètres par maximum de vraisemblance, algorithme numérique d'optimisation (méthode quasi-Newton), simulation de paysage, lois de Mendel, loi de Hardy-Weinberg.