Diplôme d'Etudes Approfondies
de Biostatistique

Université de Montpellier II

« Base génétique de l'architecture de caféiers arabica – Exploration d'une base de données architecturales »

par

Guillaume Pérouel

Encadrement : **Christian Cilas, Christophe Godin, Yann Guédon**

CIRAD – UPR 31 – UMR DAP

Soutenu le 13 juin 2008, devant le jury composé de :

M. Abraham
M. Daures
M. Ducharme
M. Guédon

REMERCIEMENTS

Tout d'abord, je tiens à remercier mes tuteurs lors de ce stage, Messieurs Christian CILAS, Christophe GODIN et Yann GUEDON. En me permettant de travailler sur cette étude avec eux, j'ai acquis une expérience très instructive et enrichissante. Les discussions que nous avons eues ont été très utiles et intéressantes. Je les remercie d'avoir partagé leurs connaissances avec moi et de la confiance qu'ils m'ont montrée.

Mes remerciements vont également à toutes les personnes des équipes 31 et DAP pour leur disponibilité lors de mes défaillances ainsi que pour leur gentillesse, leur sourire et leur bonne humeur quotidienne.

En général, merci beaucoup pour la bonne atmosphère dans laquelle j'ai pu travailler pendant ces 5 mois.

Et enfin, je voudrais remercier ma famille (merci pitite sœur pour ton expérience en rapport de stage), mon amie et mes collègues : l'artiste, le ramasseur de melons, le petit, le véreux et le plus barbu des vietnamiens (ou l'inverse) pour leur soutien et leur confiance durant ces longues et interminables années d'étude (courage pour ceux qui n'ont 'toujours' pas fini !).

Table des matières

INTRODUCTION

Le CIRAD, Centre de coopération Internationale en Recherche Agronomique pour le Développement, est un organisme français de recherche agronomique au service du développement durable des pays du Sud et de l'outre-mer français. Les domaines de recherche du CIRAD sont aussi divers que les sciences du vivant et les sciences sociales appliquées à l'agriculture, la forêt, l'élevage, la gestion des ressources naturelles, l'agroalimentaire, aux écosystèmes et aux sociétés du Sud.

Au sein de l'UMR (Unité Mixte de Recherche) DAP (Développement et Amélioration des Plantes), les chercheurs travaillent sur des espèces tropicales et tempérées, comme le riz, le cacaoyer ou encore le caféier, étudié lors de ce stage. Leur but est d'étudier les caractères prioritaires pour la gestion des cultures et la création de nouvelles variétés.

Lors de ce stage, les études auront été faites en collaboration avec l'UPR (Unité Propre de Recherche) Maîtrise des bioagresseurs des cultures pérennes. Les travaux de cette unité sont centrés sur l'épidémiologie et la dynamique des populations de bioagresseurs. Ils permettent d'élaborer des modèles plantes-bioagresseurs pour les principaux organismes nuisibles, qui servent à définir des systèmes de gestion adaptés aux situations socio-économiques des producteurs.

Le contenu de ce stage porte sur des caféiers arabica du Costa-Rica, plantés et observés par Christian Cilas (UPR 31) et Christophe Godin (UMR DAP). A partir de plusieurs bases de données sur ces caféiers issus d'un plan diallèle de croisement, il s'agira d'extraire des paramètres pertinents permettant d'étudier l'héritabilité de plusieurs caractères d'architecture. Pour cela, il nous sera nécessaire d'utiliser le logiciel VPlants avec l'aide de Yann Guédon (UMR DAP), afin d'étudier et d'analyser statistiquement nos données, de construire des modèles statistiques et apporter une réponse à notre recherche d'héritabilité des caractères architecturaux des caféiers. Afin d'y parvenir, il nous sera nécessaire d'extraire des séquences de ces attributs le long des différents axes d'une plante.

La génétique quantitative a pour objet l'analyse des caractères à variation continue et à déterminisme complexe, c'est-à-dire gouvernés par plusieurs facteurs génétiques et plusieurs facteurs non-génétiques (environnementaux). Cette approche peut donc être appliquée à des variables quantitatives résumant l'architecture de nos caféiers.

Cependant, la pérennité des arbres engendre des difficultés d'ordre méthodologique, rendant ardu le phénotypage pour les caractères architecturaux. On appelle phénotype, l'ensemble des caractères observables d'un organisme dû aux caractères héréditaires, le génotype, et aux modifications apportées par le milieu environnant. De plus, à cause d'effets ontogéniques importants, l'ontogénie regroupant l'ensemble des processus conduisant un organisme végétal de la cellule œuf à un adulte reproducteur, les années successives ne peuvent pas être considérées comme des répétitions et ne peuvent donc être utilisées pour séparer les effets génotypiques et environnementaux (Costes *et al.*, 2003 ; Renton *et al.*,

2006). Ainsi, la méthodologie du phénotypage de plantes pérennes a également été étudiée, en prenant en compte d'une part des contraintes liées aux analyses génétiques et d'autre part les contraintes liées à la complexité du matériel et à l'analyse architecturale. Le modèle statistique utilisé mélange donc les facteurs génétiques et la structure génétique de nos plantes.

Nous allons maintenant voir comment on modélise classiquement l'action du génotype et de l'environnement sur le phénotype et ainsi introduire la notion d'héritabilité au sens large. Le phénotype est la manière dont un organisme nous apparaît et/ou peut être mesuré pour un niveau d'observation donné. Ainsi, on appelle valeur phénotypique le résultat de la mesure effectuée sur un individu (Gallais, 1989).

L'observation des répétitions d'un même génotype dans diverses conditions environnementales montre alors une variation des valeurs phénotypiques qui peut se décomposer en fonction donc du génotype et de l'environnement, sous la forme d'un modèle linéaire mixte (Gallais, 1989) :

$$P_{ij} = \mu + G_i + e_{ij},$$

avec,

- P_{ij}, la valeur phénotypique de la répétition j du génotype i,
- μ, la moyenne des valeurs phénotypiques de la population,
- G_i, la valeur du génotype i,
- e_{ij}, l'effet de l'environnement sur les répétitions j du génotype i.

Les valeurs phénotypiques et génotypiques sont alors considérées comme des variables aléatoires et l'effet de l'environnement comme un résidu aléatoire dont la distribution est la même pour tous les génotypes. Le modèle précédent peut alors être étendu à un modèle additif de variances (Gallais, 1989):

$$\sigma_P^2 = \sigma_G^2 + \sigma_e^2,$$

avec,

- σ_P^2, la variance des valeurs phénotypiques,
- σ_G^2, la variance des valeurs génotypiques, et correspond donc à la variance des effets aléatoires,
- σ_e^2, la variance des effets environnementaux, soit la variance résiduelle.

Sur la base de ce modèle, l'héritabilité au sens large a été introduite pour mesurer la part de variabilité phénotypique qui est d'origine génétique (Hanson, 1963 ; Comstock & Moll, 1963 ; Gallais, 1989) :

$$h^2 = \frac{\sigma_G^2}{\sigma_P^2}$$
$$= \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2}.$$

Le concept d'héritabilité au sens large peut donc être utilisé pour évaluer la part de variabilité observée pour un caractère quantitatif qui est d'origine génétique.

Cependant, dans le cadre de cette étude, le phénotype n'est plus résumé par de simples variables quantitatives mais traduit explicitement la structure de ramification des axes principaux de caféiers sous forme de séquences discrètes. De ce fait, on souhaite explorer les variations de ces structures de ramifications en fonction du génotype. Pour tout cela, on a alors recours à des modèles de types semi-markoviens cachés.

Pour expliquer la démarche statistique employée sur nos données, nous présenterons dans la première partie du rapport, les données ainsi que la procédure de modélisation de l'architecture des plantes utilisée. Ensuite, après avoir présenté les séquences sur lesquelles nous avons travaillé, nous étudierons la méthode d'analyse de ces séquences et notamment la modélisation à l'aide d'un modèle semi-markovien caché. Nous explorons également l'espace des séquences d'états possibles trouvées à partir de notre modèle. Toutes ces recherches nous mènerons à l'étude des séquences en fonction des parents dont nos arbres sont issus afin de conclure en la présence ou non d'un effet des géniteurs sur les attributs étudiés.

1ère partie :
Modélisation de l'architecture des arbres et procédure d'extraction de caractères.

I. Codage des plantes.

1- Présentation rapide de VPlants.

Afin d'analyser l'architecture et la croissance des plantes, on peut utiliser le logiciel VPlants (anciennement AMAPmod, Godin *et al.*, 1997) qui, à l'aide d'un ensemble de méthodes, permet la constitution et l'exploration de base de données d'architectures de plantes, ainsi que l'analyse statistique de ces données. A l'aide de différents modules présents dans le logiciel (MTG, STAT_TOOL, ...), il est alors possible de numériser les plantes et d'en faire une représentation virtuelle au moyen du module PLANTGL (Boudon *et al.*, 2001), de les décrire à différentes échelles (entre-nœuds, axes,...), d'analyser statistiquement les données et d'en construire des modèles statistiques. Tous ces logiciels sont libres et regroupés sur la plate-forme OpenAlea (Pradal & Dufour-Kowalski, 2007).

2- Fichier de codages des plantes.

Pour arriver à utiliser le logiciel, il est nécessaire de fournir une représentation formelle de nos plantes, sous la forme de graphes arborescents multi-échelle ou MTG (Multiscale Tree Graphs, Godin & Caraglio, 1998). Dans ces graphes sont représentés les différents organes de la plante, définis à plusieurs échelles, ainsi que leurs connections les uns aux autres, avec des relations de succession, symbolisé par '<', et de ramification, symbolisé par '+'.

Le fichier de codage associé aux plantes et décrivant donc leur MTG est composé de 2 parties :

- une entête qui contient les informations générales relatives à la description des plantes,
- la description des plantes.

a) Entête.

Cette partie regroupe toutes les classes d'entités utilisées dans le codage ainsi que les niveaux auxquelles elles appartiennent. L'utilisateur doit également préciser l'ensemble des attributs que l'on souhaite attacher aux entités et le type de valeur associé à chaque attribut (INT, REAL, ...).

Pour le cas des caféiers, les différents attributs qui ont été mesurés sont : le nombre de branches par entre-nœud le long de la tige, le nombre de feuilles et de rameaux sur les entre-nœuds des branches, la présence ou non de fruits sur ces mêmes entre-nœuds ainsi que sur les rameaux, la longueur de la tige et des branches et enfin l'évolution du diamètre le long de la tige. On peut ajouter a cela des attributs 'qualitatifs', comme le nom de la famille à laquelle l'arbre appartient, ainsi que les noms des parents, ou l'état éventuel d'une branche (exemple : morte).

CODE:	FORM-A			
CLASSES:				
SYMBOL	SCALE	DECOMPOS	INDEXATION	DEFINITION
$	0	FREE	FREE	IMPLICIT
P	1	CONNECTED	FREE	IMPLICIT
A	2	<-LINEAR	FREE	EXPLICIT
E	3	NONE	FREE	IMPLICIT
DESCRIPTION:				
LEFT	RIGHT	RELTYPE	MAX	
A	A	+	?	
E	E	+	1	
E	E	<	1	
FEATURES:				
NAME	TYPE			
NbFeuil	INT			
Longueur	INT			
MTG:				

Figure 1 : Exemple d'entête.

b) Description du MTG.

Le MTG décrit une plante sous la forme d'une arborescence que l'on représente à l'aide de colonnes dans lesquelles sont d'abord donnés les noms des entités et ensuite les éventuelles valeurs de chacun des attributs attachés à ces entités. La première colonne correspond à la plante tout entière. La seconde représente sa tige principale avec le détail de tous ces entre-nœuds. Ensuite la 3ème colonne correspond aux branches de l'arbre et enfin la dernière aux rameaux. Cette arborescence décrit en fait une succession de graphe réduit de composantes connexes, d'une échelle à l'autre.

Le label d'une entité (P1, A1, E0, ...), permettant de localiser de manière non ambigüe cette entité, est formé tout d'abord d'une lettre signifiant le type de l'entité : P pour plante, A pour axe (tige/branche/rameau) et E pour entre-nœud. L'indice correspond quand à lui, au rang de cette entité dans l'arborescence, en notant que la tige est parcouru en partant de l'entre-nœud le plus bas (E3 dans l'exemple qui suit Figure 3).

Le symbole '^' placé avant les E signifie que tous les entre-nœuds répertoriés après un A, forment l'axe en question. Les symboles '+' et '<' représentent respectivement, les relations de ramifications et de successions entre 2 entités.

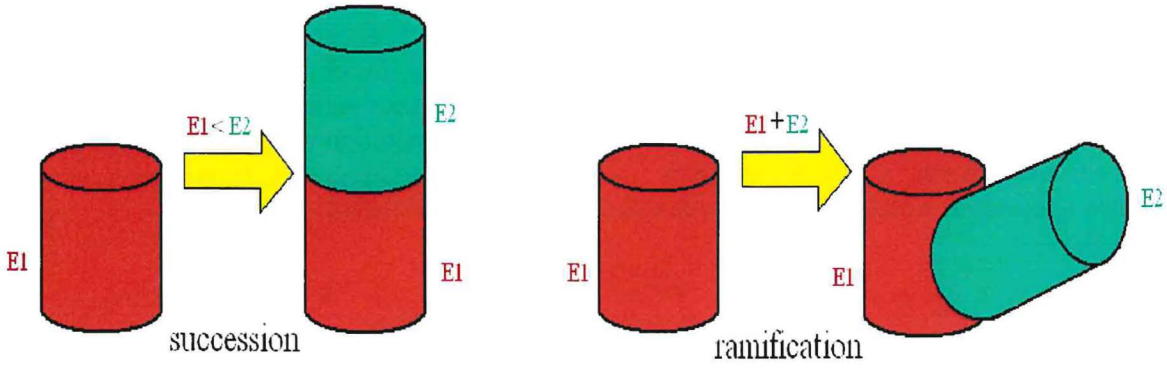
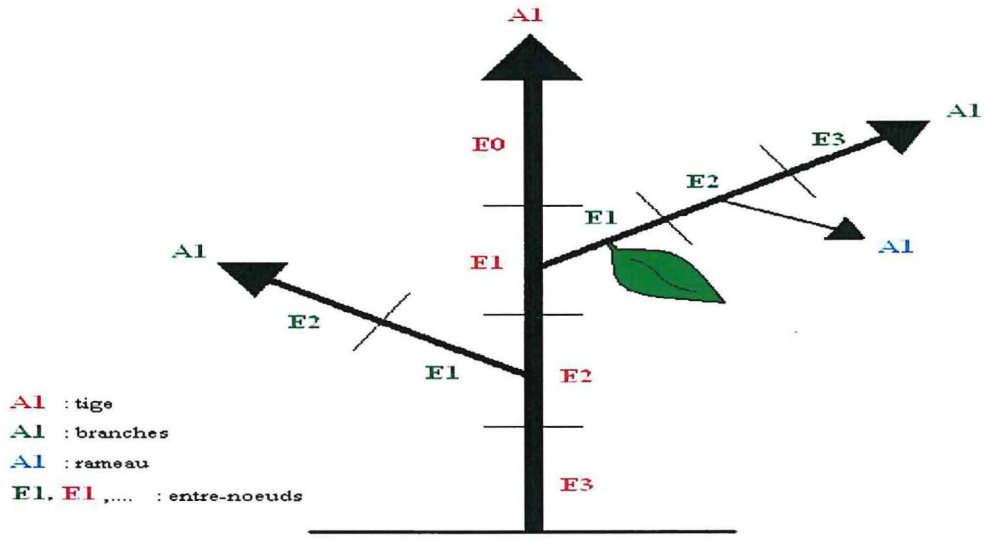


Figure 2 : Relation entre entités.



attributs

ENTITY-CODE					NbFeuil	Longueur
/P1						
	/A1					45
	^/E3					
	^<E2					
		+A1				10
		^/E1				
		^<E2				
	^<E1					
		+A1			1	20
		^/E1			1	
		^<E2				
			+A1			5
		^<E3				
	^<E0					

valeurs des attibuts

tige de la plante

branches de la plante

rameaux de la plante

Figure 3 : Exemple de MTG et son codage.

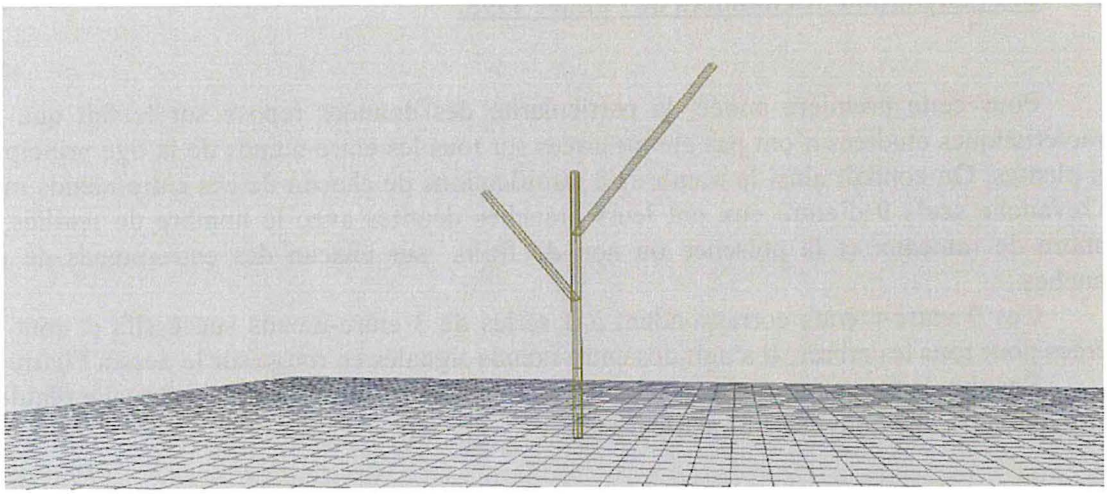


Figure 4 : Représentation virtuelle de la plante avec le package PlantGL 3D Viewer.

II. Présentation des données.

Il s'agit d'une base de données sur l'architecture de caféiers arabica du Costa-Rica. L'étude porte sur un échantillon de 513 arbres qui ont été mesurés sur 3 années successives (de 1998 à 2000), avec une variation du nombre de plantes étudiées puisque 306 l'ont été en 1999 et 346 la dernière année. Le but de ce projet est d'extraire des paramètres pertinents permettant d'étudier l'héritabilité de plusieurs caractères d'architectures.

De ce fait, sur chacune des plantes étudiées, on a extrait différents attributs, comme nous l'avons vu précédemment.

De plus, les arbres sont issus d'un plan diallèle de croisements. Ce plan est basé sur un échantillon de 10 variétés différentes. Nous avons donc des plein-frères, issus du même croisement, des demi-frères, issus d'un même parent mais dont le second parent est différent, et des individus n'ayant aucun lien de parenté. Ainsi, il a pu être créé 45 familles différentes.

ANNEE 1998		PERE										TOTAL
		1	2	3	4	5	6	7	8	9	10	
MERE	1	t8666 => 4		a9 => 12	a23 => 21	a19 => 10	a4 => 7	a12 => 12	a17 => 11	a2 => 7		84
	2		t18121 => 12	b20 => 8	b19 => 14	b8 => 15	b16 => 12	b4 => 11	b11 => 12	b1 => 7	b24 => 12	103
	3			t18140 => 11	e13 => 15	e22 => 12	e27 => 12	e6 => 13	e7 => 12	e3 => 18		93
	4				t18138 => 12	d19 => 8	d23 => 12	d9 => 8	d12 => 12		p2 => 20	72
	5					t18130 => 10	c16 => 17	c7 => 11	c11 => 8		c22 => 12	58
	6						t18141 => 11	f5 => 12	f11 => 10		f15 => 12	45
	7							t17933 => 8	x5 => 11			19
	8								t17930 => 7		t7 => 12	19
	9									t17931 => 12		12
	10										t8667 => 8	8
TOTAL		4	12	31	62	55	71	75	83	44	76	513

Figure 5 : Tableau de croisement des plantes pour l'année 1998.

1- Présentation des données de l'année 1998.

Pour cette première année, la particularité des données repose sur le fait que les caractéristiques étudiées n'ont pas été mesurées sur tous les entre-nœuds de la tige principale des plantes. On connaît ainsi le nombre de ramifications de chacun de ces entre-nœuds mais en revanche seuls 9 d'entre eux ont leurs branches décrites avec le nombre de feuilles, le nombre de rameaux et la présence ou non de fruits sur chacun des entre-nœuds de ces branches.

Ces 9 entre-nœuds correspondent à 3 séries de 3 entre-nœuds successifs et sont les mêmes pour tous les arbres. Il s'agit des entre-nœuds signalés en rouge sur le dessin Figure 6.

Ensuite, la longueur et le diamètre ne sont également mesurés qu'à intervalle régulier, comme montré une nouvelle fois sur le dessin.

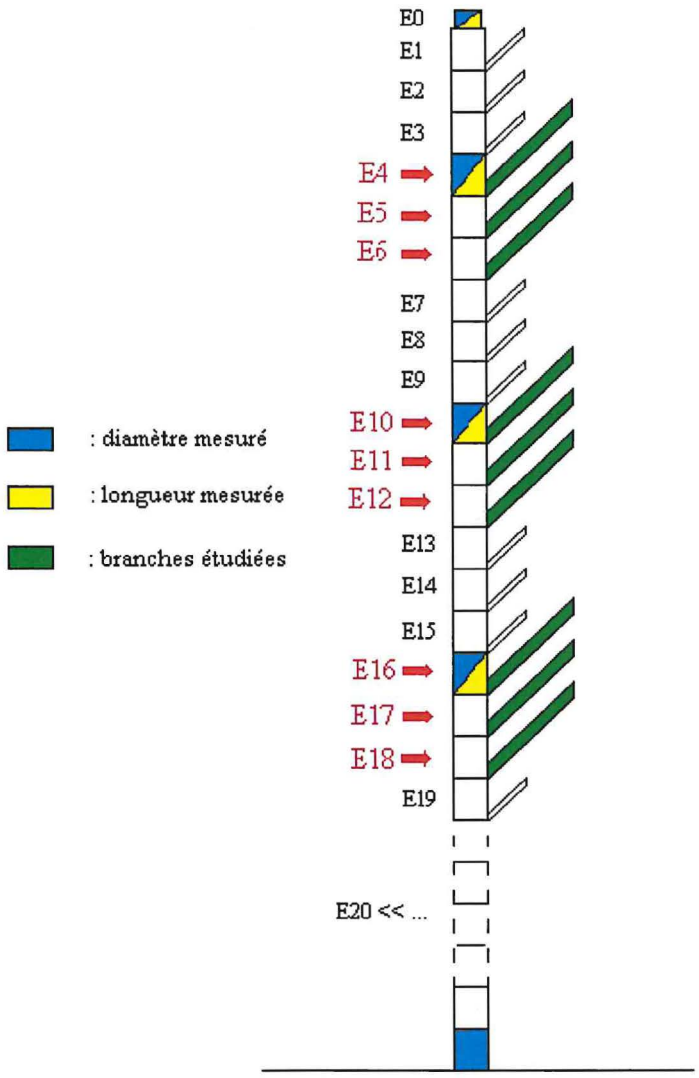


Figure 6 : Schéma d'une plante avec les différents entre-nœuds étudiés en 1998.

2- Présentation des données de l'année 1999.

En ce qui concerne l'année 1999, les mesures ont été similaires à celle de 1998. En effet, les ramifications ont été comptabilisées sur tous les entre-nœuds de la tige et les branches ont été mesurées sur les entre-nœuds situés exactement aux mêmes endroits sur la tige qu'en 1998. Du coup, nous sommes donc à une distance différente de l'apex en croissance. Il ne s'agit donc pas des mêmes entre-nœuds que ceux étudiés l'année précédente. Il s'agira alors d'une comparaison, d'une année sur l'autre, de caractéristiques situées au même endroit de l'arbre et donc des mêmes entités de la plante.

3- Présentation des données de l'année 2000.

Enfin, pour cette année 2000, les seules branches qui ont été étudiées en ce qui concerne le nombre de feuilles et la présence ou non de fruits sur leurs entre-nœuds, sont les branches portées par les entre-nœuds E4, E5 et E6, comme représentés sur le dessin comparatif des mesures effectuées sur les 3 années. Les études faites lors de cette 3ème année auront donc pour but de faire une comparaison entre ces entre-nœuds et les derniers mesurés lors de l'année 1998, qui sont bien entendu situés plus bas, du fait de la croissance de la plante durant 2 années.

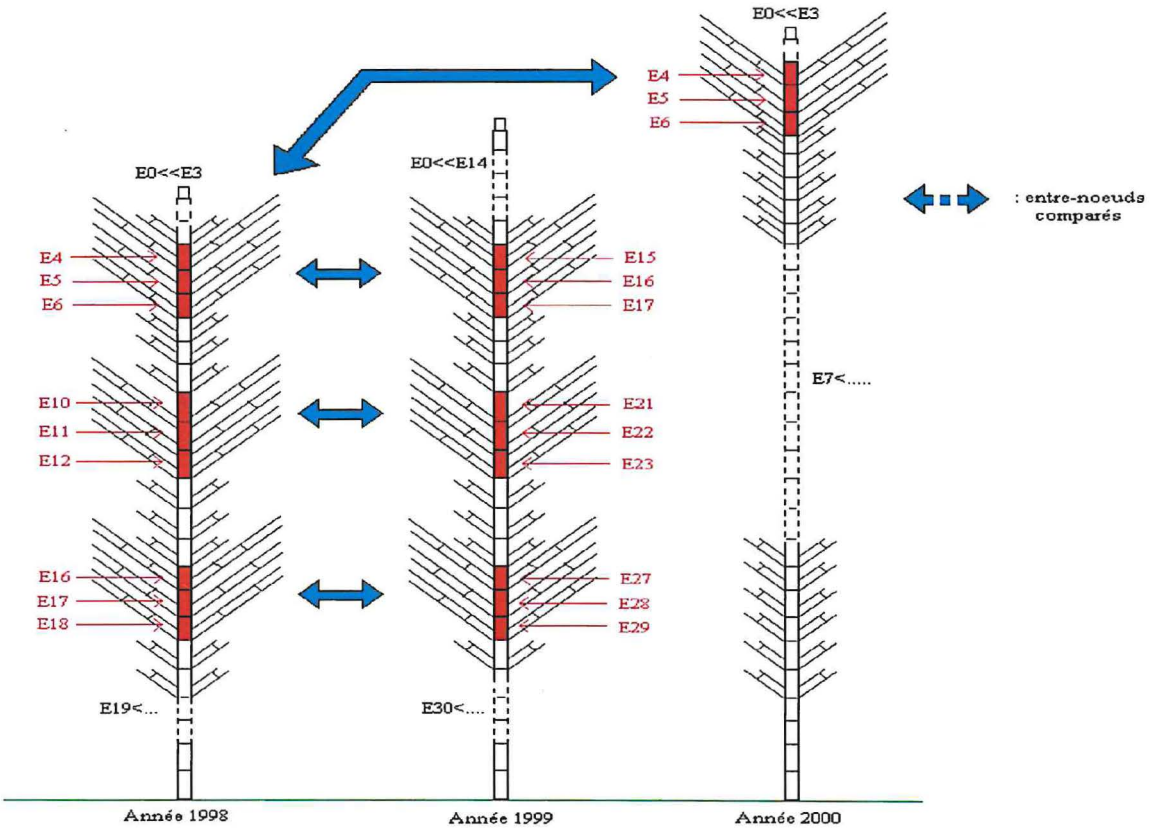


Figure 7 : Comparatif des mesures sur les 3 années.

III .Programme d'extraction.

1- Utilisation du langage Python.

Lors d'une étude précédente, menée sur des caféiers robusta de Côte d'Ivoire, un programme d'extraction de caractéristiques de ces arbres avait été écrit par Christophe Godin à l'aide du langage informatique AML (AMAP Modelling Language). Il s'agit d'un langage fonctionnel développé avec le logiciel AMAPmod.

Dans le cas de ce projet, le programme d'extraction a été écrit en Python (Van Rossum, 1989), simple d'utilisation, car très intuitif, très complet avec des interfaces graphiques, des tests et de la documentation sur ses composants.

2- Programme.

Dans cette étude, la première partie de mon travail a consister à l'écriture de programmes afin d'extraire des séquences de nos différents attributs le long des divers axes de la plante. Il peut s'agir par exemple, de la séquence du nombre de branches sur chacun des entre-nœuds de la tige de la plante, ou encore la séquence binaire de la présence ou non de fruits sur les entre-nœuds d'une branche.

Séquences															
Entre-noeuds de la tige		E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	.....	BASE
	nombre de branches	0	0	0	1	2	2	2	2	1	2	2	2	2	.....	0

Séquences													
Entre-noeuds d'une branche		E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
	présence ou non de fruits	0	1	1	1	1	1	1	0	1	1	1	1	0

Figure 8 : Exemples de séquences extraites des données de 1998.

2ème partie :

Exploration des séquences.

Séquences du nombre de branches par entre-nœuds de la tige.

I .Présentation des séquences.

1- Années 1998 et 1999 avec proposition de segmentation des séquences.

Ces séquences sont organisées en une succession de 0, de 1 et de 2 car, pour les caféiers, un entre-nœud ne peut porter qu'un maximum de 2 branches.

Il faut ensuite essayer de trouver une segmentation correcte de ces séquences afin d'identifier des modèles statistiques appropriés.

En observant les données de 1998 et 1999, on peut penser à une partition en 4 zones :

- une première ne comprenant que des 0 et se trouvant à la base des plantes,
- une seconde avec un mélange de 0, de 1 et de 2,
- une troisième avec une longue série de 2, et quelques 0 et/ou des 1,
- et une dernière avec un état final 0, correspondant à l'entre-nœud pas encore ramifié situé au sommet de la plante.

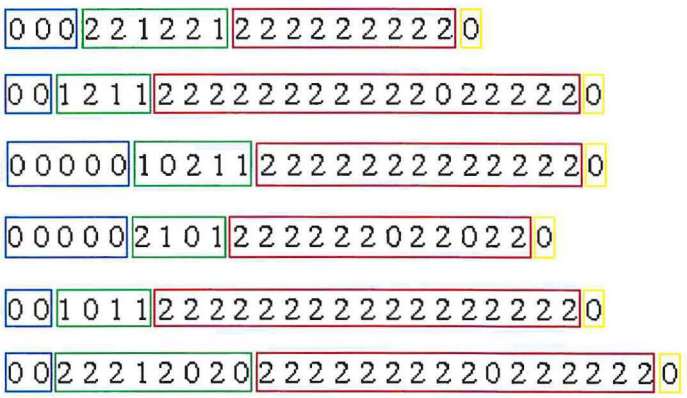


Figure 9 : Exemples de séquences extraites des données en 1998, avec une proposition de segmentation empirique.

Cette proposition de segmentation semble correcte mais reste néanmoins délicate. En effet, ce qui correspond à la seconde zone est difficile à bien cerner car il existe une incertitude sur les limites de ce segment car une observation, par exemple 2, peut être observée dans des zones adjacentes. On aura donc des difficultés à interpréter cette partie de la séquence.

2- Année 2000 avec proposition de segmentation des séquences.

En revanche, pour l'année 2000, on a plus de mal à proposer une segmentation correcte. Les séquences obtenues cette année là sont en effet beaucoup plus longues que les années précédentes. On peut y voir cependant la même configuration des séquences au début, à savoir, une série de 0 suivie d'une série mélangeant des 0, des 1 et des 2. On retrouve également l'observation finale 0. Par contre, pour ce qui est du reste de la séquence, c'est-à-dire la longue, voir très longue, série de 2 entrecoupée de quelques 0 et 1, elle est plus difficile à modéliser car cette zone est plus hétérogène dans sa composition avec plus ou moins de 0 et de 1. En effet, en comparaison à 1998, certaines branches ont pu mourir par exemple, ce qui peut expliquer la présence davantage de 0 et de 1 dans cette dernière zone. Cependant, nous allons continuer à proposer cette segmentation.

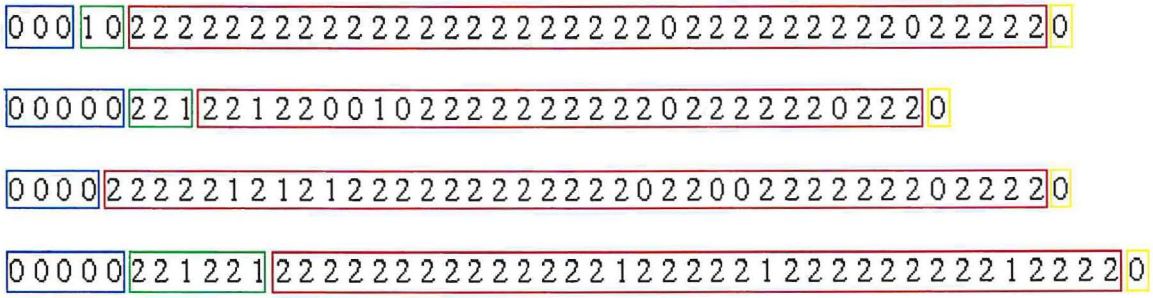


Figure 10 : Exemples de séquences extraites des données en 2000, avec une proposition de segmentation empirique.

II. Méthode d'analyse des séquences.

1- Utilisation d'une semi-chaîne de Markov cachée.

a) Choix d'une semi-chaîne de Markov.

Nous allons utiliser des modèles probabilistes pour l'analyse de nos séquences afin de modéliser au mieux la succession de zones obtenue. Nous sommes donc confrontés à une succession de zones homogènes, caractérisées par une certaine composition 'intra-zone' de 0, de 1 et de 2 et possédant une structure cachée, car une observation peut être observée dans différentes zones. La seule connaissance de l'observation ne permet donc pas d'identifier le segment dans lequel on se trouve.

Dans notre cas, nous utiliserons des semi-chaînes de Markov cachées. En effet, une chaîne de Markov cachée permet de modéliser une séquence en la segmentant en un nombre fini de zones qui se succèdent. On a recours à des modèles markoviens cachés car on ignore a priori, les longueurs et les positions de ces différentes régions. Les chaînes de Markov cachées ont été décrites pour la première fois dans une série de publication statistique par Leonard E. Baum et d'autres auteurs après 1965. Elles ont été ensuite appliquées à divers domaines tels que la reconnaissance de la parole (Rabiner, 1989), la modélisation du langage (Gauvain *et al.*, 1994), la reconnaissance de caractères manuscrits (Anigbogu, 1992), le traitement du signal (Saerens, 1993) ou l'analyse des images (Devijver & Dekesel, 1988 ; Mao & Kung, 1990). Dans la seconde moitié des années 1980, les chaînes de Markov cachées ont commencé à être appliquées à l'analyse biologique, en particulier à l'ADN, notamment dans la détection et la prédiction de gènes, la recherche de motifs, ... (Durbin *et al.*, 1998 ; Baldi & Brunak, 1998).

De plus, ces modèles seront semi-markoviens cachés car ainsi on pourra précisément prendre en compte les lois des longueurs de chacune de ces zones. On aura alors autant d'états dans notre semi-chaîne de Markov cachée que de zones.

b) Définition d'une semi-chaîne de Markov cachée.

Une chaîne de Markov peut être vue comme un ensemble d'états entre lesquels s'effectuent des transitions. Cette définition est cependant beaucoup trop simpliste pour pouvoir modéliser nos séquences. De plus, les modèles markoviens sont basés sur l'hypothèse d'homogénéité de la séquence, avec donc une probabilité similaire d'observer une observation tout au long de la séquence. Dans notre cas, nous sommes confrontés à des séquences formées de plusieurs zones aux caractéristiques différentes. Il nous faudrait donc utiliser un modèle capable de segmenter une séquence en un nombre fini de régions homogènes pouvant avoir des différences de structures, ou de composition. C'est pour cela que nous privilégions dans un premier temps les chaînes de Markov cachés (Rabiner, 1989), qui permettent de modéliser une séquence par un ensemble fini de modèles qui se succèdent le long de la séquence. Ils sont d'ailleurs très largement utilisés et notamment pour la recherche de régions homogènes par Churchill (1989, 1992), puis Muri (1997, 1998) et Boys *et al.* (2000).

Notre chaîne de Markov, cachée puisque comme nous l'avons vu précédemment, nous ignorons la longueur et l'emplacement de nos zones, modélisera donc la succession des zones homogènes le long de nos séquences. Elle se caractérise alors par 2 processus stochastiques $\{S_t, X_t; t = 0,1,...\}$ tels que :

- $\{S_t\}$ soit le processus caché, appelé processus d'état, et qui est une chaîne de Markov d'ordre 1, d'espace d'états fini $\{0,1,...,J-1\}$,
- $\{X_t\}$ soit le processus observé, appelé processus d'observation ou d'émission.

Ainsi, à chaque position t de la séquence X_t, sera associé un état caché S_t. Cela fonctionne comme si le processus d'observation était lié au processus caché par une fonction probabiliste f avec $X_t = f(S_t)$. Cette fonction est supposée telle qu'une observation puisse être émise dans différents états.

Le couple de processus $\{S_t, X_t; t = 0,1,...\}$ est une chaîne de Markov d'ordre 1 cachée si la relation de dépendance suivante est vérifiée :

$$\begin{aligned} & P\left(X_t = x_t, S_t = s_t \mid X_0^{t-1} = x_0^{t-1}, S_0^{t-1} = s_0^{t-1}\right) \\ & = P\left(X_t = x_t, S_t = s_t \mid S_{t-1} = s_{t-1}\right) \\ & = P\left(X_t = x_t \mid S_t = s_t\right) P\left(S_t = s_t \mid S_{t-1} = s_{t-1}\right) \end{aligned}$$

Ensuite, on peut définir notre chaîne de Markov cachée par un ensemble de paramètres.

Nous avons tout d'abord les paramètres d'une chaîne de Markov d'ordre 1 à J états:

- les probabilités initiales :

$$\pi_j = P\left(S_0 = j\right), \qquad j = 0,1,...,J-1, \qquad \text{avec} \quad \sum_{j-0}^{J-1} \pi_j = 1,$$

- les probabilités de transition:

$$p_{ij} = P\left(S_t = j \mid S_{t-1} = i\right), \qquad \forall t, \quad i, j = 0,1,...,J-1, \qquad \text{avec} \quad \sum_{j-0}^{J-1} p_{ij} = 1,$$

Auxquels on ajoute les probabilités d'observation, qui mettent en relation le processus $\{X_t\}$ avec la chaîne de Markov $\{S_t\}$:

$$b_j(y) = P\left(X_t = y \mid S_t = j\right), \qquad j = 0,1,...,J-1 \ ; \ y = 0,1,...,Y-1,$$
$$\text{avec} \quad \sum_{\mathrm{j}} b_j(y) = 1.$$

Un ordre 1 signifie que tout le passé du processus est résumé dans l'état précédent.

Cependant, la modélisation par chaîne de Markov cachée d'une séquence implique que la loi du temps de séjour dans un état est une loi géométrique de paramètre $(1 - p_{ii})$. Cette hypothèse est souvent très irréaliste pour modéliser des longueurs de zones. Il serait plus réaliste d'utiliser une loi qui possède une densité avec une forme de cloche avec une traîne à droite. De ce fait, on va souhaiter imposer une loi particulière, adaptée à notre problème, pour les temps de séjour de chaque état caché visité. C'est pour cela que nous allons finalement utiliser les modèles de semi-chaîne de Markov cachée qui permettent précisément de modéliser explicitement les longueurs de chaque zone.

Une semi-chaîne de Markov cachée est construite à partir d'une chaîne de Markov cachée sous-jacente. Nous allons donc maintenant définir les différents paramètres de cette chaîne de Markov cachée :

- des probabilités initiales :

$$\pi_j = P(S_0 = j), \qquad j = 0,1,...,J-1, \qquad \text{avec} \quad \sum_j \pi_j = 1,$$

- des probabilités de transition :

.pour un état i non absorbant:

$$p_{ij} = P(S_t = j \mid S_t \neq i, S_{t-1} = i), \qquad \forall j \neq i \quad \text{avec} \quad \sum_{j \neq i} p_{ij} = 1,$$

$$p_{ii} = 0,$$

.pour le cas particulier d'un état i absorbant :

$$p_{ii} = P(S_t = i \mid S_{t-1} = i) = 1,$$
$$p_{ij} = 0, \qquad \forall j \neq i,$$

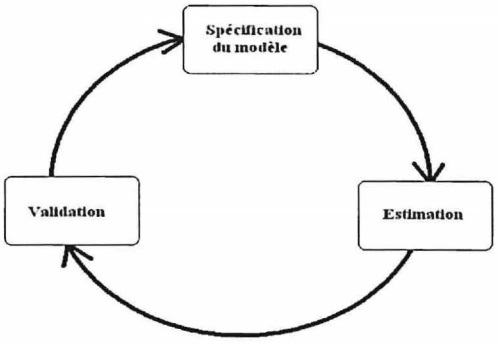
- des lois d'occupation, ou temps de séjour dans un état, attachées à chaque état non-absorbant de la chaîne de Markov:

$$d_j(u) = P(S_{t+u+1} \neq j, S_{t+u+v} = j; v = 0,...,u-2 \mid S_{t+1} = j, S_t \neq j), \qquad u = 1, 2, \ldots$$

- des lois d'observation, identiques à celles du modèle markovien caché vu précédemment.

2- Démarche de modélisation statistique.

La démarche de modélisation statistique se décompose en 3 étapes que l'on itère jusqu'à obtenir un modèle correctement ajusté aux données. Dans un premier temps, il faut spécifier le modèle que l'on va utiliser. Un modèle peut être vu comme une formalisation mathématique supposée reproduire de manière approchée la réalité d'un phénomène dans le but d'en reproduire le fonctionnement. Un modèle est défini par des paramètres qu'il est nécessaire d'estimer, ce qui correspond à la seconde étape de notre démarche. Et enfin, on valide ou non le modèle estimé. Si ce n'est pas le cas, on recommence la démarche jusqu'à ce que le modèle nous paraisse correct.



a) Spécification des modèles.

Pour les modèles markoviens, la spécification du modèle se réduit au choix de la famille de modèles candidats, par exemple une chaîne de Markov d'ordre élevé ou une semi-chaîne de Markov. En revanche, pour les modèles markoviens cachés, cela consiste principalement à dimensionner la chaîne de Markov sous-jacente et à faire des hypothèses sur les propriétés structurelles sur la base des distributions caractéristiques extraites des séquences.

Afin de modéliser nos séquences, différents modèles ont été testés. Tout d'abord, pour l'année 1998, on peut proposer 2 modèles de semi-chaîne de Markov cachée. Ces modèles respectent la segmentation adoptée. On en propose alors un premier comprenant 4 états correspondant à nos 4 zones. Cependant, comme nous l'avons vu précédemment, les limites d'une zone ne sont souvent pas simples à définir. Nous l'avons déjà évoqué pour celles du second segment. De plus, nous souhaitons arriver à isoler la zone apicale non ramifiée des plantes. De ce fait, on peut proposer un second modèle avec un état final artificiel sur lequel on impose une contrainte. Cette contrainte est que l'on ne peut atteindre cet état de fin fictif que depuis l'état 3 correspondant donc à cet entre-nœud pas encore ramifié situé au sommet de la plante. Ce procédé d'ajout d'un état final, qui est une méthode très utilisé avec les modèles markoviens, permet ainsi de bien modéliser les longueurs de nos régions (Durbin *et al.* 1998).

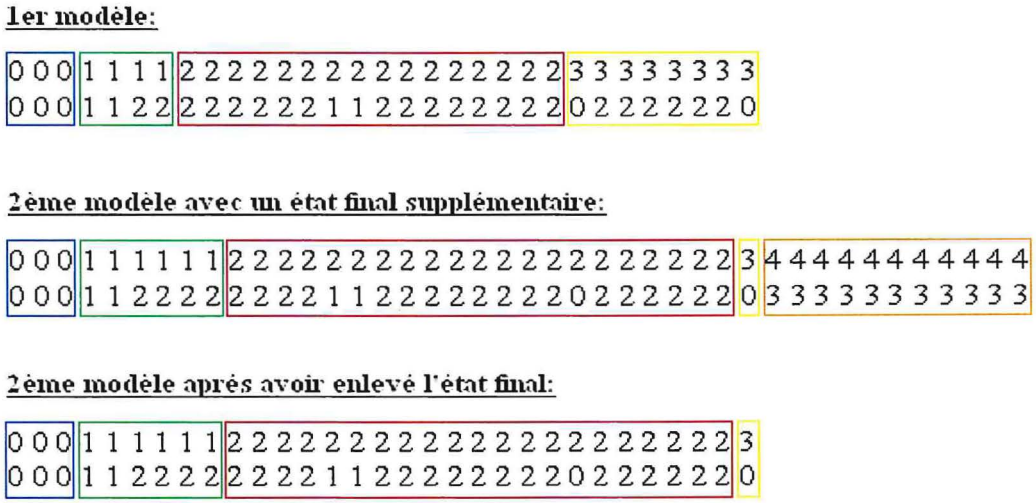


Figure 11 : Comparaison des différents modèles sur une même séquence extraite en 1998.

Sur cette figure sont représentées, la séquence des observations sur la seconde ligne et la modélisation proposée de cette séquence selon le modèle adopté sur la première ligne.

On voit bien sur ce comparatif des 2 modèles proposés, qu'il y a une différence flagrante dans la délimitation des zones. En effet, avec la semi-chaîne de Markov cachée avec une spécification usuelle, la fin de la séquence ne coïncide pas avec la sortie du dernier état visité. On est donc en présence de censure à droite du dernier temps de séjour. En revanche, pour le modèle avec un état final artificiel, la fin de la séquence coïncide bien avec la sortie du dernier état visité. Il n'y a donc aucune censure. On utilisera donc ce second modèle pour estimer nos données.

Ensuite, pour l'année 1999, on pourra comparer les 2 mêmes modèles, avec toujours une préférence pour le second.

En revanche pour l'année 2000, il est nécessaire de créer 2 nouveaux modèles. Nous gardons la même démarche, avec un second modèle calqué sur le premier, mais auquel on ajoute un état fictif afin de bien isolé l'état final de nos séquences observées dans un état de notre semi-chaîne de Markov cachée. Néanmoins, on va redimensionner ces modèles. En effet, pour modéliser la longue série de 2 qui, comme nous l'avons vu dans la présentation des séquences, est beaucoup plus hétérogènes, on fait s'opposer en parallèle 2 états. La différence entre ce qui sera donc un état 2 et un état 3, se trouve dans le fait que pour l'un d'eux, l'état 2 en l'occurrence, on retrouvera peu de 0 et/ou de 1 insérés parmi la série de 2, alors que pour l'autre état, on en aura davantage. De ce fait, on se retrouve donc avec un état supplémentaire par rapport aux autres modèles, l'état final devenant donc l'état 4. Par conséquent, on ne pourra jamais rencontrer à la fois un état 2 et un état 3 dans la segmentation de nos séquences.

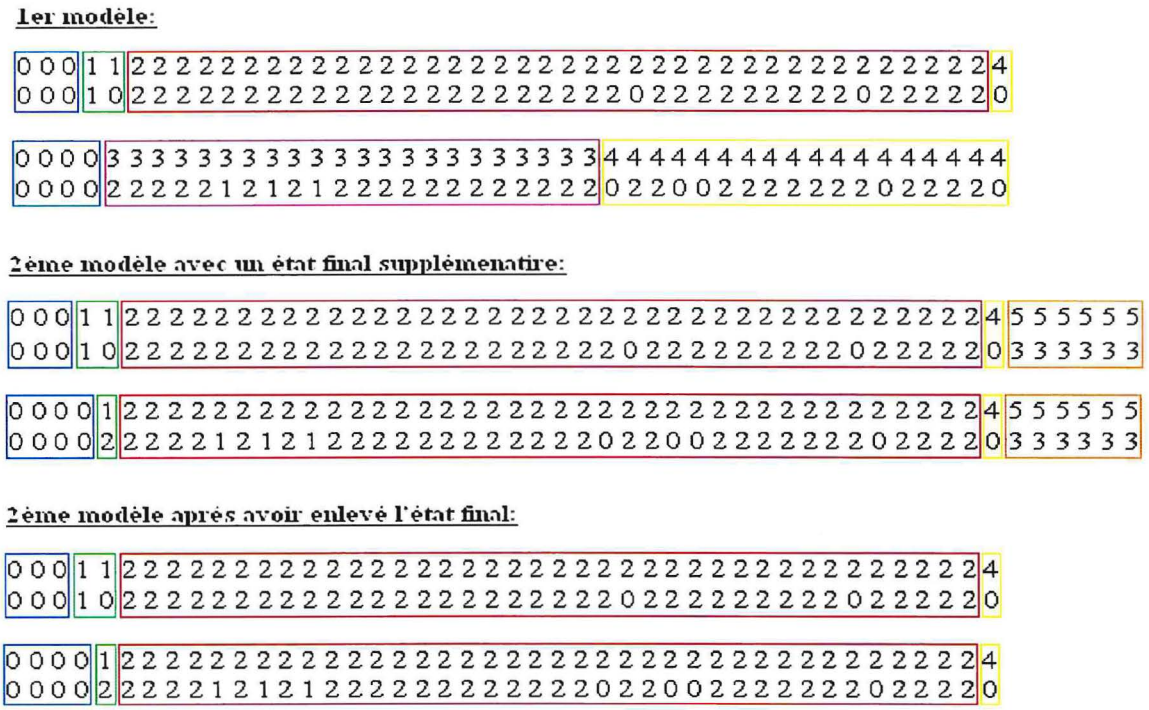


Figure 12 : Comparaison des différents modèles sur 2 séquences extraites en 2000.

Une nouvelle fois, on voit bien l'importance du second modèle qui parvient à isoler en un seul état l'entre-nœud apical des plantes. Notons simplement que l'utilisation de ce modèle peut entrainer une différence au niveau de l'état dans lequel on entrera parmi les 2 mis en parallèle.

b) Estimation – Algorithme EM.

L'algorithme EM (Expectation-Maximisation) est une méthodologie statistique qui s'applique principalement à l'estimation par maximum de vraisemblance dans le cadre de problèmes aux données incomplètes. Cet algorithme s'applique donc parfaitement aux semi-chaînes de Markov cachées puisque les états ne sont pas observables.

Cet algorithme alterne des étapes d'évaluation de l'espérance, l'étape E, et une étape de maximisation, l'étape M (Baum *et al.*, 1970 ; Dempster *et al.* 1977 ; McLachlan & Krishnan, 2008).

L'étape E consiste à estimer la log-vraisemblance des données complètes en calculant l'espérance de cette log-vraisemblance à l'étape k conditionnellement à ce qui est observé et que l'on note $Q\left(\theta\middle|\theta^{(k)}\right)$, avec θ le paramètre à estimer.

Calcul de $Q\left(\theta\middle|\theta^{(k)}\right)$:

$$Q\left(\theta\middle|\theta^{(k)}\right)=E\left\{\log f\left(S_0^{\tau-1},X_0^{\tau-1};\theta\right)\middle|X_0^{\tau-1}=x_0^{\tau-1};\theta^{(k)}\right\}.$$

avec la vraisemblance des donnés complètes qui s'écrit de la manière suivante :

$$f\left(s_0^{\tau-1},x_0^{\tau-1};\theta\right)=P\left(S_0^{\tau-1}=s_0^{\tau-1},X_0^{\tau-1}=x_0^{\tau-1};\theta\right).$$

L'étape M permet d'estimer le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée lors de l'étape précédente. En fait, la prochaine valeur de θ, $\theta^{(k+1)}$, est choisie telle que, $\forall\theta\in\Theta$, Θ étant l'espace des paramètres :

$$Q\left(\theta^{(k+1)}\middle|\theta^{(k)}\right)\geq Q\left(\theta\middle|\theta^{k}\right),$$

ce qui peut encore se dire :

$$Q^{(k+1)}=\arg\max_{\theta}\left\{Q\left(\theta\middle|\theta^{(k)}\right)\right\}.$$

c) Validation.

La précision du modèle estimé est principalement évaluée par l'ajustement des distributions caractéristiques calculées à partir des paramètres du modèle sur les distributions caractéristiques empiriques correspondantes, extraites des séquences observées. Parmi tous les critères possibles pour effectuer la sélection du modèle, nous pouvons citer par exemple le BIC (Bayesian Information Criterion, Schwarz, 1978 ; Katz, 1981)

Cependant, ces critères de sélection de modèle ne s'appliquent que si la chaîne de Markov sous-jacente est ergodique, c'est-a-dire constituée d'une seule classe récurrente et apériodique. De ce fait, on ne peut pas les utiliser pour nos modèles car il s'agit ici de modèles 'gauche-droite', c'est-à-dire formés d'une succession d'états transitoires et d'un état final absorbant.

3- Présentation des modèles.

Comme nous l'avons vu précédemment, pour créer une semi-chaîne de Markov cachée, il est nécessaire de spécifier les probabilités initiales de ses états, les probabilités de transition entre états, les lois d'occupation de chacun d'eux et enfin les lois d'observations des observations au sein d'un état. Tous ces paramètres correspondent à des contraintes structurelles sur la semi-chaîne de Markov cachée et se définissent de la manière suivante:

- Le nombre d'états, qui correspond au nombre de zones dans nos séquences observées.

- Les probabilités initiales, sur lesquelles on fait l'hypothèse d'une loi uniforme. Par commodité, on fait l'hypothèse de lois proches de lois uniformes, dans le sens où l'on n'attribue pas précisément, dans le cas de 4 états, une probabilité de 1/4 à chacune d'entre elles.

- La matrice de transitions avec les probabilités de transition entre états. Cette dernière est triangulaire supérieure avec une diagonale principale nulle, excepté pour l'état absorbant p_{33} dans les cas 1998 et 1999, et p_{44} pour 2000. En effet, comme nous l'avons vu précédemment, nous sommes dans une configuration 'gauche-droite' avec des états transitoires qui se succèdent, d'où la triangulaire supérieure, et un état final absorbant, ce qui justifie la structure de la diagonale principale. Pour les deuxièmes modèles, on impose également d'avoir la dernière colonne avec des valeurs nulles sauf pour p_{J-1J} et p_{JJ}, ce qui force le passage par l'état correspondant au dernier entre-nœud de la tige, pour entrer dans l'état final. Enfin, sur une ligne, on fait de nouveau une hypothèse de lois proches de lois uniformes.

- Les lois d'occupation de chacun des états, pour lesquelles on attribue généralement une loi binomiale négative $NB(1,1,p)$, avec p d'une valeur raisonnable par rapport à nos données, et qui correspond à une reparamétrisation sous forme de semi-chaîne de Markov cachée d'une simple chaîne de Markov cachée.

- Les lois d'observation dans chacun des états. On pose là encore des lois proches de lois uniformes entre les différentes observations, exception faîte lorsque l'on désire obtenir un état formé d'une seule observation, comme c'est le cas pour le dernier d'entre eux, et dans ce cas là, on pose une probabilité de 1 à la seule observation souhaitée.

Dans les figures 13 et 14 qui suivent, sont présentés les 4 modèles évoqués précédemment, ceux utilisés en 1998 et en 1999 dans la Figure 13 et ceux de 2000 avec la Figure 14. A chaque fois, le modèle situé à droite correspond à celui qui fait intervenir un état fictif supplémentaire afin de forcer le passage dans l'état final absorbant.

```
HIDDEN_SEMI-MARKOV_CHAIN

4 STATES

INITIAL_PROBABILITIES
0.3 0.3 0.2 0.2

TRANSITION_PROBABILITIES
0.0 0.4 0.3 0.3
0.0 0.0 0.5 0.5
0.0 0.0 0.0 1.0
0.0 0.0 0.0 1.0

STATE 0 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 1 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 2 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1


1 OUTPUT_PROCESS

OUTPUT_PROCESS 1 : NONPARAMETRIC

STATE 0 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.4
OUTPUT 1 : 0.3
OUTPUT 2 : 0.3

STATE 1 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.4
OUTPUT 2 : 0.3

STATE 2 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.3
OUTPUT 2 : 0.4

STATE 3 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 1.0
```

```
HIDDEN_SEMI-MARKOV_CHAIN

5 STATES

INITIAL_PROBABILITIES
0.3 0.3 0.2 0.2 0.0

TRANSITION_PROBABILITIES
0.0 0.4 0.3 0.3 0.0
0.0 0.0 0.5 0.5 0.0
0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 1.0

STATE 0 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 1 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 2 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.05


1 OUTPUT_PROCESS

OUTPUT_PROCESS 1 : NONPARAMETRIC

STATE 0 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.4
OUTPUT 1 : 0.3
OUTPUT 2 : 0.3

STATE 1 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.4
OUTPUT 2 : 0.3

STATE 2 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.3
OUTPUT 2 : 0.4

STATE 3 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 1.0

STATE 4 OBSERVATION_DISTRIBUTION
OUTPUT 3 : 1.0
```

Figure 13 : Modèles utilisés pour les 2 années de 1998 et 1999.

```
HIDDEN_SEMI-MARKOV_CHAIN

5 STATES

INITIAL_PROBABILITIES
0.2 0.2 0.2 0.2 0.2

TRANSITION_PROBABILITIES
0.0 0.3 0.3 0.2 0.2
0.0 0.0 0.4 0.3 0.3
0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 1.0

STATE 0 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 1 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 2 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 3 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

1 OUTPUT_PROCESS

OUTPUT_PROCESS 1 : NONPARAMETRIC

STATE 0 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.4
OUTPUT 1 : 0.3
OUTPUT 2 : 0.3

STATE 1 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.4
OUTPUT 2 : 0.3

STATE 2 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.3
OUTPUT 2 : 0.4

STATE 3 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.4
OUTPUT 2 : 0.3

STATE 4 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 1.0
```

```
HIDDEN_SEMI-MARKOV_CHAIN

6 STATES

INITIAL_PROBABILITIES
0.2 0.2 0.2 0.2 0.2 0.0

TRANSITION_PROBABILITIES
0.0 0.3 0.3 0.2 0.2 0.0
0.0 0.0 0.4 0.3 0.3 0.0
0.0 0.0 0.0 0.5 0.5 0.0
0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.0 1.0

STATE 0 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 1 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.1

STATE 2 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.05

STATE 3 OCCUPANCY_DISTRIBUTION
NEGATIVE_BINOMIAL    INF_BOUND : 1
PARAMETER : 1    PROBABILITY : 0.05

1 OUTPUT_PROCESS

OUTPUT_PROCESS 1 : NONPARAMETRIC

STATE 0 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.4
OUTPUT 1 : 0.3
OUTPUT 2 : 0.3

STATE 1 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.4
OUTPUT 2 : 0.3

STATE 2 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.3
OUTPUT 1 : 0.3
OUTPUT 2 : 0.4

STATE 3 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 0.4
OUTPUT 1 : 0.3
OUTPUT 2 : 0.3

STATE 4 OBSERVATION_DISTRIBUTION
OUTPUT 0 : 1.0

STATE 5 OBSERVATION_DISTRIBUTION
OUTPUT 3 : 1.0
```

Figure 14 : Modèles utilisés pour l'année 2000.

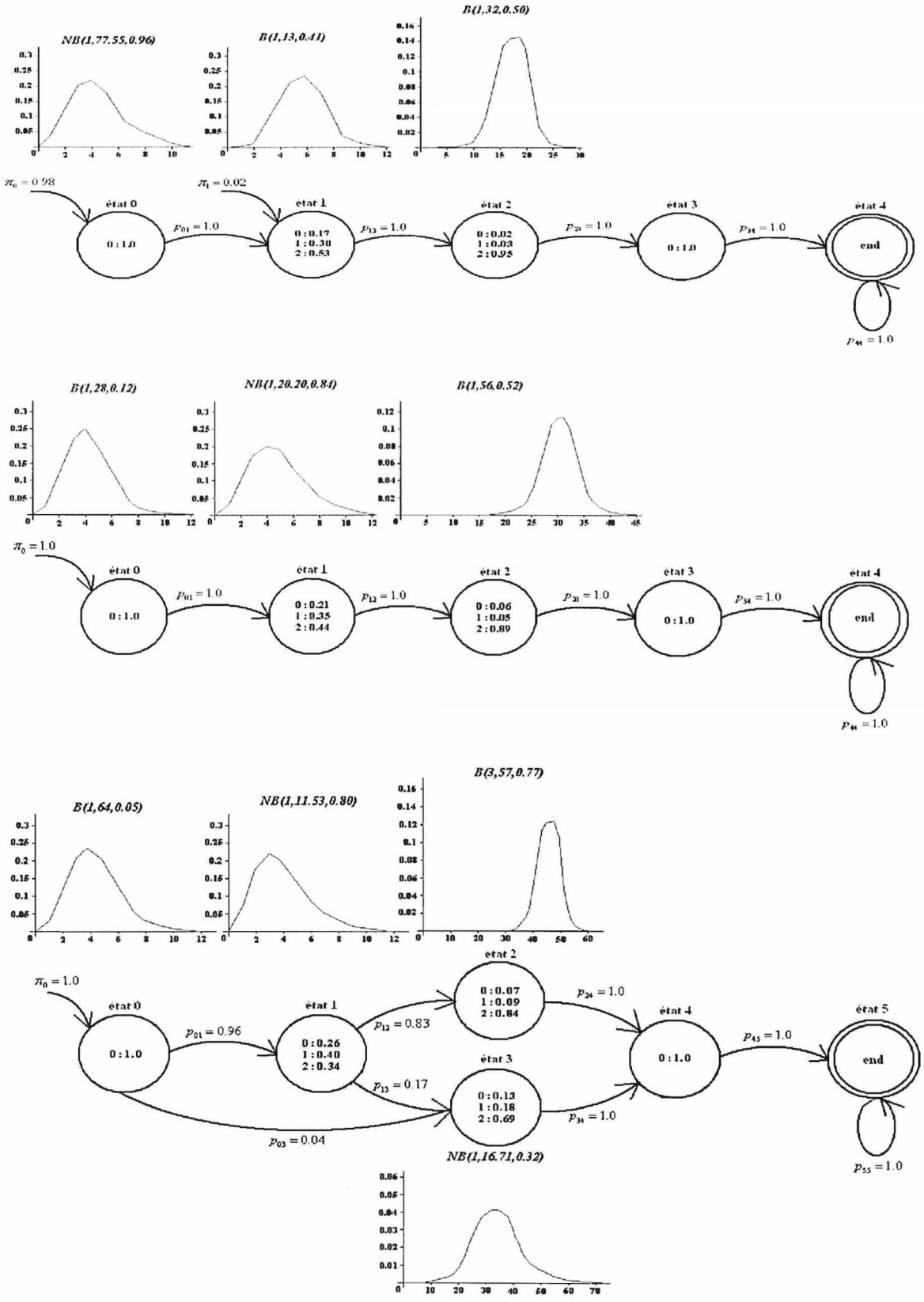


Figure 15 : Représentation des modèles avec un état supplémentaire, estimés sur les données de 1998 (graphe 1), 1999 (graphe 2) et 2000 (graphe 3).

Ces modèles à convergence de l'algorithme EM sont obtenus après environ 50 itérations. Les paramètres estimés sont représentés sur ces 3 graphiques. Les cercles représentent les états de notre modèle avec, à l'intérieur, les lois d'observations arrondies à $10^{\wedge -2}$, au-dessus, on retrouve les lois d'occupation, ainsi que les probabilités de transition p_{ij} et les probabilités initiales π_j supérieures à 0.01. On reconnaît bien la structure 'gauche-droite' des modèles, puisqu'à l'exception de l'année 2000 et ses 2 états en parallèle, on a $p_{ii+1} = 1$ pour tous les états i transitoires. Ces modèles estimés montrent également que les seules probabilités initiales non nulles sont celles qui permettent d'entrer dans l'état 0. On note aussi que les lois d'occupation des états sont loin d'être des lois géométriques. De plus, on peut remarquer que la dispersion de ces lois d'occupation est relativement faible. On peut en conclure que les zones sont bien définies.

4- Qualité des modèles.

Il est nécessaire de vérifier la qualité des modèles utilisés sur nos séquences. Pour cela, nous pouvons regarder l'ajustement des lois caractéristiques, comme les probabilités d'observation des différents états, les temps de retour ainsi que les temps de séjour dans chacun d'eux, calculées à partir des paramètres du modèle utilisé, sur les caractéristiques correspondantes extraites des séquences observées.

Nous nous limiterons ici aux probabilités des observations ajustées sur nos données, en utilisant, comme nous l'avons vu précédemment, les 2 modèles faisant intervenir un état fictif supplémentaire.

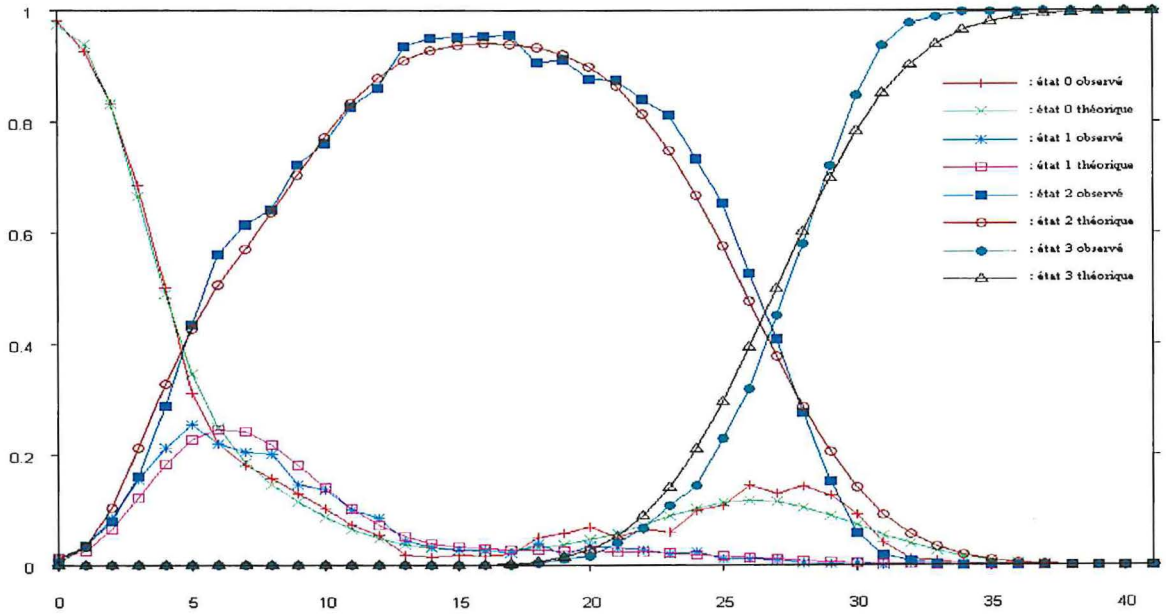


Figure 16 : Probabilités des observations ajustées sur les données de 1998.

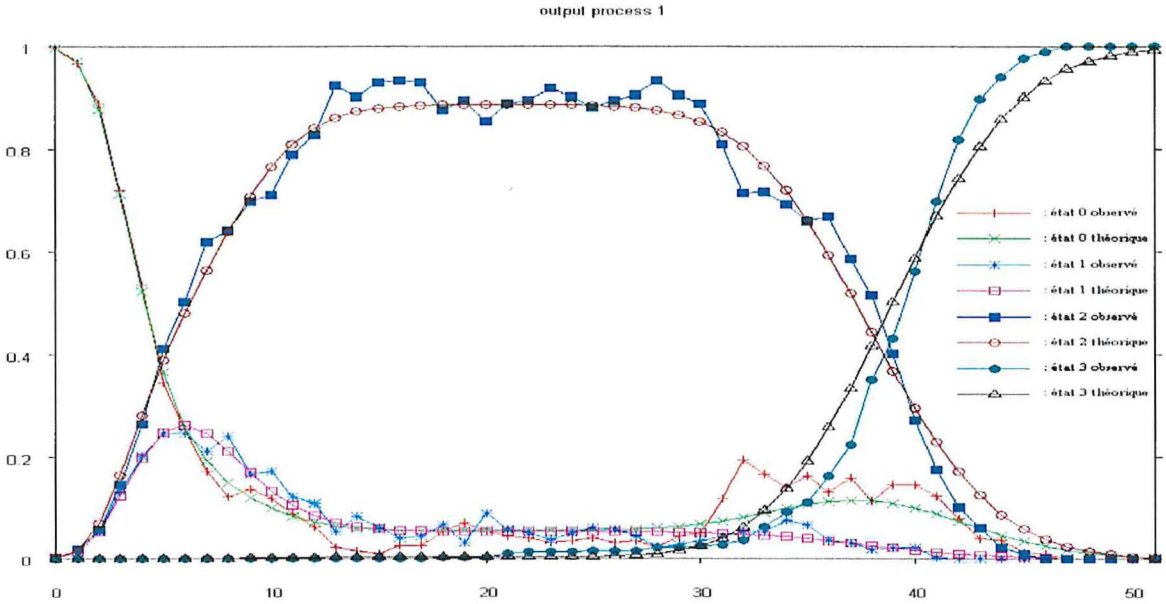


Figure 17 : Probabilités des observations ajustées sur les données de 1999.

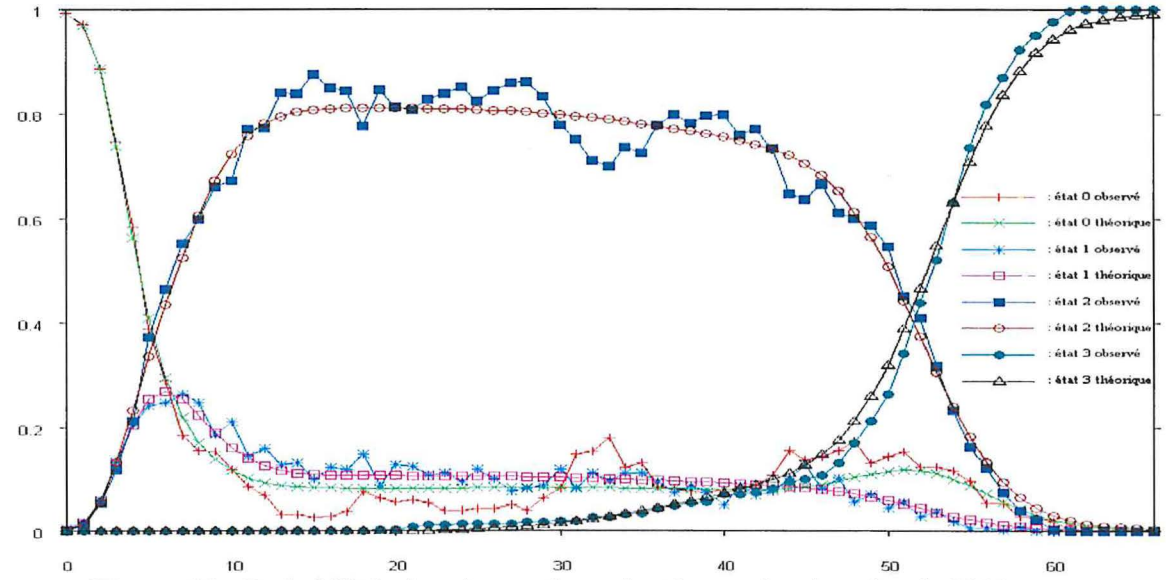


Figure 18 : Probabilités des observations ajustées sur les données de 2000.

On voit clairement, à l'aide de ces 3 graphiques, la qualité des modèles utilisés. En effet, les lois théoriques sont très proches de nos observations et ceci témoigne donc que les modèles sont bien ajustés aux données.

De plus, avec ces graphiques, on distingue nettement la succession des différentes zones homogènes de nos séquences observées. En effet, on voit bien qu'au début des séquences, on ne rencontrera que l'observation 0. Ensuite, en parallèle avec la diminution de la proportion de 0, on a une augmentation des proportions de 1 et de 2 qui nous mène au second état, formé du mélange des observations 0, 1 et 2. Suit ensuite une longue zone avec une très grande proportion de 2 et quelques 0 et 1. Et enfin, alors que toutes ce proportions diminuent, on voit celle de l'observation 3 augmenter progressivement, jusqu'à ne pouvoir observer plus que lui.

III. Méthode d'exploration de l'espace des séquences d'états possibles.

Les méthodes d'exploration de l'espace des séquences d'états possibles se décomposent en 3 catégories. Nous n'évoquerons ici que 2 d'entre elles :

- l'énumération des séquences d'états les plus probables, qui nécessite les algorithmes de Viterbi et de Viterbi généralisé,
- le calcul des profils d'états résumant les séquences d'états, qui nécessite les algorithmes 'Avant-Arrière' (ou 'Forward-Backward') et de Viterbi 'Avant-Arrière'.

1- Calcul des séquences d'états les plus probables.

a) Algorithme de Viterbi pour trouver la séquence d'états la plus probable.

Dans diverses situations, il est intéressant de connaître la séquence d'états la plus probable, c'est-à-dire celle qui explique au mieux la séquence observée pour un modèle donné. Ceci est réalisé par un algorithme de programmation dynamique, appelé algorithme de Viterbi (Viterbi, 1967). La programmation dynamique est une méthode de résolution de problème d'optimisation qui repose sur une propriété de décomposition de la fonction à optimiser.

Cet algorithme est utilisé dans le cas où le processus sous-jacent est modélisable comme une semi-chaîne de Markov discrète à états finis. Le problème est alors, étant donnée une séquence d'observation $x_0^{\tau-1}$, de trouver la séquence d'états $s_0^{\tau-1}$ pour laquelle la probabilité a posteriori $P\left(S_0^{\tau-1}=s_0^{\tau-1}\middle|X_0^{\tau-1}=x_0^{\tau-1}\right)$ est maximale. L'algorithme de Viterbi donne la solution de ce problème d'estimation par maximum a posteriori.

Comme le processus d'états $\{S_t\}_t$ est une semi-chaîne de Markov, on a la décomposition suivante :

$$\max_{s_0,\ldots,s_{\tau-1};s_{t+1}\neq s_t} P\left(S_0^{\tau-1}=s_0^{\tau-1},X_0^{\tau-1}=x_0^{\tau-1}\right)$$
$$=\max_{s_t}\left\{\max_{s_0,\ldots,s_{t-1}} P\left(S_0^{t}=s_0^{t},X_0^{t}=x_0^{t}\right)\times \max_{s_{t+1},\ldots,s_{\tau-1}} P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1},S_{t+1}^{\tau-1}=s_{t+1}^{\tau-1}\mid S_{t+1}\neq s_t,S_t=s_t\right)\right\}.$$

Soit la quantité suivante:

$$\alpha_j(t)=\max_{s_0,\ldots,s_{t-1}} P\left(S_{t+1}\neq j,S_t=j,S_0^{t-1}=s_0^{t-1},X_0^{t}=x_0^{t}\right),$$

La décomposition précédente s'écrit alors:

$$\max_{s_0,\ldots,s_{\tau-1};s_{t+1}\neq s_t} P\left(S_0^{\tau-1}=s_0^{\tau-1},X_0^{\tau-1}=x_0^{\tau-1}\right)$$
$$=\max_j\left[\max_{s_{t+1},\ldots,s_{\tau-1}} P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1},S_{t+1}^{\tau-1}=s_{t+1}^{\tau-1}\mid S_{t+1}\neq j,S_t=j\right)\alpha_j(t)\right].$$

L'algorithme de Viterbi, basé sur les quantités $\alpha_j(t)$, s'écrit alors de la façon suivante,
$t=0,\ldots,\tau-2 \quad ; \quad j=0,\ldots,J-1$:

$$\alpha_j(t)=\max_{s_0,\ldots,s_{t-1}} P\left(S_{t+1}\neq j,S_t=j,S_0^{t-1}=s_0^{t-1},X_0^t=x_0^t\right)$$
$$=b_j(x_t)\max\left[\max_{1\le u\le t}\left[\left\{\prod_{v=1}^{u-1}b_j(x_{t-v})\right\}d_j(u)\max_{i\neq j}\left\{p_{ij}\alpha_i(t-u)\right\}\right],\right.$$
$$\left.\left\{\prod_{v=1}^{t}b_j(x_{t-v})\right\}d_j(t+1)\pi_j\right].$$

La censure à droite du temps de séjour dans le dernier état visité rend le cas $t=\tau-1$ particulier. On a alors,
$j=0,\ldots,J-1$:

$$\alpha_j(\tau-1)=\max_{s_0,\ldots,s_{\tau-1}} P\left(S_{\tau-1}=j,S_0^{\tau-2}=s_0^{\tau-2},X_0^{\tau-1}=x_0^{\tau-1}\right)$$
$$=b_j(x_{\tau-1})\max\left[\max_{1\le u\le\tau-1}\left[\left\{\prod_{v=1}^{u-1}b_j(x_{\tau-1-v})\right\}D_j(u)\max_{i\neq j}\left\{p_{ij}\alpha_i(\tau-1-u)\right\}\right],\right.$$
$$\left.\left\{\prod_{v=1}^{\tau-1}b_j(x_{\tau-1-v})\right\}D_j(\tau)\pi_j\right].$$

avec $D_j(u)=\sum_{v\ge u}d_j(v)$, la fonction de survie du temps de séjour dans l'état j.

Les sous-séquences d'états optimales s_0^t se terminant dans un état donné se déduisent ainsi des $t\times J$ sous-séquences d'états optimales s_0^{t-u} se terminant dans les différents états calculées aux étapes précédentes.

L'état précédent optimal est donné par :

$$\psi_j(t)=\arg\max_i\left\{p_{ij}\alpha_i(t-u)\right\}.$$

La probabilité de la séquence d'états optimale associée à la séquence observée $x_0^{\tau-1}$ est égale à :

$$\max_j\left\{\alpha_j(\tau-1)\right\},$$

et l'état final optimal est donné par :

$$\tilde{s}_{\tau-1} = \arg\max_j \{\alpha_j(\tau-1)\}.$$

La séquence d'états optimale est alors extraite par une procédure de chaînage arrière :

$$\tilde{s}_t = \psi_{\tilde{s}_{t+1}}(t+u).$$

On arrive ainsi à calculer la probabilité maximale de générer la séquence $x_0^{\tau-1}$, à l'aide d'une équation de récurrence. Cet algorithme de Viterbi nous permet donc de trouver la séquence qui maximise la probabilité a posteriori de la manière suivante :

$$\max_{s_0,\ldots,s_{\tau-1}} P\left(S_0^{\tau-1} = s_0^{\tau-1} \mid X_0^{\tau-1} = x_0^{\tau-1}\right) = \frac{\max_{s_0,\ldots,s_{\tau-1}} P\left(S_0^{\tau-1} = s_0^{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1}\right)}{P\left(X_0^{\tau-1} = x_0^{\tau-1}\right)}$$
$$= \frac{\max_j \{\alpha_j(\tau-1)\}}{P\left(X_0^{\tau-1} = x_0^{\tau-1}\right)},$$

avec $P\left(X_0^{\tau-1} = x_0^{\tau-1}\right)$ qui se calcule à l'aide de la récurrence Avant de l'algorithme 'Avant-Arrière' que nous verrons plus tard.

Terminons la présentation de cet algorithme en précisant sa complexité. Elle est de $O\left(J\tau\left(J+\tau\right)\right)$ en temps pour le pire des cas et de $O\left(J\tau\right)$ en espace.

b) Algorithme de Viterbi généralisé pour trouver les L séquences d'états les plus probables.

Il est également possible de trouver les L séquences d'états les plus probables pour une séquence observée. Pour cela, on utilise un nouvel algorithme, l'algorithme de Viterbi généralisé. Ce dernier permet de calculer les L séquences d'états les plus probables dans l'ordre décroissant de leur probabilité a posteriori, puisque comme nous venons de le voir, la séquence d'états la plus probable sera celle qui maximisera la probabilité a posteriori.

Dans le cas d'une chaîne de Markov cachée (Foreman, 1993), l'optimalité de cet algorithme repose sur le fait que les L séquences d'états partielles les plus probables soient dans l'état j au temps t, nécessitant au maximum le calcul des L séquences dans chaque état au temps $t-1$, la situation extrême étant le cas où les L séquences d'états partielles les plus probables étant dans l'état j au temps t, sont construites à partir des L séquences d'états partielles les plus probables étant dans un état i donné au temps $t-1$. Ce principe se transpose directement au cas semi-markovien (Guédon, 2007) où les L séquences d'états partielles les plus probables quittant l'état j au temps t, nécessite au maximum le calcul des L séquences quittant chaque état aux temps $t-u$ pour $u=1,\ldots,t$. La principale différence entre cet algorithme et l'algorithme de Viterbi présenté précédemment, est la gestion des rangs des séquences d'états partielles aux temps $t-u$.

Nous noterons L_t le nombre de séquences d'états partielles au temps t. Ce nombre augmente à chaque pas de temps jusqu'à ce qu'il dépasse L, auquel cas, il reste fixé à L.

Pour un état j, la récurrence Avant est la même que celle de l'algorithme de Viterbi, à l'exception que $\alpha_i^{(t-u)}$ est remplacé par $\alpha_i^{r(t-u,i)}$, pour $t-u=0,...,\tau-1$, où pour chaque temps t, les rangs des séquences d'états partielles $\left(r(t-u,j);u=1,...,t;j=0,...,J-1\right)$ sont initialisés à 1.

Cette quantité $\alpha_j^n(t)$ est la probabilité de la séquence partielle observée x_0^t conjointement avec la $n^{ième}$ séquence d'état partielle quittant l'état j au temps t.

L'état précédent optimal est donc donné par :

$$\psi_j(t)=\arg\max_i\left\{p_{ij}\alpha_j^{r(j)}(t-u)\right\}.$$

Dans ce cas, la probabilité de la séquence d'états optimale associée à la séquence observée $x_0^{\tau-1}$ conjointement avec la $n^{ième}$ séquence d'état la plus probable est alors égale à :

$$\max_j\left\{\alpha_j^{r(j)}(\tau-1)\right\}$$

où les rangs des séquences d'états $\left(r(j);j=0,...,J-1\right)$ sont initialisés à 1 et le rang de la séquence d'état sélectionnée est incrémentée à 1 pour éviter de ne resélectionner la même séquence.

On obtient donc l'état final optimal suivant:

$$\tilde{s}_{\tau-1}=\arg\max_j\left\{\alpha_j^{r(j)}(\tau-1)\right\}.$$

Et enfin, la séquence d'états optimale est la suivante :

$$\tilde{s}_t=\psi_{\tilde{s}_{t+1}}(t+u).$$

Pour cet algorithme, la complexité en temps est de $O\left(LJ\tau(J+\tau)\right)$ et celle en complexité est de $O(LJ\tau)$.

2- Calcul des profils d'états résumant les séquences d'états.

a) Algorithme 'Avant-Arrière' pour calculer les profils d'états par sommation.

L'algorithme 'Avant-Arrière' (Churchill, 1989) permet de calculer les probabilités lissées $L_j(t)$, qui sont les probabilités de chaque état, calculées avec les paramètres estimés

sur l'ensemble de la période en utilisant l'ensemble des données disponibles. Rappelons également que cet algorithme permet d'implémenter l'étape E de l'algorithme EM et de calculer la vraisemblance des séquences observées $P\left(X_0^{\tau-1}=x_0^{\tau-1};\theta\right)$, ce que nous verrons un peu plus tard.

Cet algorithme repose sur la décomposition suivante de ces probabilités lissées :

$$\begin{aligned} L_j(t) &= P\left(S_{t+1}\neq j, S_t = j \middle| X_0^{\tau-1}=x_0^{\tau-1}\right) \\ &= \frac{P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1}\middle| S_{t+1}\neq j, S_t=j\right)}{P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1}\middle| X_0^{t}=x_0^{t}\right)} P\left(S_{t+1}\neq j, S_t=j\middle| X_0^t=x_0^t\right) \\ &= \qquad B_j(t) \qquad \times \qquad F_j(t). \end{aligned}$$

De ce fait, il se compose de 2 calculs récursifs (Guédon & Cocozza-Thivent, 1990). Le premier de *0* à $\tau-1$, correspond à la récurrence Avant et permet de calculer la vraisemblance de la séquence observée $x_0^{\tau-1}$ en calculant les probabilités filtrées $F_j(t)$ (Devijver, 1985). Le second de $\tau-1$ à *0*, correspond alors à la récurrence Arrière qui permet de calculer, soient les quantités $B_j(t)$, soient directement les quantités $L_j(t)$.

La récurrence Avant est donnée par,
$t=0,...,\tau-2$; $j=0,...,J-1$:

$$\begin{aligned} F_j(t) &= P\left(S_{t+1}\neq j, S_t=j\middle| X_0^t=x_0^t\right) \\ &= \frac{b_j(x_t)}{N_t}\left[\sum_{u=1}^{t}\left\{\prod_{v=1}^{u-1}\frac{b_j(x_{t-v})}{N_{t-v}}\right\} d_j(u)\sum_{i\neq j} p_{ij}F_i(t-u)\right. \\ &\qquad \left. +\left\{\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right\} d_j(t+1)\pi_j\right], \end{aligned}$$

où N_t est le facteur de normalisation tel que $N_t = P\left(X_t=x_t\middle| X_0^{t-1}=x_0^{t-1}\right)$.

La censure au temps $\tau-1$, du temps de séjour dans le dernier état visité est donnée par ,
$j=0,...,J-1$:

$$\begin{aligned} F_j(\tau-1) &= P\left(S_{\tau-1}=j\middle| X_0^t=x_0^t\right) \\ &= \frac{b_j(x_{\tau-1})}{N_{\tau-1}}\left[\sum_{u=1}^{\tau-1}\left\{\prod_{v=1}^{u-1}\frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}}\right\} D_j(u)\sum_{i\neq j} p_{ij}F_i(\tau-1-u)\right. \\ &\qquad \left. +\left\{\prod_{v=1}^{\tau-1}\frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}}\right\} D_j(\tau)\pi_j\right]. \end{aligned}$$

Notons que le temps exact passé dans le dernier état visité est inconnu, et seul le temps de séjour minimum est connu.

Revenons maintenant au facteur de normalisation N_t. Il peut être directement obtenu durant cette récurrence Avant. En effet, on obtient,
$t = 1, ..., \tau - 1$:

$$\begin{aligned} N_t &= P\left(X_t = x_t \middle| X_0^{t-1} = x_0^{t-1}\right) \\ &= \sum_j P\left(S_t = j, X_t = x_t \middle| X_0^{t-1} = x_0^{t-1}\right) \\ &= \sum_j \left[b_j(x_t) \left\{ \sum_{i \neq j} p_{ij} F_i(t-1) - F_j(t-1) + P\left(S_{t-1} = j \middle| X_0^{t-1} = x_0^{t-1}\right) \right\} \right] \end{aligned}$$

Comme nous l'avons signalé précédemment, cette récurrence permet de calculer la vraisemblance de la séquence observée pour le paramètre θ. En effet, nous avons :

$$\begin{aligned} P\left(X_0^{\tau-1} = x_0^{\tau-1}; \theta\right) &= P(X_0 = x_0) \prod_{t=1}^{\tau-1} P\left(X_t = x_t \middle| X_0^{t-1} = x_0^{t-1}; \theta\right) \\ &= \prod_{t-0}^{\tau-1} N_t. \end{aligned}$$

Ensuite, la récurrence Arrière consiste donc à calculer, soit $B_j(t)$, soit $L_j(t)$ pour chaque état j, en partant de $\tau - 1$ jusqu'au temps $t = 0$. Elle est initialisée pour $t = \tau - 1$ par, $j = 0, ..., J - 1$:

$$\begin{aligned} L_j(\tau - 1) &= P\left(S_{\tau-1} = j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= F_j(\tau - 1), \end{aligned}$$

d'où

$$B_j(\tau - 1) = 1.$$

Ensuite, on peut réécrire les probabilités $L_j(t)$, en les décomposant en 3 termes :

$$\begin{aligned} L_j(t) &= P\left(S_t = j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= P\left(S_{t+1} \neq j, S_t = j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) + P\left(S_{t+1} = j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &\quad - P\left(S_{t+1} = j, S_t \neq j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= L1_j(t) + L_j(t+1) - P\left(S_{t+1} = j, S_t \neq j \middle| X_0^{\tau-1} = x_0^{\tau-1}\right) \end{aligned}$$

La récurrence Arrière est basée sur les quantités $L1_j(t)$, calculées de la manière suivante :

$$L1_j(t)=\left[\sum_{k\neq j}\left[\sum_{u=1}^{\tau-2-t}\frac{L1_k(t+u)}{F_k(t+u)}\left\{\prod_{v=0}^{u-1}\frac{b_k(x_{t+u-v})}{N_{t+u-v}}\right\}d_k(u)\right.\right.$$
$$\left.\left.+\left\{\prod_{v=0}^{\tau-2-t}\frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}}\right\}D_k(\tau-1-t)\right]p_{jk}\right]F_j(t).$$

Cependant, puisque pour tout $t<\tau-1$, $L1_j(t)=B_j(t)F_j(t)$, la récurrence Arrière basée sur les quantités $B_j(t)$ sera directement déduite de l'expression de $L1_j(t)$ que nous venons de voir :

$$B_j(t)=\sum_{k\neq j}\left[\sum_{u=1}^{\tau-2-t}B_k(t+u)\left\{\prod_{v=0}^{u-1}b_k(x_{t+u-v})\right\}d_k(u)\right.$$
$$\left.+\left\{\prod_{v=0}^{\tau-2-t}\frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}}\right\}D_k(\tau-1-t)\right]p_{jk}.$$

La complexité de cette récurrence arrière est cubique pour le temps. Cependant, il est possible d'avoir une complexité de $O\left(J\tau\left(J+\tau\right)\right)$ en temps et de $O\left(J\tau\right)$ en espace.

b) Algorithme de Viterbi 'Avant-Arrière' pour calculer les profils d'états par maximisation.

Comme alternative aux profils d'états donnés par les probabilités lissées $L_j(t)$ calculées lors de l'algorithme 'Avant-Arrière', on peut proposer de calculer à l'aide de cet algorithme de Viterbi 'Avant-Arrière', les séquences d'états les plus probables passant par un état j à un instant t associées à la séquence observée. Cet algorithme repose sur la décomposition suivante :

$$\max_{s_0,\ldots,s_{t-1}}\max_{s_{t+1},\ldots,s_{\tau-1}}P\left(S_0^{t-1}=s_0^{t-1},S_t=j,S_{t+1}^{\tau-1}=s_{t+1}^{\tau-1},X_0^{\tau-1}=x_0^{\tau-1}\right)$$
$$=\left\{\max_{s_{t+1},\ldots,s_{\tau-1}}P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1},S_{t+1}^{\tau-1}=s_{t+1}^{\tau-1}\middle|S_{t-1}\neq j,S_t=j\right)\right.$$
$$\left.\times\max_{s_0,\ldots,s_{t-1}}P\left(S_t=j,S_0^{t-1}=s_0^{t-1},X_0^t=x_0^t\right)\right\}$$
$$=\beta_j(t)\alpha_j(t),$$

où les quantités $\alpha_j(t)$ seront calculées dans la récurrence Avant tandis que les quantités $\beta_j(t)$ seront calculées dans la récurrence Arrière.

La récurrence Avant de cet algorithme est alors la même que celle de l'algorithme de Viterbi.

Et ensuite, la récurrence Arrière, est initialisée pour $t=\tau-1$ par, $j=0,...,J-1$:

$$\beta_j(\tau-1)=1,$$

On introduit alors une nouvelle quantité :

$$\xi_k(t+1)=\max_{s_{t+2},...,s_{\tau-1}} P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1},S_{t+2}^{\tau-1}=s_{t+2}^{\tau-1}\middle|S_{t+1}=k,S_t\neq k\right),$$

pour obtenir la récurrence Arrière telle que :
$t=\tau-2,...,0 \quad ; \quad j=0,...,J-1$:

$$\begin{aligned}\beta_j(t)&=\max_{s_{t+1},...,s_{\tau-1}} P\left(X_{t+1}^{\tau-1}=x_{t+1}^{\tau-1},S_{t+1}^{\tau-1}=s_{t+1}^{\tau-1}\middle|S_{t+1}\neq j,S_t=j\right)\\&=\max_{k\neq j}\left[\max_{1\le u\le\tau-2-t}\left[\beta_k(t+u)\left\{\prod_{v=0}^{u-1}b_k\left(x_{t+u-v}\right)\right\}d_k(u)\right],\right.\\&\qquad\left.\left\{\prod_{v=0}^{\tau-2-t}b_k\left(x_{\tau-1-v}\right)\right\}D_k(\tau-1-t)\right]p_{jk}\\&=\max_{k\neq j}\left\{\xi_k(t+1)p_{jk}\right\},\end{aligned}$$

On retrouve ici la récurrence Arrière de l'algorithme 'Avant-Arrière' à la différence que l'on a remplacé la sommation par une maximisation et que l'on ne fait plus intervenir le facteur de normalisation.

Cet algorithme est donc tout simplement le mélange de la récurrence Avant de l'algorithme de Viterbi avec la récurrence Arrière de l'algorithme 'Avant-Arrière' avec une complexité identique à celle de l'algorithme de Viterbi.

3- Illustration de ces algorithmes.

A l'aide de fonctions du logiciel V-Plants, il est possible de connaître les L séquences d'états les plus probables pour une séquence observée, L étant fixé. Notons simplement ici, que le fait de devoir fixer au préalable la valeur de L, constitue un inconvénient de l'algorithme de Viterbi généralisé puisque l'on ne connait pas les probabilités a posteriori de ces séquences.

Pour notre cas, nous fixerons $L=10$, et de ce fait, nous pouvons donc sortir les 10 séquences d'états les plus probables pour chacune de nos séquences observées.

Séquence d'états observée :

0 0 2 1 0 0 1 0 1 2 2 2 2 2 2 0 2 2 2 2 2 2 2 0

Les 10 séquences d'états les plus probables :

0 0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3	1	(0.67368	0.67368	24)
0 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 3	2	(0.154785	0.828465	25)
0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3	3	(0.0981505	0.926616	26)
0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3	4	(0.0227401	0.949356	27)
0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 3	5	(0.0226233	0.971979	28)
0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3	6	(0.013999	0.985978	29)
0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3	7	(0.0044863	0.990464	29)
0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3	8	(0.00379017	0.994255	29)
0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3	9	(0.00207404	0.996328	30)
0 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3	10	(0.000877537	0.997206	30)

ordre de la séquence d'état

probabilité a posteriori de la séquence

probabilités a posteriori cumulées

Figure 19 : Exemple d'une séquence observée lors de l'année 1998 avec ses 10 séquences les plus probables calculées à l'aide de l'algorithme de Viterbi généralisé.

Dans cet exemple, il s'agit de la séquence observée ayant la probabilité a posteriori de la séquence d'états la plus probable associée, la plus forte de tout l'échantillon, soit 0.67368.

On peut également obtenir une représentation graphique de l'algorithme de Viterbi 'Avant-Arrière' pour une séquence observée. Sur le graphe de la Figure 20 suivante, on a une représentation de la séquence d'états la plus probable avec sa probabilité a posteriori, ainsi que les différentes alternatives successives à cette séquence. En effet, on retrouve bien la séquence d'états la plus probable en haut du graphe. Ensuite, on a une représentation des différentes séquences individuelles calculées par maximisation sur toutes les séquences d'états les plus probables. On remarque que l'on retrouve aussi les résultats de l'algorithme de Viterbi généralisé sur ce graphe. En effet, les différents points présents sur les courbes correspondent aux autres séquences d'états les plus probables. Par exemple, pour la seconde séquence d'états la plus probable, à partir de la Figure 19, on remarque que la différence provient du fait que l'on a un état 1 au lieu d'un état 2 au temps $t = 9$. On retrouve alors cette différence sur le graphe de la Figure 20, puisque pour $t = 9$, on a un symbole sur la courbe de l'état 1 avec une probabilité 0.154785, qui est la probabilité a posteriori de la seconde séquence d'états la plus probable.

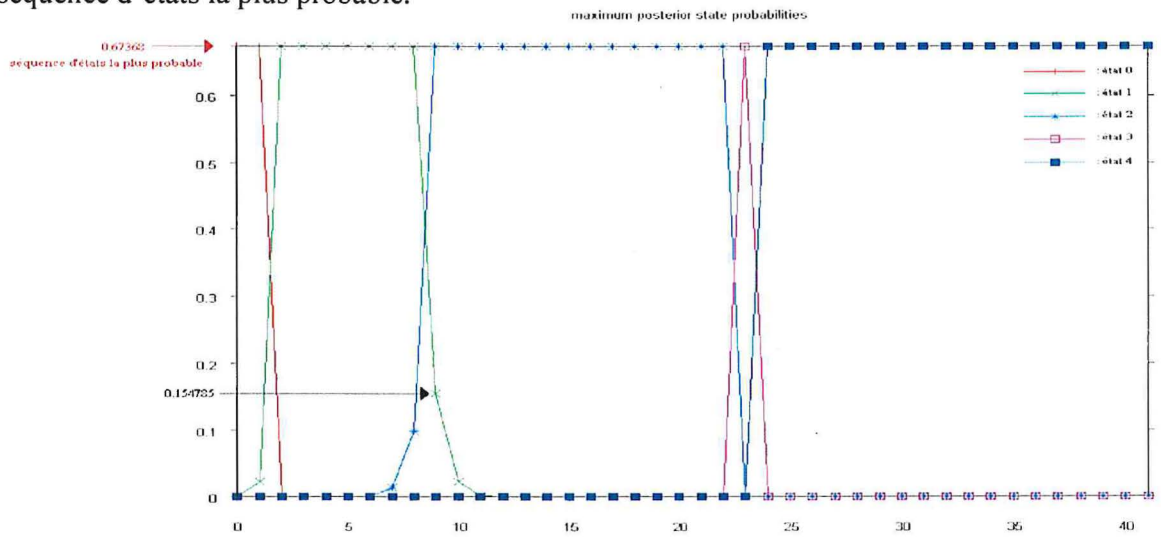


Figure 20 : Représentation graphique de l'algorithme de Viterbi 'Avant-Arrière' sur une séquence observée en 1998.

En complément de l'analyse de la probabilité a posteriori de ces différentes séquences d'états les plus probables, il est également important d'étudier le troisième indicateur présent dans les parenthèses de la Figure 21. Il peut s'interpréter comme un nombre de cases remplies dans un tableau (nombre d'états J × longueur séquence τ) correspondant à la taille du faisceau des séquences d'états. En effet, 24 est la longueur de cette séquence. Ensuite, en considérant la seconde séquence d'états la plus probables, on 'rempliera' une nouvelle case, celle correspondant à l'état 1 au temps $t = 9$. Toutes ces modifications sont signalées en rouge dans la figure suivante. Et on continue, ainsi de suite jusqu'à la $10^{ème}$ séquence qui n'engendrera pas de nouvelles cases par rapport aux 9 autres précédentes. On peut schématiser ce mécanisme comme dans la Figure 22.

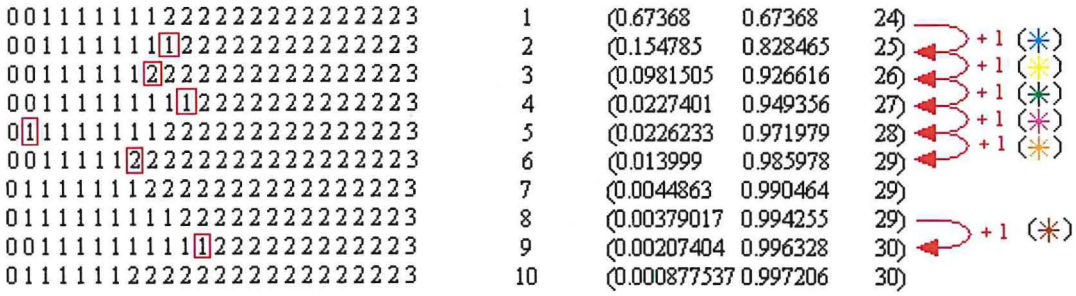


Figure 21 : Reprise des séquences de la Figure 19 en précisant les différentes modifications.

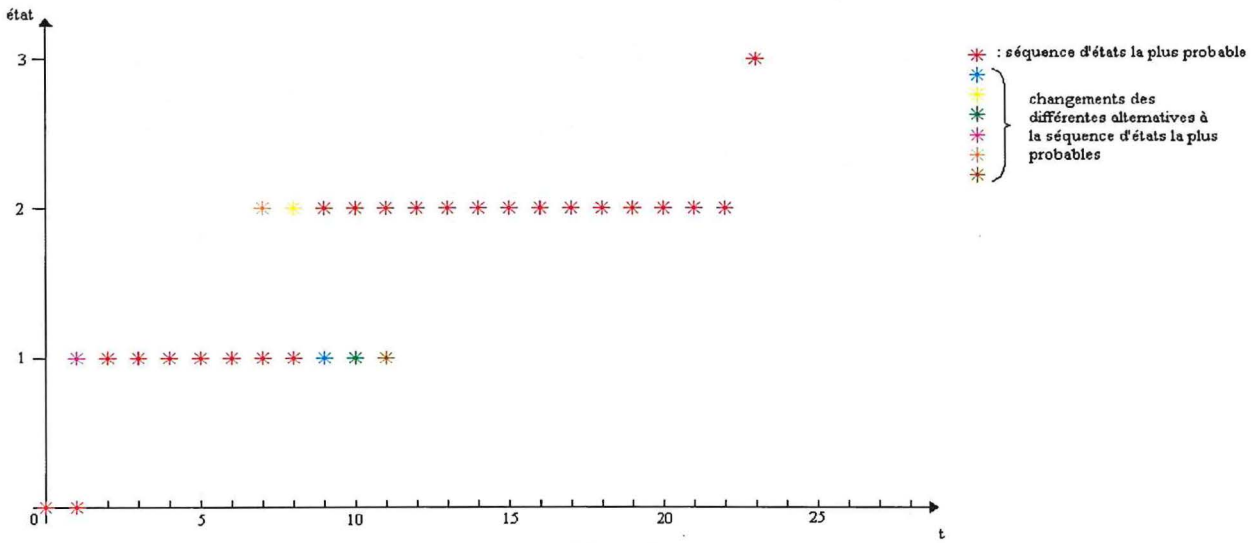


Figure 22 : Schématisation de la procédure de 'remplissage de cases'.

Nous allons maintenant construire un histogramme sur les valeurs de ces probabilités a posteriori pour toutes les séquences des 3 années étudiées. Pour chaque classe, on donne sa proportion dans l'échantillon total. On a alors une idée sur le poids de la séquence d'états la plus probables de chacune de nos séquences. En effet, si les probabilités sont faibles, cela signifie que l'on aura très certainement des alternatives avec des poids très proches de celui de la séquence d'états la plus probable. De ce fait, cela ne sera pas pertinent. En revanche, plus une probabilité sera forte, plus la séquence d'états la plus probable sera significative et aura des alternatives avec des probabilités faibles par rapport à elle, comme on peut le voir sur la Figure 21.

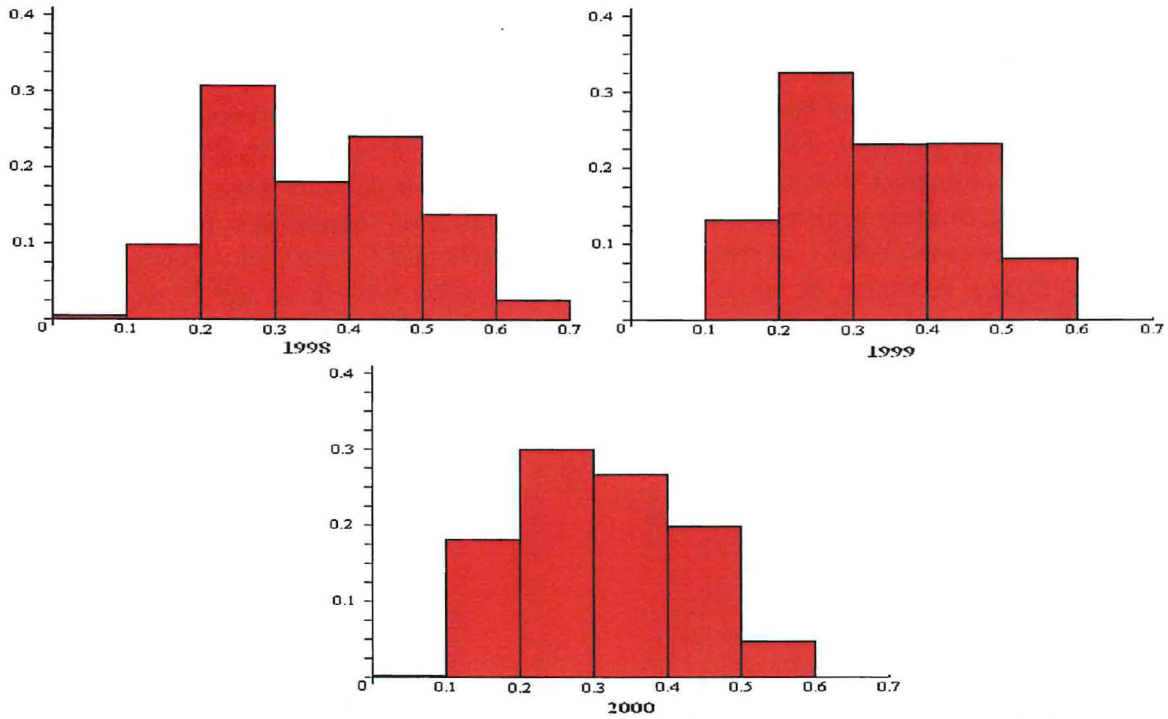


Figure 23 : Histogrammes des fréquences des probabilités a posteriori de nos séquences pour les 3 années.

On voit que la répartition est sensiblement la même pour les 3 années. On peut également voir qu'il y a peu de probabilités a posteriori inférieures à 0.1. Plus de 80% de ces probabilités sont comprises entre 0.2 et 0.5. On peut donc conclure que ces résultats sont relativement satisfaisants puisque la séquence d'états la plus probables de chacune de nos séquences observées a un poids assez fort. De ce fait, on pourra travailler avec ces séquences afin d'étudier la structure de ramification des plantes et l'éventuelle influence que peuvent avoir les parents. On peut néanmoins ajouter que si ces probabilités ne sont pas plus élevées, c'est principalement du à l'incertitude liée à la localisation de la transition entre l'état 1 et 2, et l'état 1 et 2 ou 3 pour l'année 2000.

Nous allons maintenant étudier l'homogénéité de nos séquences selon les parents des plantes.

IV. Etude de l'influence des parents sur la structure de ramification.

Nous allons désormais rechercher les éventuels effets héréditaires de nos attributs. Pour cela, nous allons travailler avec un modèle global, ceux évoqués dans le paragraphe *II*, tout en partitionnant notre échantillon en différents groupes. En procédant ainsi, on pourra directement comparer les différents sous-groupes formés.

Dans un premier temps, nous travaillerons sur les parents de notre tableau de croisement. Cela correspond en fait, aux individus issus des familles présentes dans la diagonale du tableau de la Figure 5. Il est intéressant de faire cette étude au préalable, car ainsi, si l'on ne constate pas de différences entre ces familles, on pourra penser qu'il n'y en aura pas non plus lors de la comparaison de toutes les autres familles.

On se contentera de faire les comparaisons sur la seule année 1998. En effet, l'échantillon est plus important cette année que les 2 autres, et, de plus, s'il existe des différences entre les parents en 1998, elles se retrouveront les années suivantes.

Après avoir estimé notre modèle global à l'aide de l'algorithme EM et étudier les séquences d'états les plus probables de chacune de nos séquences observées avec l'algorithme de Viterbi et le calcul des probabilités a posteriori, nous allons extraire les lois empiriques correspondant aux paramètres du modèle en utilisant les séquences d'états les plus probables calculées précédemment. Ces lois empiriques sont les lois de temps d'occupation des états et les lois d'observations au sein de ces états. Comme nous venons de le voir, nous effectuerons une partition selon les parents des individus.

1- Effectifs pour chaque parent.

PARENT	T18121	T18140	T18138	T18130	T18141	T17931	T17933	T17930	T8667	T8666
EFFECTIF	12	11	12	10	11	12	8	7	8	4

2- Lois empiriques.

Il nous est très difficile ici de comparer graphiquement les différentes lois empiriques. En effet, les effectifs sont trop faibles pour que l'interprétation soit pertinente.

En revanche, on peut toujours faire une comparaison des temps de séjour moyens et de leur variance dans chacun des états et pour chacune de nos familles, pour essayer de voir s'il existe une homogénéité entre ces groupes.

FAMILLE	LOIS D'OCCUPATION					
	ETAT 0		ETAT 1		ETAT 2	
	moyenne	variance	moyenne	variance	moyenne	variance
T18121	4.58	1.35	5.75	3.3	18.08	1.36
T18140	5.63	3.05	5.45	3.27	14.91	13.9
T18138	3.33	2.42	5.5	0.45	17.67	4.97
T18130	3.8	8.84	6.2	3.95	13.8	9.29
T18141	3.67	2	5.1	1.1	15.45	2.67
T17931	4.82	1.16	6	1.67	16.93	8.97
T17933	5.38	4.84	5	1.71	15.12	8.98
T17930	4.14	0.81	6.28	3.24	17.57	4.95
T8667	4.5	1.14	6	3.14	18	11.14
T8666	3.5	1.67	7	2	18.75	2.25

Figure 27 : Tableau des temps de séjour moyen et leur variance.

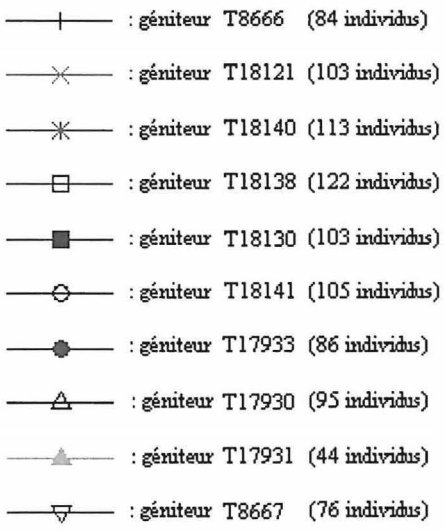
Comme nous l'avons dit, on peut difficilement conclure en une éventuelle influence des parents sur la structure de ramification des plantes. Cependant, le tableau précédent pourrait nous faire croire qu'il existe une homogénéité entre les 10 parents car les moyennes des temps d'occupation ne sont pas très différentes pour chaque état.

Ne pouvant pas aller plus loin dans l'étude de l'influence des parents, nous allons maintenant partitionner notre échantillon selon les géniteurs

V. Etude de l'influence des géniteurs sur la structure de ramification.

Nous allons désormais effectuer des groupes selon les géniteurs de nos individus. Dans cette étude, un couple père/mère est équivalent à un couple mère/père. De ce fait, dans les différents sous-échantillons que nous allons utiliser, on retrouvera tous les individus qui ont le même géniteur. Pour reprendre le tableau de croisement de la Figure 5, on ne tient désormais plus compte des lignes et des colonnes, mais pour un géniteur i, on répertoriera toutes les plantes de la colonne i et celles de la ligne i comme étant un seul groupe. Cependant, en procédant ainsi, on se retrouve donc avec des individus présents dans 2 groupes différents. De ce fait, on émet un biais au profit de l'homogénéisation de nos sous-échantillons en diminuant les différences. De ce fait, nous nous retrouvons avec des effectifs importants, et malgré le biais que l'on ajoute, cela devient intéressant de comparer ces groupes afin de voir s'il y a un effet sur la structure de ramification du au géniteur de la plante.

1- Légende utilisée pour les différents géniteurs avec les effectifs respectifs.



2- Lois d'occupation des états.

a) Représentation graphique des lois.

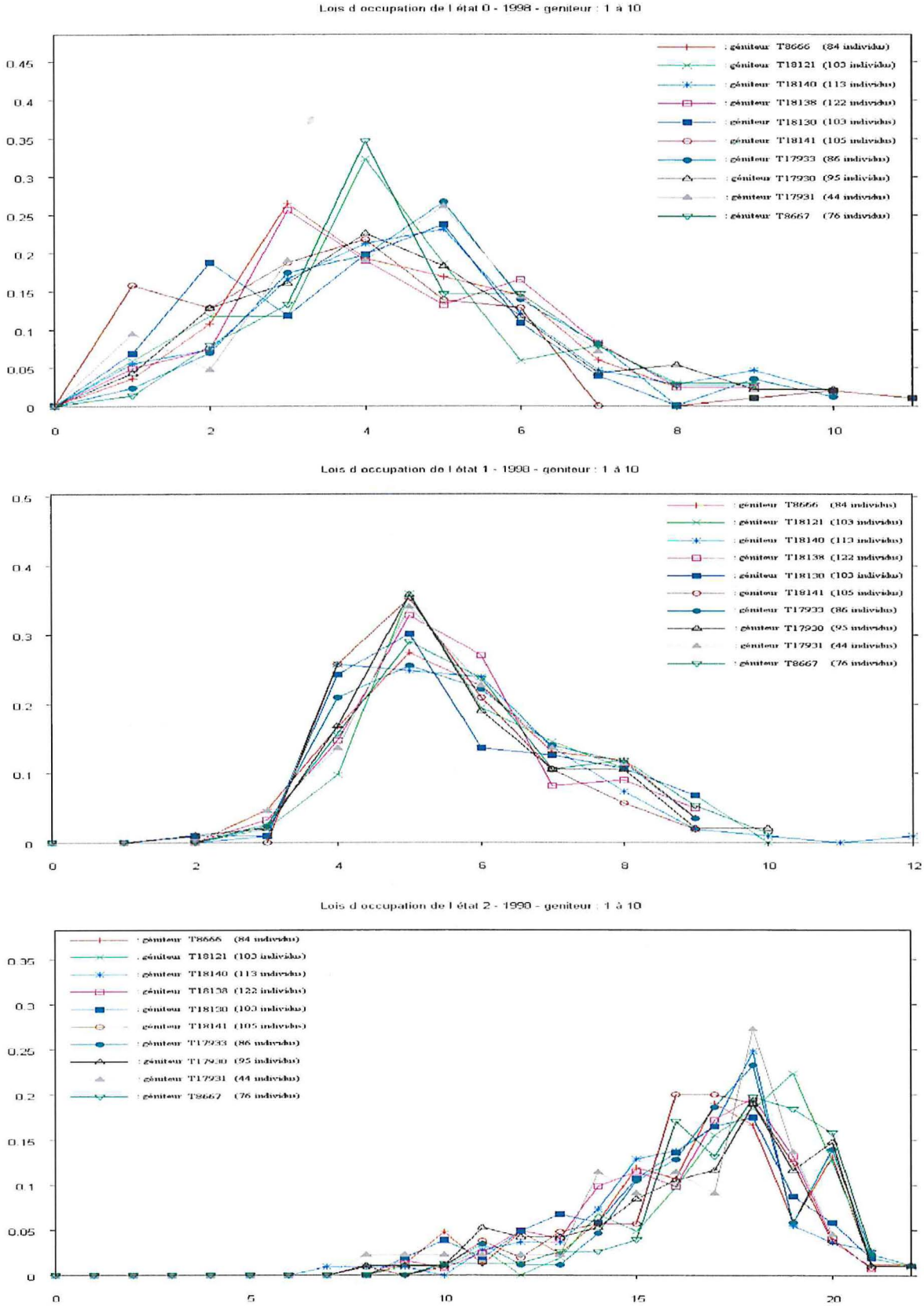


Figure 29 : Lois d'occupation des états 0, 1 et 2 pour les 10 géniteurs en 1998.

Dans le cas des lois d'occupations de l'état 0, on a quelques différences entre nos 10 géniteurs. Néanmoins, on peut y voir des similitudes importantes. De ce fait, et ceux malgré une légère différence au niveau de l'allure des courbes, on peut dire qu'il y a très peu de différences entre les géniteurs.

En ce qui concerne les lois d'occupations de l'état 1, la ressemblance entre chaque sous-groupe est beaucoup plus marquée. On peut alors conclure qu'il n'y a pas de différences entre nos 10 groupes.

Et enfin, pour les lois d'occupations de l'état 2, bien que l'on ne puisse pas conclure en l'absence de différences entre les sous-échantillons, cela reste tout de même relativement faible.

b) Comparaison des moyennes et des variances.

Afin d'affirmer nos conclusions sur cette homogénéité entre les géniteurs, nous pouvons effectuer des comparaisons entre les différentes moyennes de nos groupes.

	LOIS D'OCCUPATION					
	ETAT 0		ETAT 1		ETAT 2	
GENITEUR	moyenne	variance	moyenne	variance	moyenne	variance
T8666	4.14	2.71	5.73	2.25	16.43	8.05
T18121	4.27	3.41	6.01	2.48	17.34	6.07
T18140	4.58	4.06	5.64	2.32	16.22	6.75
T18138	4.38	3.44	5.69	2.02	16.25	6.37
T18130	4.11	4	5.69	2.61	16.02	7.7
T18141	3.71	4.19	5.41	1.55	16.44	5.71
T17933	4.65	3.12	5.73	2.17	17.05	5.72
T17930	4.46	4.12	5.66	2.35	16.58	8.91
T17931	4.19	2.74	5.61	1.78	16.16	8.37
T8667	4.59	2.81	5.87	2.44	17.47	5.05

Figure 30 : Tableau des temps de séjour moyen et leur variance.

On retrouve clairement dans ces résultats, l'homogénéité qu'il existe entre nos groupes. Les temps de séjour moyen dans chacun des états sont proches pour chacun des géniteurs, ainsi que leur variance.

c) Analyse de variances – ANOVA.

L'ANOVA, ou analyse de variances, est une technique statistique qui permet de comparer les moyennes μ_i de plusieurs populations. Le but de ce test paramétrique est alors de dire si, parmi l'ensemble des groupes étudiés, au moins un d'entre eux diffère des autres. Mais si c'est le cas, on ne sera ni combien, ni lequel ou lesquels.

L'hypothèse nulle de ce test est donc :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_{10},$$

testée contre l'hypothèse alternative suivante :

H_1 : *il y a au moins une moyenne différente.*

Nous allons appliquer cette analyse sur nos 10 groupes pour chacune des lois d'occupation précédente. Les résultats seront répertoriés dans le tableau suivant avec, pour chaque groupe, l'effectif et la valeur de la statistique F du test avec la probabilité critique p associée. C'est cette dernière qui va permettre de confirmer ou non l'hypothèse alternative. En effet, il s'agit de la probabilité, ou encore le risque, de commettre une erreur en déclarant qu'il existe une différence entre nos groupes. Il suffit alors de comparer cette valeur à un seuil de signification qui est généralement égal à 0.05, et qui correspond à l'erreur de première espèce. Si la probabilité critique trouvée est supérieure à 0.05, on accepte l'hypothèse et on conclut ainsi qu'il n'y a aucune différence significative entre nos 10 groupes. En revanche, si elle est inférieure, on rejettera l'hypothèse nulle.

		ETAT 0		ETAT 1		ETAT 2	
GENITEUR	EFFECTIF	F	p	F	p	F	p
T8666	84	2.175	0.02172	1.1	0.36	3.45	0.00038
T18121	103						
T18140	113						
T18138	122						
T18130	103						
T18141	105						
T17933	86						
T17930	95						
T17931	44						
T8667	76						

Figure 31 : ANOVA pour les lois d'occupation des états 0, 1 et 2.

On peut donc conclure qu'il n'y a aucune différence significative entre les 10 groupes en ce qui concerne les lois d'occupations de l'état 1, puisque 0.360018>0.05. En revanche, on rejette cette hypothèse pour les 2 autres états. On peut alors comparer nos groupes 2 à 2 afin de voir d'où provient cette différence. Dans les Figures 32 et 33, les cases colorés (▭) correspondent aux cas où l'hypothèse nulle d'égalité des moyennes entre les 2 groupes est rejetée.

	T8666	T18121	T18140	T18138	T18130	T18141	T17933	T17930	T17931	T8667
T8666		0.25 0.624	2.60 0.104	0.87 0.355	0.017 0.865	2.41 0.118	3.71 0.053	1.19 0.275	0.02 0.855	2.79 0.093
T18121			1.34 0.247	0.18 0.674	0.375 0.548	4.22 0.039	2.02 0.15	0.41 0.532	0.065 0.787	1.33 0.248
T18140				0.63 0.433	2.92 0.085	9.6 0.002	0.060 0.79	0.21 0.650	1.26 0.262	0.0001 0.988
T18138					1.1 0.297	6.49 0.011	1.116 0.29	0.07 0.779	0.34 0.565	0.62 0.439
T18130						1.93 0.162	3.8 0.05	1.40 0.236	0.054 0.801	2.81 0.0913
T18141							11.057 0.001	6.36 0.012	1.8 0.179	9.12 0.003
T17933								0.49 0.492	2 0.156	0.056 0.799
T17930									0.53 0.473	0.21 0.649
T17931										1.52 0.218
T8667										

Figure 32 : ANOVA pour les géniteurs pris 2 à 2, pour la loi d'occupation de l'état 0.

Au vue de ces résultats, on peut penser que la différence survenue lors de l'ANOVA générale effectuée sur les 10 groupes pourrait provenir du géniteur T18141.

On effectue le même procédé pour l'état 2.

	T8666	T18121	T18140	T18138	T18130	T18141	T17933	T17930	T17931	T8667
T8666		5.52 0.0189	0.28 0.603	0.24 0.633	0.98 0.324	0.0006 0.93	2.36 0.122	0.12 0.729	0.26 0.619	6.57 0.0109
T18121			10.34 0.0016	10.72 0.0014	13.03 0.0005	7.18 0.008	0.68 0.415	3.85 0.0483	6.36 0.0122	0.14 0.709
T18140				0.006 0.898	0.3 0.594	0.40 0.531	5.22 0.022	0.84 0.363	0.02 0.867	11.62 0.001
T18138					0.41 0.53	0.34 0.565	5.30 0.0212	0.79 0.378	0.035 0.83	12.04 0.0008
T18130						1.36 0.243	7.27 0.0075	1.87 0.169	0.076 0.774	14.05 0.0003
T18141							3.06 0.0779	0.14 0.711	0.37 0.55	8.7 0.0037
T17933								1.33 0.248	3.47 0.061	1.36 0.243
T17930									0.61 0.443	4.7 0.03
T17931										7.69 0.0064
T8667										

Figure 33 : ANOVA pour les géniteurs pris 2 à 2, pour la loi d'occupation de l'état 2.

Dans ce cas là, le rejet de l'hypothèse d'égalité entre 2 moyennes fait intervenir plus de géniteurs. Cependant, on peut penser que les géniteurs T18121 et T8667 sont plus fortement responsables du rejet de l'hypothèse nulle lors de l'ANOVA générale.

Nous allons maintenant refaire une ANOVA en enlevant pour les 2 états les géniteurs cités.

LOIS D'OCCUPATION	F	p
ETAT 0 SANS GENITEUR T18141	1.051	0.3958
ETAT 2 SANS GENITEURS T18121 ET T8667	1.273	0.26

Figure 34 : Nouvelle ANOVA des lois d'occupation des états 0 et 2.

On obtient alors des conclusions différentes puisque désormais on accepte l'hypothèse nulle d'égalité des moyennes dans les 2 cas. Ce rejet du géniteur T18141 pour obtenir une homogénéisation des moyennes de nos groupes est peu évident sur le premier graphique de la Figure 29 bien que la courbe —⊖— aborde une allure légèrement différente des autres. Il en est de même pour le troisième graphique de cette Figure 29. En effet, on voit que les géniteurs T18121 (—✕—) et T8667 (—▽—) semblent être sensiblement différents des autres courbes, mais cela est très faible.

Nous allons maintenant étudier les lois d'observations au sein de ces états.

3- Lois d'observation empirique dans un état.

Comme nous l'avons vu dans le paragraphe *IV*, nous nous limiterons à l'étude des lois d'observations dans l'état 1.

a) Représentation graphique.

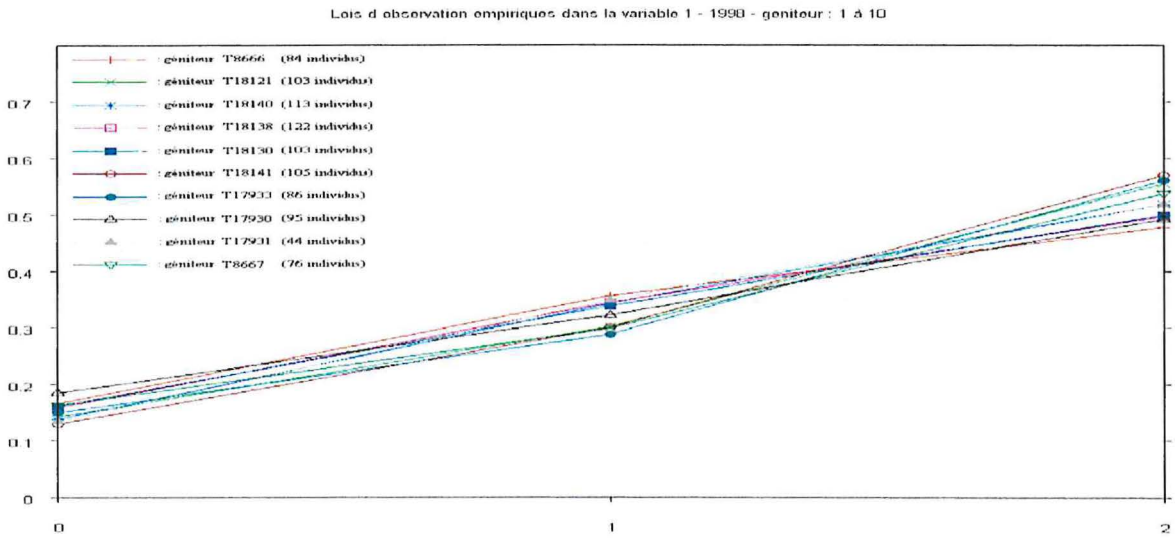


Figure 35 : Lois d'observation dans l'état 1 pour les 10 géniteurs en 1998.

En revanche, contrairement donc à ce que nous avions pu voir dans le paragraphe précédent, ici les lois sont très similaires. En effet, on a des proportions très proches pour chacune des observations.

Puisque les lois d'occupations sont également très similaires dans les états 0 et 2, on conclut qu'il n'y a pas ou très peu de différences entre nos groupes.

b) Comparaison des distributions – Test de Kruskal-Wallis.

Dans ce cas, nous utiliserons le test de Kruskal-Wallis. En effet, ce test est utilisé lorsqu'il faut décider si k échantillons indépendants sont issus d'une même population. Il est non-paramétrique et de ce fait on travaille désormais sur les rangs des observations. La statistique du test est alors construite à partir des moyennes des rangs des observations dans les différents sous-échantillons étudiés.

Nous allons alors dresser le tableau avec la valeur de la statistique du test et la probabilité critique p associée pour les 10 groupes, avec les hypothèses suivantes :

H_0 : *les 10 échantillons sont extraits d'une même population*

H_1 : *il y a au moins un groupe issu d'une population différente des autres*

GENITEUR	EFFECTIF	χ^2	p
T8666	84	19.699	0.0194
T18121	103		
T18140	113		
T18138	122		
T18130	103		
T18141	105		
T17933	86		
T17930	95		
T17931	44		
T8667	76		

Figure 36: Test de Kruskal-Wallis pour les lois d'observations dans l'état 1.

La valeur de la probabilité critique est inférieure à 0.05. On rejette alors l'hypothèse nulle et donc le fait que les densités de nos 10 groupes sont égales. On peut alors effectuer le test de Kruskal-Wallis en prenant nos groupes 2 à 2, comme vu précédemment.

	T8666	T18121	T18140	T18138	T18130	T18141	T17933	T17930	T17931	T8667
T8666		5.9 0.0121	2.29 0.132	0.4 0.52	0.39 0.525	8.77 0.002	5.43 0.015	0.003 0.906	1.49 0.227	2.2 0.14
T18121			1.03 0.307	3.97 0.0442	3.55 0.056	0.36 0.545	0.003 0.9	5.70 0.0133	0.53 0.479	0.59 0.453
T18140				0.95 0.328	0.83 0.364	2.58 0.11	1.004 0.31	2.14 0.146	0.002 0.911	0.02 0.858
T18138					0.0004 0.927	6.62 0.0088	3.635 0.053	0.34 0.558	0.6 0.447	1 0.314
T18130						5.98 0.0117	3.28 0.065	0.33 0.562	0.55 0.471	0.89 0.345
T18141							0.25 0.611	8.52 0.002	1.42 0.239	1.69 0.196
T17933								5.22 0.0167	0.55 0.471	0.6 0.45
T17930									1.37 0.249	2.06 0.154
T17931										0.004 0.899
T8667										

Figure 37 : Test de Kruskal-Wallis pour les géniteurs pris 2 à 2, pour les lois d'observations dans l'état 1.

De la même manière que lors de l'étude de l'ANOVA sur les lois d'occupations de l'état 2, les cas où l'hypothèse nulle est rejetée font intervenir plusieurs géniteurs. Nous pouvons alors faire un tableau montrant les différentes probabilités obtenues lorsque l'on enlève certains géniteurs dans notre test.

GENITEURS REJETES	χ^2	p
T8666	15.97	0.0425
T18121	16.83	0.0319
T18138	17.99	0.0206
T18130	18.39	0.0178
T18141	13.50	0.0957
T17933	17.33	0.027
T17930	16.12	0.0405

Figure 38 : Tableau des statistiques du test de Kruskal-Wallis et de la probabilité critique associée pour les différents cas étudiés.

On voit donc qu'il suffit d'enlever une nouvelle fois le géniteur T18141 pour obtenir une probabilité critique supérieure à 0.05 et ainsi valider l'hypothèse d'égalité des distributions pour les autres groupes.

Si l'on reprend le graphe de la Figure 35, on voit que ce choix n'est pas du tout évident, mais en analysant plus précisément les courbes, on remarque que la courbe 6 (—⊖—) correspond au géniteur qui a la plus petite proportion de 0 et la plus grosse proportion de 2 dans l'état 1. Ce qui, sans en faire un échantillon à part, peut illustrer le fait que sans lui, il n'y a pas de différences significatives entre les autres groupes.

4-Conclusion.

Malgré quelques différences observées lors de ces études, notamment lorsque l'on prend en compte le géniteur T18141 dans nos différents tests de comparaison de moyennes, on voit qu'il y a une homogénéité claire dans tous nos résultats. On peut donc penser qu'il n'y a pas effet des géniteurs sur ces séquences pour toutes les plantes étudiées.

CONCLUSION ET PERSPECTIVES

Un des objectifs initiaux de ce travail de stage était de mettre en place une modélisation statistique sur nos séquences d'attributs afin de rendre compte ou non du caractère héritable de ces attributs pour notre base de données de caféiers arabica. Le caractère sur lequel nous nous sommes attardés est le nombre de ramification sur chacun des entre-nœuds de la tige principale de nos plantes. Nous en avons alors extrait des séquences, sur lesquelles nous avons fait l'hypothèse d'un modèle. Tout au long de ce stage, nous avons travaillé avec des semi-chaînes de Markov cachées, ce qui s'est avéré être un bon choix au vue de la qualité de l'ajustement sur nos données. A partir de là, nous avons pu étudier les différentes lois empiriques correspondant aux paramètres de ce modèle. Afin de voir s'il existait des effets génétiques sur les différents attributs étudiés, nous avons fait un échantillonnage de notre population selon les géniteurs dont les plantes sont issues. Enfin, après avoir fait des analyses statistiques, on a pu conclure qu'il y avait peu de différences significatives entre tous les géniteurs présents dans notre échantillon au niveau de l'architecture des tiges principales de ces plantes. Il serait alors intéressant de reproduire cette démarche de modélisation à une autre échelle, comme par exemple, sur les branches.

Cependant, on pourrait penser que le modèle ultime pour modéliser nos données, aurait été un modèle qui combinerait une semi-chaîne de Markov cachée et un modèle linéaire mixte généralisé. En effet, nous avons affaire à des séquences structurées en phases successives, asynchrones entre individus, modélisée par une semi-chaîne de Markov, et des données influencées par des covariables pouvant varier dans le temps, et présentant une hétérogénéité inter-individuelle, ce qui correspond à un modèle linéaire mixte. Ces effets aléatoires peuvent être introduits au sein de modèles linéaires généralisés, qui eux permettent d'analyser des données discrètes. Un modèle linéaire mixte peut alors être vu comme un cas particulier des modèles linéaires mixtes généralisés. Ces derniers seront associés aux états de la chaîne de Markov sous-jacente. Pour chacune des phases, la tendance et les covariables seront modélisés par des effets fixes, et un effet aléatoire modélisera l'hétérogénéité entre les individus.

L'étude de la nature des variations observées dans des données relevées sur une population est l'un des objectifs principaux de la statistique. Lorsqu'une population se divise en sous-groupes, il est intéressant de savoir si les observations se comportent de manière identique d'un sous-groupe à l'autre. Si c'est le cas, cette répartition en sous-population sera dite homogène. A contrario, constater une hétérogénéité peut constituer une source d'information importante lors de l'analyse de données.

Le découpage en sous-groupes peut-être précis ou à l'inverse flou. En effet, dans certaines situations une hétérogénéité des données est pressentie sans bien savoir à quel découpage elle correspond précisément ; autrement dit, si telle donnée appartient ou non à tel sous-groupe. Les modèles de mélange apparaissent alors comme un outil naturel permettant de prendre en compte ce type d'hétérogénéité. Des lois sont supposées pour chaque classe, et en affectant à chaque donnée une certaine probabilité d'appartenance aux différentes classes,

ces modèles permettent de respecter la non connaissance exacte du découpage Dietz (1992), Dietz & Böhning (1995).

Nous allons maintenant présenter brièvement les différents modèles évoqués depuis le début de ce paragraphe.

Tout d'abord, les modèles linéaires mixtes permettent d'étudier la variabilité que présentent des données, de façon plus précise et plus élaborée que le simple modèle linéaire classique en y autorisant les effets aléatoires (Searle *et al.*, 1992). Au cours d'une expérience, diverses sources de variations peuvent influencer les valeurs de la variable réponse. Ces différentes sources de variation sont modélisées par des facteurs qui peuvent avoir 2 natures : fixe ou aléatoire. On parle alors de facteurs à effets fixes et de facteurs à effets aléatoires (Eisenhart, 1947).

- les facteurs à effets fixes ont en général un nombre fini de niveaux et les données se répartissent sur ces différents niveaux. On souhaite en tirer une information concernant l'effet de chaque niveau sur la variable d'intérêt.
- les facteurs à effets aléatoires ont un nombre potentiellement infini de niveaux et les données (en nombre fini) se répartissent sur un échantillon aléatoire de ces niveaux. La façon dont chacun des niveaux influe sur le résultat ne présente pas d'intérêt. En revanche, on souhaite connaître la part de variabilité induite par ces effets.

Ensuite, les modèles linéaires généralisés (Wedderburn & Nelder, 1972), qui sont, comme leur nom l'indique, une généralisation des modèles linéaires classiques en termes de loi de probabilité, mais également en termes de lien à la linéarité. Cette famille de modèles permet d'étudier la liaison entre une variable dépendante et un ensemble de variables explicatives.

Et enfin, les modèles linéaires mixtes généralisés, qui sont donc des modèles à la croisée de 2 types d'extension des modèles linéaires classiques. La première est une extension en termes de loi et donne naissance à la classe des modèles linéaires généralisés, et la deuxième est une extension en termes d'introduction d'effets aléatoires pour aboutir à la classe des modèles linéaires mixtes.

Revenons maintenant à notre choix d'utiliser un modèle linéaire mixte généralisé. La détermination d'un modèle provient de la loi marginale de la variable réponse étudiée. Dans le cadre de cette étude, nous avons travaillé sur les lois d'observations, et principalement celles observées dans l'état 1 puisque ce dernier correspond à un mélange des différentes observations, et les temps de séjour. Or, dans ce cas, les lois d'observation ne sont pas gaussiennes. On pourrait alors se demander s'il serait raisonnable de mettre un modèle linéaire généralisé sur nos états. Ce dernier pourrait être soit ordinal, soit cardinal, du fait du caractère catégoriel de nos observations. Cependant, en procédant ainsi, on ne tient pas compte de l'effet géniteur (et même de celui des parents, en considérant que les effectifs sont suffisants pour conclure qu'il y a homogénéité des résultats, comme nous l'avons vu dans le paragraphe *IV* de la seconde partie). Il serait alors plus judicieux de leurs associer des modèles linéaires mixtes généralisés. En incluant ces effets aléatoires, on influencera ainsi les différentes lois empiriques.

On peut alors conclure qu'il serait intéressant d'utiliser une combinaison semi-markovienne de modèles linéaires mixtes généralisés. Cette famille de modèles hérite donc de la structure cachée de chacun des 2 modèles qui le composent et ainsi de 2 types de variables non-

observables: les effets aléatoires et les états cachés. De ce fait, les paramètres de ce genre de modèle sont de 2 types. Il y a, d'une part, les paramètres liés à la semi-chaîne de Markov sous-jacente, c'est-à-dire les probabilités initiales ainsi que les probabilités de transition, et, d'autre part, les paramètres associés aux J modèles linéaires mixtes généralisés.

Il serait donc intéressant d'utiliser un modèle de ce genre sur nos données et de reproduire la même démarche statistique.

BIBLIOGRAPHIE

[1] CIRAD.
Website: www.cirad.fr

[2] Formation à OpenAlea.
Website: http://elearning.cirad.fr

[3] AmapMod.
Website: http://amap.cirad.fr/amapmod/refermanual15/accueil.html

[4] Python.
Website: www.python.org.

[5] Python, tutoriel en français.
Website: www.tuxfinder.com/french/Python

[6] BAUM L. E., PETRIE T., SOULES G., WEISS N. (1970)
A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41, 164-171.

[7] BROUARD T., SLIMANE M., ASSELIN DE BEAUVILLE J-P., VENTURINI G. (1998)
Apprentissage d'une chaîne de Markov cachée. Problèmes numériques liés à l'application à l'image. *Revue de Statistique Appliquée*, 83-108.

[8] CHURCHILL G. A. (1989)
Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, **51**, 79-94.

[9] CILAS C., GODIN C., BERTRAND B., BAILLERES H. (2006)
Genetic study on the physical properties of Coffea Arabica L.Wood. *Trees-Structure and Function*. **20**(5), 587-592.

[10] DEMPSTER A. P., LAIRD N. M., RUBIN D. B. (1977)
Maximum likehood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*, 1-38.

[11] DURBIN R., EDDY S., KROGH A., MITCHISON G. (1998)
Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press.*

[12] GODIN C. (1998)
Le codage des plantes utilisé dans AMAPMOD. Version 1. *Document de travail du programme Modélisation des plantes*, 3-98.

[13] GUEDON Y. (2003)
Estimating hidden semi-Markov chains from discrete sequences. *J. Comput. Graphical Stat* **12**(3), 604-639.

[14] GUEDON Y. (2007)
Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computationnal Statistics and Data Analysis*, **51**(5), 2379-2409.

[15] GUEDON Y., BARTHELEMY D., CARAGLIO Y., COSTES E. (2001)
Pattern analysis in branching and axillary flowering sequences. *Journal of theoretical biology*, **212**, 418-520.

[16] GUEDON Y., HEURET P., COSTES E. (2003)
Comparison methods for branching and axillary flowering sequences. *Journal of theoretical biology*, **225**, 301-325.

[17] MACKAY-ALTMAN R. (2007)
Mixed hidden Markov models: an extension of the hidden Markov models to the longitudinal data settings. *Journal of the American Statistical Association*, **102**(477), 201-210.

[18] MURI-MAJOUBE F., PRUM B. (2001)
Une approche statistique de l'analyse des génomes. *La Gazette des Mathématiciens*, **78**, 63-98.

[19] RABINER L. (1989)
A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEEE*, 77(2).

[20] RENTON M., GUEDON Y., GODIN C., COSTES E. (2006)
Similarities and gradients in growth unit branching patterns during ontogeny in "Fuji" apple trees: a stochastic approach. *Journal of Experimental Botany*, **57**(12), 3131-3143.

[21] SEGURA V. (2007)
Etude des déterminismes génétiques de l'architecture du pommier. *Thèse de doctorat, Montpellier SupAgro*.

[22] SEGURA V., CILAS C, LAURENS F., COSTES E. (2006)
Phenotyping progenies for complex architectural traits : a strategy for 1-year-old apple trees (*Malus x domestica* Borkh.). *Tree Genetics & Genomes* **2**, 140-151.

[23] SWINNEN G. (2005)
Apprendre à programmer avec Python.

CS 26-06-08

Les travaux actuels du CIRAD visent à développer le SIG sur l'ensemble de l'aire grégarigène. L'essentiel des travaux repose sur la création d'une carte des biotopes acridiens concernant toute l'aire grégarigène. Parallèlement à ceci et en relation au caractère d'outil d'aide à la décision du SIG une réflexion a été engagée concernant l'incertitude associée au diagnostic de risque. C'est au sein de cette problématique visant à replacer le diagnostic établi dans un contexte de fiabilité et à lui associer des considérations de confiance dans les résultats que s'intègrent les travaux réalisés durant ce stage.

Le présent travail s'intéresse aux couches d'information pluviométrique et acridienne. L'interpolation des données ponctuelles est en effet source d'incertitude quant aux résultats. La démarche s'articule autour de la mesure de la précision de ces interpolations. Les objectifs poursuivis sont triples. Ils visent à disposer d'outils permettant d'estimer l'incertitude sur ces couches d'informations acquises par interpolation, à déterminer les moyens nécessaires pour garantir leur fiabilité, et à intégrer au SIG et à ses prévisions une évaluation de leur qualité.

En conséquence, le plan de ce mémoire s'organise en 3 parties avec tout d'abord la présentation du contexte technique du SIG, de la problématique de l'interpolation des données ponctuelles, et des outils disponibles pour tenter d'estimer l'incertitude associée aux résultats. La deuxième partie présente la méthodologie développée pour sélectionner une méthode d'interpolation et des paramétrages optimaux, pour identifier des réseaux minimaux de collecte de l'information capables de garantir une incertitude acceptable sur les couches d'information acquises par interpolation, et pour la réalisation et l'intégration au sein du SIG d'applications renseignant sur la qualité et la fiabilité des informations pluviométriques et acridiennes utilisées pour construire le diagnostic de risque de pullulation. La troisième partie présente les résultats.

Résumé