Master Biostatistique et Statistiques Industrielles

Rapport de stage

Modélisation de la variabilité génétique de la transpiration des fruits de pêcher

---

Auteur
TUZIN Nicolas

Lieu du stage
Institut National de la Recherche Agronomique

Maîtres de stage
Mme. QUILOT-TURION Bénédicte
Mr. MEMMAH Mohamed

Encadrant universitaire
Mr. POULIN Nicolas

Année universitaire 2016-2017

Résumé

Ce rapport présente une approche pour estimer les paramètres du sous-modèle conductance en y intégrant la variabilité génétique de ces paramètres. Le but est de simplifier la paramétrisation du modèle en déterminant les paramètres dépendants du génotype et les paramètres indépendants du génotype.

L'utilisation des méthodes de régression non paramétrique pour la masse fraîche en fonction du temps et pour la conductance totale en fonction de la masse fraîche était nécessaire pour pouvoir utiliser le sous-modèle conductance lors des estimations des paramètres.

Ensuite, une analyse de sensibilité a été appliquée dans le but d'évaluer l'influence des paramètres du sous-modèle conductance sur la conductance totale.

Enfin, pour calibrer le modèle, un algorithme à évolution différentielle a été utilisé. Comme c'est un algorithme stochastique, plusieurs répétitions ont été effectuées.

Mots-clés : variabilité génétique, paramétrisation du modèle, régression non paramétrique, analyse de sensibilité, algorithme à évolution différentielle.

Remerciements

Au terme de ce travail, je tiens à exprimer ma gratitude et mes remerciements pour toutes les personnes qui ont contribué à sa réalisation.

Je tiens d'abord à remercier mes tuteurs, Madame QUILOT-TURION Bénédicte et Monsieur MEMMAH Mohamed, pour leurs nombreux conseils et disponibilités.

J'adresse mes remerciements à GIS Fruits pour le financement de mon stage.

Je remercie également tous les membres de l'unité PSH, en particulier Monsieur VALSESIA Pierre.

Je voudrais aussi remercier mon encadrant, Monsieur POULIN Nicolas, pour avoir suivi mon travail.

Présentation de l'entreprise

L'INRA (Institut National de la Recherche Agronomique) est un organisme lié à la recherche en agriculture, alimentation et environnement. Premier institut de recherche agronomique en Europe, et deuxième dans le monde en sciences agricoles, l'INRA compte 8165 agents titulaires et un budget de 881.5 millions d'euros en 2015.

L'INRA PACA, situé dans la région Provence-Alpes-Côte d'Azur, est l'un des 17 centres de recherche régionaux de l'INRA. Le centre est tourné vers l'agro écologie des systèmes de culture sous serres et en vergers et la modélisation de l'impact régionalisé du changement climatique à l'échelle du paysage.

Il existe plusieurs unités dans ce centre, dont PSH (Plantes et Systèmes de culture Horticoles). Cette dernière a pour mission de contribuer à la mise au point de systèmes de culture des fruits et légumes afin d'améliorer la qualité des produits récoltés et le respect de l'environnement. L'objectif principal est de comprendre et quantifier l'impact des pratiques agricoles et des facteurs de l'environnement sur le fonctionnement de la plante et de ses organes, ainsi que sur les populations de bio agresseurs et d'auxiliaires des cultures. Les travaux de recherche combinent des approches expérimentales et de modélisation afin d'atteindre les objectifs. Les objets d'étude sont les cultures de tomate et de laitue en maraichage et les vergers de pêcher et de pommier en arboriculture.

Le stage s'est déroulé en étroite collaboration avec une autre unité, GAFL (Génétique et Amélioration des Fruits et Légumes) du centre INRA PACA.

Table des matières

Chapitre 1

Introduction

1.1 Contexte général

Produire des fruits de qualité en quantité tout en préservant l'environnement est un des grands défis auquel les producteurs fruitiers Européens sont aujourd'hui confrontés. Pour répondre à de telles exigences, la production fruitière intégrée est devenue incontournable. La production intégrée, c'est l'utilisation de techniques alternatives comme la lutte biologique ou l'utilisation de compensations écologiques couplées à des méthodes chimiques.

C'est dans ce contexte que des recherches sont menées par l'INRA pour lutter contre la moniliose, qui est une maladie causée par un champignon qui s'attaque aux fleurs et aux fruits. Le champignon infecte le fruit via les microfissures qui apparaissent sur l'épiderme du fruit. En effet, ces microfissures pouvant atteindre 10% de la superficie des fruits, sont des opportunités d'infection. La moniliose est parmi les maladies les plus dommageables pour les vergers de pêcher, elle peut entraîner des pertes de récolte de pêche allant jusqu'à 40% (Gibert et al., 2009). L'estimation de la conductance de la cuticule des fruits est une bonne évaluation de l'état de la surface du fruit et de la présence de microfissures, portes ouvertes pour le champignon.

A ce jour, on ne dispose que de fongicides pour lutter contre la moniliose et aucun moyen de lutte alternative ne semble pouvoir aboutir rapidement. Ce genre de pratiques pose deux problèmes : environnemental (beaucoup de traitements chimiques) et sanitaire (résidus dus aux traitements proches de la récolte).

Face à ces problèmes, des chercheurs de l'INRA travaillent actuellement sur l'élaboration de nouvelles stratégies de production. Une solution serait de trouver les meilleures combinaisons génotypes et pratiques culturales pour un environnement, donc les recherches visent à optimiser les interactions $G \times E \times P$ (Génotype, Environnement, Pratique culturale).

Afin de faciliter leurs recherches, et de réduire leur coûts d'expérimentation, les chercheurs ont opté pour la modélisation. Ils ont mis au point des modèles de simulation, permettant de suivre la croissance des fruits. Le modèle Fruit Virtuel (FV) (Génard et al., 2007) développé par l'unité PSH, permet de simuler la croissance et la qualité du fruit. Ce modèle couplé avec des algorithmes d'optimisation efficaces peut aider à mettre en place des stratégies de production innovantes permettant de lutter contre la moniliose.

Des études (Quilot-Turion et al., 2012, Memmah et al., 2014) ont été menées pour concevoir des idéotypes variétaux adaptés à des pratiques culturales et à un environnement donné (Avignon, France). Des résultats assez intéressants ont été obtenus. Cepen-

dant, ces études ne prenaient pas en compte d'une manière explicite le contrôle génétique des paramètres du modèle FV. Ces études ont conclu sur l'intérêt de l'intégration de l'information génétique dans la conception assistée par modèles. Une façon de procéder est d'identifier et d'estimer les paramètres dépendants du génotype pour régler le modèle afin qu'il corresponde au mieux aux données disponibles des différents génotypes.

1.2 Objectif du stage et approche

Le stage a porté sur l'un des sous-modèles du Fruit Virtuel, décrivant la conductance cuticulaire du fruit. Il y a 15 paramètres qui interviennent dans ce sous-modèle. A partir de données de 156 génotypes, l'objectif est de déterminer quels sont les paramètres génotype-dépendants/génotype-indépendants, et d'estimer ces paramètres.

Pour ce faire, dans un premier temps nous avons utilisé des méthodes de régression afin de lisser des nuages de points, dans le but d'avoir une relation continue entre des variables du modèle.

La deuxième étape a consisté à effectuer une analyse de sensibilité. Cette dernière permet de décrire les paramètres influents la variable de sortie qui nous intéresse.

La dernière étape a consisté à paramétrer le sous-modèle de la conductance pour différents génotypes. Une approche à plusieurs étapes a été adoptée pour l'estimation des paramètres et définir ceux qui sont génotype-dépendants et génotype-indépendants. D'abord aucune contrainte n'a été définie dans le problème d'optimisation, puis combinée à une analyse de sensibilité, à une analyse de corrélation et à une analyse de variabilité, une nouvelle optimisation a été réalisée avec des paramètres en contraintes. A chaque étape, l'algorithme à évolution différentielle nous a permis d'estimer les paramètres.

Chapitre 2

Modèle et données

2.1 Présentation du modèle Fruit Virtuel

Le modèle Fruit Virtuel a été développé pour simuler la croissance du fruit, le niveau de maturité et l'évolution des teneurs en matière sèche et en ses principaux sucres et acides au cours de la croissance du fruit (Génard et al., 2010).

Il est très utile pour prédire le comportement des cultures dans leur environnement. L'utilisation de ce modèle pour le pêcher a montré certains phénomènes : l'application d'un stress hydrique après une période de bonne irrigation diminue fortement la croissance ou encore un changement de la quantité de feuilles par fruit modifie une part importante du fonctionnement du fruit (Génard et al., 2010).

Le modèle est basé sur différents sous-modèles (Figure 2.1) décrivant l'accumulation de la masse sèche, des sucres et de l'eau dans la pulpe du fruit. Ils représentent aussi des processus tels que la respiration et la transpiration en prenant en compte les microfissures et la conductance de la cuticule du fruit.

A partir des variables d'entrée, le modèle renvoie plusieurs variables de sortie. Il prend en compte les effets de certains facteurs climatiques (rayonnement, température et humidité de l'air), de quelques pratiques culturales (éclaircissage, irrigation), et du génotype sur la croissance et la qualité du fruit. Les entrées du modèle sont le rayonnement, la température quotidienne moyenne, l'humidité relative de l'air et les potentiels hydriques des tiges.

Les valeurs des paramètres ont été définies à partir de la littérature ou estimées selon plusieurs études.

2.2 Présentation du sous-modèle conductance

Le modèle simule la valeur de la conductance totale à un pas de temps quotidien. Cette valeur est simulée à partir du sous-modèle conductance, qui prend comme entrée la masse fraîche du fruit, $M(t)$ (g). La conductance totale, $g(t)$ $(cm.day^{-1})$ pour un temps t, est la somme de trois composantes : conductance stomatique, $g_{sto}(t)$, conductance cuticulaire, $g_{cut}(t)$, conductance des microfissures, $g_{ck}(t)$ (Gibert et al., 2005) :

$$g(t) = g_{sto}(t) + g_{cut}(t) + g_{ck}(t) \tag{2.1}$$

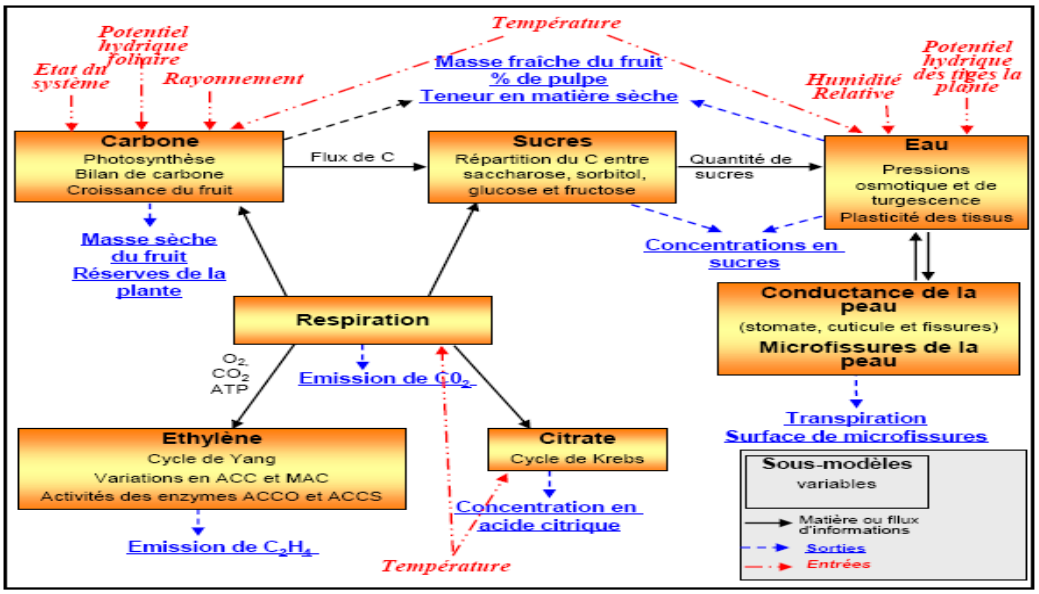

FIGURE 2.1 – Représentation schématique des relations entre les sous-modèles. Adapté de Génard et al. (2010)

La conductance stomatique est obtenue par le produit de la conductance stomatique spécifique, g'_{sto} $(cm^3.day^{-1})$, et du nombre de stomates n divisé par la surface du fruit, $S_f(t)$ (cm^2) :

$$g_{sto}(t) = \frac{n}{S_f(t)} \times g'_{sto} \tag{2.2}$$

La surface du fruit est donnée par la formule :

$$S_f(t) = \gamma \times M(t)^{\eta} \tag{2.3}$$

où γ et η sont des paramètres dépendants de la géométrie du fruit.

La conductance cuticulaire est égale à la proportion de surface de cuticule sur le fruit multipliée par sa conductance spécifique, $g'_{cut}(t)$ $(cm.day^{-1})$:

$$g_{cut}(t) = \frac{S_f(t) - S_{sto} - S_{ck}(t)}{S_f(t)} \times g'_{cut}(t) \tag{2.4}$$

où S_{sto} (cm^2) est la surface occupée par les stomates et $S_{ck}(t)$ (cm^2) la surface des microfissures.

La conductance cuticulaire spécifique est égale à :

$$g'_{cut}(t) = gc_1 e^{-gc_2 t} + \frac{gc_3 e^{-gc_4(t-gc_5)}}{(1 + e^{-gc_4(t-gc_5)})^2} \tag{2.5}$$

avec gc_1 $(cm.day^{-1})$, gc_2 $(cm.day^{-1})$, gc_3 (day^{-1}), gc_4 (day^{-1}), gc_5 (day) les paramètres de la conductance cuticulaire.

Similairement, la conductance des microfissures est égale à la proportion de fissures cuticulaires à l'échelle du fruit multipliée par la conductance spécifique, g'_{ck} $(cm.day^{-1})$:

$$g_{ck}(t) = \frac{S_{ck}(t)}{S_f(t)} \times g'_{ck} \tag{2.6}$$

La surface de microfissures présentes est donnée par :

$$S_{ck}(t) = \sum_{i=1}^{n} \left(\frac{dSnewck(i)}{di} \times (1 - \delta_i)^{t-i} \right) \tag{2.7}$$

où δ_i est le ratio instantané de cicatrisation, égal à :

$$\delta_i = \delta_{h1} e^{\delta_{h2} RI_i} \tag{2.8}$$

avec δ_{h1} (cm^2) et δ_{h2} (sans dimension) les paramètres du processus de cicatrisation, et RI (Ripeness Index) l'indice de maturité calculé chaque jour en divisant la date de jour (DAFB) par la date de récolte (DAFB).

La variable $\frac{dSnewck(i)}{di}$ représente la nouvelle surface de microfissures générée par jour lorsque la différence entre le taux d'expansion de la surface de la pulpe, $\frac{dS_{pulpe}}{dt}$ $(cm^2.day^{-1})$, et le taux d'expansion de la surface de la cuticule, $\frac{dS_{cut}}{dt}$ $(cm^2.day^{-1})$, est positive :

$$\frac{dS_{newck}}{dt} = \begin{cases} \frac{dS_{pulpe}}{dt} - \frac{dS_{cut}}{dt} & \text{si} \quad \frac{dS_{pulpe}}{dt} > \frac{dS_{cut}}{dt} \\ 0 & \text{sinon} \end{cases} \tag{2.9}$$

Le taux d'expansion de la surface de la pulpe est estimé par le taux d'expansion de la surface du fruit, qui est une variable d'entrée.

Le taux d'expansion de la surface de la cuticule est égal au produit de la surface de la cuticule, $S_{cut}(t)$ (cm^2) avec le taux d'expansion relatif de la surface de la cuticule, $RER(t)$ (day^{-1}) :

$$\frac{dS_{cut}}{dt} = S_{cut}(t) \times RER(t) \tag{2.10}$$

Le taux d'expansion relatif de la surface de la cuticule est donné par :

$$RER(t) = -cut_1 M(t) + cut_2 \tag{2.11}$$

où cut_1 (g^{-1}) et cut_2 (sans dimension) sont les paramètres du taux d'expansion de la cuticule.

2.3 Description des données

Cette population est issue d'un croisement entre un pêcher sauvage, *Prunus davidiana* (clone P1908), source de résistance face à plusieurs ravageurs (sharka, oïdium, puceron vert et cloque) mais possédant une très faible valeur agronomique, et Summergrand (SG). Suite à ce croisement, un hybride SD40 présentant un bon niveau de résistance à l'oïdium a été sélectionné pour effectuer un rétrocroisement avec la variété Summergrand donnant lieu à la famille BC1 (Back Cross 1). Un mélange des pollens de cette famille a servi à féconder la variété commerciale Zéphyr (ZE). C'est la population issue de ce dernier croisement qui est appelée BC2 (Back Cross 2) .

L'étude a été réalisée au centre de recherche INRA d'Avignon (sud de la France). Les génotypes BC2 et les trois parents ont été plantés de façon aléatoire dans un verger avec un arbre par génotype. Les arbres avaient 1 an en 1999. Tous les génotypes ont été greffés sur des porte-greffes de GF305.

Diverses expériences ont été menées de 2006 à 2015 sur différents individus de la population à différents stades de croissance des fruits afin d'obtenir des données d'un stade jeune fruit jusqu'à maturité. Une base de données a été créée à partir des mesures pluriannuelles de la conductance de la surface du fruit, de la masse fraiche et de la surface des fruits. Ces données sont disponibles pour 156 génotypes de la population. Cependant, certains génotypes possèdent moins de données que d'autres.

Chapitre 3

Méthodologie et approches proposées

3.1 Lissage

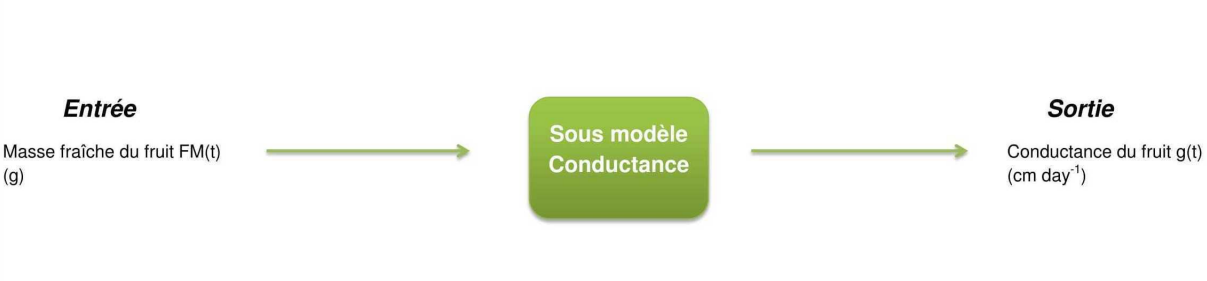

FIGURE 3.1 – Entrée et sortie du sous-modèle conductance

Le sous-modèle conductance du modèle Fruit Virtuel permet de calculer la conductance totale du fruit, en prenant comme entrée la masse fraîche du fruit (Figure 3.1). Cette dernière est simulée à partir d'autres sous-modèles du modèle Fruit Virtuel, pour un certain temps t (DAFB). Lors de l'estimation des paramètres du sous-modèle conductance, nous serons amenés à utiliser ce sous-modèle. Pour ceci, il faut définir une masse fraîche pour chaque temps t, qui est défini à pas de temps quotidien. Or, avec les données expérimentales (Annexe A), il y a plusieurs valeurs de masse fraîche pour une même valeur de temps. C'est pourquoi, nous allons utiliser les méthodes de régression afin de trouver une courbe lisse pour la relation entre la masse fraîche et le temps, et donc pour avoir une seule valeur de la masse fraîche pour un temps donné. Ces méthodes de régression sont également appliquées pour la conductance totale en fonction de la masse fraîche.

Pour la suite du rapport, on considère un échantillon composé des couples (X, Y) tel que $X = (X_1, ..., X_n)^t$ est le vecteur des variables explicatives et $Y = (Y_1, ..., Y_n)^t$ est le vecteur des variables réponses, avec n le nombre d'individus.

En l'absence de toute hypothèse sur la fonction de régression, nous allons utiliser les méthodes de régression non paramétriques :

$$Y_i = f(X_i) + \epsilon_i, \qquad i = 1, ..., n \tag{3.1}$$

où

— f est la fonction de régression,
— les ϵ_i sont les erreurs aléatoires.

La fonction f peut être définie ainsi :

$$f(x) = \mathbb{E}[Y|X = x] \tag{3.2}$$

Le but de la régression non paramétrique est d'estimer la fonction f. Pour ceci, il existe plusieurs méthodes, dans ce qui suit nous présenterons quelques unes.

3.1.1 Régression loess

La régression loess a été introduite par Cleveland (1979) et développée par Cleveland et Devlin (1988). Les moindres carrés pondérés sont utilisés pour obtenir un estimateur de la fonction f.
L'idée de la régression loess consiste à ajuster un modèle de régression défini uniquement dans un voisinage du point d'intérêt. Ce voisinage est déterminé grâce à la méthode des k plus proches voisins. En effet, pour chaque point $X_i, i = 1, ..., n$ du domaine, on choisit les k points qui ont les distances minimales avec ce point. La distance est déterminée par $|X_i - X_j|$, soit la valeur absolue de la différence des valeurs X_i et X_j.
Le poids p d'un point X_j au voisinage de X_i est déterminé par la fonction tricube :

$$w_j(X_i) = \begin{cases} \left(1 - \left(\frac{d_{ij}}{d_{imax}}\right)^3\right)^3 & \text{si} \quad 0 \leq \left(\frac{d_{ij}}{d_{imax}}\right) < 1 \\ 0 & \text{sinon} \end{cases} \tag{3.3}$$

où
— $d_{ij} = |X_i - X_j|$,
— d_{imax} est égale à la distance maximale du point d'intérêt X_i des points appartenants à son voisinage, c'est-à-dire la distance entre X_i et le k-ème point du domaine le plus proche de X_i.

Lors de l'estimation de la fonction de régression, le poids d'un point X_j au voisinage de X_i est d'autant plus important que sa distance à X_i est petite.

Une fois le voisinage et les poids déterminés, les ajustements locaux peuvent être lancés. Pour chaque point du domaine, une régression polynomiale de degré d est effectuée. Cette dernière consiste à lier des variables par un polynôme. Pour un modèle de la forme (3.1), on considère $f(X_i)$ comme un polynôme, on a donc :

$$Y_i = a_0 + a_1 X_i + a_2 {X_i}^2 + a_3 X_i^3 + ... + a_d X_i^d + \epsilon_i, \qquad i = 1, ..., n \tag{3.4}$$

Le but de la régression polynomiale est d'estimer les paramètres $a_0, a_1, ..., a_d$. En utilisant les poids définis précédemment, les polynômes sont ajustés en appliquant la méthode

des moindres carrés pondérés à l'ensemble des points du voisinage. Pour un point X_i, on estime les paramètres $\beta = (\beta_0, ..., \beta_d)$, tels que :

$$\hat{\beta} = \underset{\beta}{argmin} \sum_{j=1}^{k} w_j(X_i)(Y_j - \beta_0 - \beta_1 X_j - ... - \beta_d X_j^d)^2 \tag{3.5}$$

La valeur prédite au point X_i par la régression loess correspond à la valeur prédite par le polynôme ajusté localement. Elle est égale à :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + ... + \hat{\beta}_d X_i^d \tag{3.6}$$

avec $\widehat{\beta} = (\widehat{\beta_0}, ..., \widehat{\beta_d})$.
On effectue cette démarche pour tous les points du domaine.

3.1.2 Régression kernel

Sous le même principe que la régression loess, la régression kernel (Nadaraya, 1964, Watson, 1964), estime la fonction de régression au point X_0 en calculant une moyenne pondérée des observations $Y_i, i = 1, ..., n$. La pondération d'un point X_i dépend du noyau choisi ainsi que de la distance entre X_i et X_0. Un noyau est une fonction K vérifiant les propriétés suivantes :

(i) $K(u) \geq 0$

(ii) $\displaystyle\int_{-\infty}^{+\infty} K(u)\, du = 1$

(iii) $\displaystyle\int_{-\infty}^{+\infty} uK(u)\, du = 0$

Il existe différents noyaux dans la littérature. Voici ceux qui sont utilisés couramment :

Noyau	Formule
Gaussien	$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Epanechnikov	$K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{(\lvert x \rvert \leq 1)}$
Uniforme	$K(x) = \frac{1}{2}\mathbb{1}_{(\lvert x \rvert \leq 1)}$
Triangulaire	$K(x) = (1 - \lvert x \rvert)\mathbb{1}_{(\lvert x \rvert \leq 1)}$

TABLE 3.1 – Définition de différents noyaux

Considérons un réel strictement positif, h, appelé fenêtre. Ce dernier est le paramètre de lissage, dont le choix résulte d'un arbitrage biais/variance, et d'un arbitrage lissage/non lissage de la courbe de lissage.

Les choix du noyau K et de la fenêtre h permettent d'estimer la fonction de régression. Pour un échantillon $(X_1, ..., X_n)$, la fonction de répartition empirique, qui est l'estimateur de la fonction de répartition $G(x)$, est donnée par :

$$\widehat{G}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i \leq x)} \tag{3.7}$$

Sachant que la densité peut être approchée de cette façon,

$$g(x) = G'(x) \approx \frac{G(x+h) - G(x-h)}{2h} \tag{3.8}$$

On en déduit un estimateur de la densité (Rosenblatt,1956),

$$\widehat{g}(x) = \frac{\widehat{G}(x+h) - \widehat{G}(x-h)}{2h} = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}_{(-h < X_i - x \leq h)} \tag{3.9}$$

Plus généralement, l'estimateur à noyau de la densité (Parzen,1962) est défini ainsi :

$$\widehat{g}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{3.10}$$

L'estimateur de Rosenblatt n'est autre qu'un cas particulier de l'estimateur à noyau, avec $K(u) = \frac{1}{2}\mathbb{1}_{(-1<u\leq 1)}$.

En utilisant l'estimateur à noyau de la densité, un estimateur de la fonction de régression f au point X_0 est défini par :

$$\widehat{f}(X_0) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - X_0}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - X_0}{h}\right)} \tag{3.11}$$

Si l'on choisit le noyau gaussien, on donnera un poids à toutes les observations. Ainsi, toutes les observations seront utilisées lors du calcul de l'estimateur au point X_0. Néanmoins les points X_i qui ont une distance très élevée avec le point d'intérêt X_0, auront un poids quasi-négligeable. Cependant, pour un choix d'un autre noyau défini dans le tableau (3.1), seules les observations X_i tels que $|X_i - X_0| \leq h$ seront prises en compte.

Toutefois, le choix du noyau est peu important comparativement au choix de la fenêtre (Hastie et al., 1990).

3.1.3 Spline

Très utilisée dans le domaine de l'analyse numérique, une spline est une fonction définie par des polynômes par morceaux. En statistique, on l'utilise aussi pour lisser un nuage de points. Le principe du spline est de diviser l'intervalle $[a, b]$ où la fonction de spline est définie, en plusieurs sous intervalles $[a, t_1], [t_1, t_2], ..., [t_{k-1}, t_k], [t_k, b]$. Les points

$t_1, ..., t_k$ sont appelés les noeuds. Ils définissent les points de coupure dans lesquels pour chaque intervalle, un polynôme est défini. Le degré des polynômes correspond au degré du spline. Pour certains paramètres $(\beta_1, .., \beta_{m+k})$, la fonction $f(x)$ du modèle (3.1) peut être décomposée (Eubank, 1999) :

$$f(x) = \sum_{j=1}^{m+k} \beta_j B_j(x) \tag{3.12}$$

où $m-1$ est le degré des polynômes par morceaux de la fonction f et les $B_j(x)$ sont des fonctions.

On nomme les fonctions $B_j(x)$, une base de fonction. Un exemple de base de fonction est la base de fonctions de puissance tronquées :

$$B_j(x) = x^{j-1}, j = 1, ..., m \tag{3.13}$$
$$B_j(x) = (x - t_{j-m})_+^{m-1}, j = m+1, ..., m+k \tag{3.14}$$

Ainsi, avec la base de fonctions de puissance tronquées, la formule (3.12) peut s'écrire de cette manière :

$$f(x) = \sum_{j=1}^{m} \beta_j x^{j-1} + \sum_{j=1}^{k} \beta_{m+j}(x - t_j)_+^{m-1} \tag{3.15}$$

Une première possibilité d'estimer la fonction f serait d'estimer les coefficients β_j. Or, en pratique les noeuds t_k ne sont pas connus, ce qui rend l'estimation plus difficile. Néanmoins, il existe des solutions envisageables, avec chacune ses avantages et inconvénients.

Splines de lissage

Dans cette partie, nous allons traiter une méthode permettant d'estimer la fonction de régression f en utilisant les fonctions splines définies précédemment. Cette méthode utilise les valeurs des variables explicatives $X_1, ...X_n$ comme noeuds. Soient les variables $X_1, ...X_n$ appartenant à un intervalle $[a, b]$. Elle détermine une valeur de l'estimateur de la fonction f en minimisant un critère bien précis,

$$\underset{f}{argmin} \left(\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_a^b f^{(m)}(t) dt \right) \tag{3.16}$$

où λ, un réél positif, est le paramètre de lissage.

Ce critère combine la mesure de la qualité de l'ajustement (le premier terme de la somme) et une mesure de lissage (le deuxième terme de la somme). Ce dernier est une pénalité, elle a été introduite par Reinsch (1967). Le paramètre λ établit un équilibre entre ces deux termes.

La solution à ce problème de minimisation est constituée de morceux de polynômes de degré $(2m-1)$ entre les noeuds t_j (Besse et al., 1989). En général on prend $m = 2$, ce qui permet d'obtenir des splines cubiques.
Il est possible de montrer que (3.16) admet une unique solution (Hastie et al., 2009).

Comme la solution est une spline, on peut écrire :

$$f(X_i) = \sum_{j=1}^{n} N_j(X_i)\theta_j \tag{3.17}$$

où $\theta_1, ..., \theta_n$ sont des coefficients et les $N_j(x)$ représentent une base de fonction. Dans ce cas, le critère à minimiser devient :

$$\underset{\theta}{argmin} \left(||Y - N\theta||^2 + \lambda\theta^t\Omega\theta \right) \tag{3.18}$$

avec $\theta = (\theta_1, ..., \theta_n)^t$, les matrices N et $\Omega \in \mathbb{R}^{n\times n}$, avec $N_{ij} = N_j(X_i)$ et $\Omega_{ij} = \int_a^b N_i^{(m)}(t)N_j^{(m)}(t)dt$.

Le but est maintenant d'estimer le paramètre θ. Ceci étant un problème d'une régression Ridge, la solution est :

$$\widehat{\theta} = (N^tN + \lambda\Omega)^{-1}N^tY \tag{3.19}$$

L'estimateur de la fonction de régression f au point X_i par la méthode des splines de lissage est donc donné par,

$$\widehat{f}(X_i) = \sum_{j=1}^{n} N_j(X_i)\widehat{\theta_j} \tag{3.20}$$

ou encore de façon matricielle,

$$\widehat{f} = N\widehat{\theta} = N(N^tN + \lambda\Omega)^{-1}N^tY \tag{3.21}$$

Le principal avantage des splines de lissage est leur critère précis utilisé pour estimer la fonction de régression alors qu'un grand inconvénient de cette méthode est leur manque de généralité au cas multivarié (Hastie et al., 1991).

Régression basis splines

Une autre régression utilisant les splines que l'on peut aborder est le basis spline (ou B-spline). On considère de nouveau les variables $X_1, ..., X_n$ des noeuds. Une base de fonction précise est utilisée, les bases B-splines. Pour ce faire, on doit ajouter des noeuds à l'ensemble existant $t_1, ..., t_k$. Pour un spline de degré $m-1$, on ajoute $2m$ noeuds tel que :

$$t_{-(m-1)} \leq ... \leq t_0 \leq X_{(1)} \tag{3.22}$$

$$X_{(n)} \leq t_{k+1} \leq ... \leq t_{k+m} \tag{3.23}$$

où $X_{(1)} = min(X)$ et $X_{(n)} = max(X)$.

Les valeurs des noeuds supplémentaires sont arbitraires mais un choix judicieux serait de les prendre égaux respectivement à $X_{(1)}$ et $X_{(n)}$ (Hastie et al., 2009).
Soit $B_{i,m}(x)$ la $i^{\text{ème}}$ base d'ordre m de la base de fonction B-splines. Les bases sont définies récursivement :

$$B_{i,1}(x) = \begin{cases} 1 & \text{si} \quad t_i \leq x < t_{i+1} \\ 0 & \text{sinon} \end{cases} \tag{3.24}$$

pour initialiser le calcul avec $i = -(m-1), ..., k$.

$$B_{i,m}(x) = \frac{x - t_i}{t_{i+m-1} - t_i} B_{i,m-1}(x) + \frac{t_{i+m} - x}{t_{i+m} - t_{i+1}} B_{i,m-1}(x) \tag{3.25}$$

pour $i = -(m-1), ..., k$.

Soit la matrice $M = \{B_{j,m}(X_i)\}_{i=1,...,n, j=-(m-1),...,k} \in \mathbb{R}^{n \times (k+m)}$. Ainsi, la fonction de régression f a pour estimateur (Eubank,1999) :

$$\widehat{f}(X_i) = \sum_{j=1}^{m+k} \widehat{\beta}_j B_{j-m,m}(X_i) \tag{3.26}$$

où $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_{m+k})^t$ est une solution des équations normales données par :

$$M^t M \beta = M^t Y \tag{3.27}$$

avec $\beta = (\beta_1, ..., \beta_{m+k})^t$.

Régression P-spline

Un problème majeur de la régression B-spline est le choix du nombre de noeuds k. Un nombre insuffisant de noeuds conduirait à la perte d'informations pertinentes, d'où à une courbe ajustée très lisse, alors qu'un surnombre de noeuds mènerait à un surajustement, donc à une courbe ajustée très flexible (Eilers et al., 2016). Pour remédier à ce problème, O'sullivan (1986) a eu l'idée de prendre un nombre élevé de noeuds et d'introduire un terme de pénalité pour rendre la courbe moins flexible. Ce terme de pénalité ressemble fortement au terme de pénalité de Reinsch pour les splines de lissage.
En s'inspirant de cette idée, Eilers et Marx (1996) proposent une pénalité basée sur les différences d'ordre q des coefficients de la base de fonction B-splines. Donc la régression P-spline est une régression B-spline avec un nombre excessif de noeuds à laquelle on ajoute une pénalité.
Dans ce cas, on souhaite minimiser :

$$\underset{\beta}{argmin} \left(\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{k+m} \beta_j B_{j-m,m}(X_i) \right\}^2 + \lambda \sum_{j=q+1}^{m+k} (\Delta^q \beta_j)^2 \right) \tag{3.28}$$

où λ est le paramètre de lissage et $\Delta^q \beta_j = \Delta(\Delta^{q-1} \beta_j)$ avec $\Delta \beta_j = \beta_j - \beta_{j-1}$.

Si l'on note $D \in \mathbb{R}^{(m+k-q)\times(m+k)}$, la matrice représentant les termes Δ^q, le système d'équation minimisant (3.28) est :

$$M^t Y = (M^t M + \lambda D^t D)\beta \tag{3.29}$$

On en déduit que l'estimateur de f est donné par :

$$\widehat{f}(X_i) = \sum_{j=1}^{m+k} \widehat{\beta}_j B_{j-m,m}(X_i) \tag{3.30}$$

3.1.4 Régression Friedman's supersmoother

Introduit par Friedman (1984), cette régression utilise aussi les régressions locales, mais contrairement à la méthode loess où le degré des polynômes des régressions peut être défini par l'utilisateur, Friedman's supersmoother utilise seulement des régressions locales linéaires (donc des polynômes de degré 1). Un voisinage autour de chaque point doit être déterminé afin d'utiliser les régressions locales. Pour $J \in \mathbb{R}$, le voisinage de X_i, noté $N(X_i)$, est défini en prenant $J/2$ points à gauche et $J/2$ points à droite de X_i. Dans ce cas, l'estimateur de la fonction f au point X_i est donné par :

$$\widehat{f}(X_i) = \widehat{\alpha} + \widehat{\beta} X_i \tag{3.31}$$

où $\widehat{\alpha}$ et $\widehat{\beta}$ sont les paramètres estimés de la régression locale.

Pour estimer ces paramètres, Friedman propose un algorithme :

$$\widehat{\beta} = \frac{C_j}{V_j} \tag{3.32}$$

$$\widehat{\alpha} = \bar{Y_j} - \widehat{\beta}\bar{X_j} \tag{3.33}$$

avec

$$\bar{X_j} = \frac{1}{J} \sum_{X_j \in N(X_i)} X_j$$

$$\bar{Y_j} = \frac{1}{J} \sum_{X_j \in N(X_i)} Y_j$$

$$\bar{C_j} = \sum_{X_j \in N(X_i)} (X_j - \bar{X_j})(Y_j - \bar{Y_j})$$

$$\bar{V_j} = \sum_{X_j \in N(X_i)} (X_j - \bar{X_j})^2$$

3.1.5 Choix du paramètre de lissage

Pour toutes les régressions décrites auparavant, certains choix doivent être faits afin que l'on puisse les utiliser. Ces choix conduisent à un compromis entre le lissage et la flexibilité de l'estimateur. On identifie ceci comme la dualité biais-variance. Augmenter la flexibilité fera diminuer le biais, mais dans ce cas la courbe obtenue sera oscillante, ce qui implique que la variance sera élevée. Donc diminuer la flexibilité entraînera une courbe plus lisse et ainsi une variance faible, mais cette fois c'est le biais qui sera élevé.

Par conséquent, le choix de la valeur du paramètre de lissage est important pour avoir un bon équilibre biais-variance. Une méthode a été mise au point pour permettre de calculer la valeur optimale du paramètre de lissage, la méthode de la validation croisée, donnée par :

$$\underset{\lambda}{argmin} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{f}_{-i}(X_i))^2 \tag{3.34}$$

avec $\widehat{f}_{-i}(X_i)$ la valeur prédite au point X_i en prenant en compte dans la régression toutes les observations sauf le couple (Y_i, X_i).

La valeur optimale est donc le λ minimisant la formule précédente.

3.2 Analyse de sensibilité

De nos jours, beaucoup de modèles sont développés afin de décrire de nombreux phénomènes dans le domaine de la biologie ou de l'agronomie par exemple. Cependant, ces modèles sont parfois difficiles à appréhender. En effet, ils prennent en compte un grand nombre de variables d'entrée. Lorsque nous souhaitons analyser plus en détail ces modèles ou les simplifier, nous pouvons avoir recours à l'analyse de sensibilité. Cette dernière permet d'étudier comment l'incertitude dans la sortie d'un modèle peut être répartie sur différentes sources d'incertitude dans l'entrée du modèle (Saltelli et al., 2008).

L'analyse de sensibilité peut être utilisée pour plusieurs raisons : mieux comprendre la relation entre les variables d'entrée et de sortie, identifier les variables d'entrée les plus/moins influentes, déterminer les variables d'entrée qui intéragissent avec d'autres variables d'entrée. Il est possible de distinguer trois groupes de méthodes d'analyse de sensibilité : les méthodes de screening, qui consistent en une analyse qualitative de la sensibilité de la variable de sortie aux variables d'entrée, les méthodes locales, qui évaluent quantitativement comment de petites perturbations autour d'une valeur de variable d'entrée se répercutent sur la variable de sortie, et les méthodes globales, qui s'intéressent à la variabilité de la sortie du modèle dans l'intégralité de son domaine de variation (Jacques, 2005). La différence entre l'analyse de sensibilité locale et l'analyse de sensibilité globale est que la première s'intéresse à la valeur de la variable de sortie tandis que la deuxième s'intéresse à sa variabilité. Nous allons utiliser trois méthodes d'analyse de sensibilité globale.

On définit pour la suite $Y = f(X)$ un modèle où $X = (X_1, ..., X_k)$ sont les k variables d'entrée indépendantes.

3.2.1 Méthode de Morris

Introduit par Morris (1991), cette méthode s'appuie sur une discrétisation de l'espace des variables, c'est-à-dire que l'exploration de la variable de sortie est réalisée sur une grille régulière des variable d'entrées. Elle fait partie des méthodes OAT (One At a Time), signifiant que les paramètres d'entrée sont variés un seul à la fois. La sensibilité de la sortie à une des variables X_i est mesurée en comparant les résultats où seule cette variable a été variée. Considérons une discrétisation avec p niveaux pour chaque variable X_i. On note Ω la grille de dimension k définie par ces valeurs sélectionnées. On définit le noeud par une valeur quelconque de X dans Ω, et la trajectoire par un ensemble de noeuds tel que seul un paramètre varie entre deux noeuds successifs. Ainsi, une trajectoire comporte $k+1$ noeuds (en comptant le noeud de départ). Le noeud de départ est choisi aléatoirement ainsi que les paramètres qui varient entre deux noeuds successifs. Pour une construction de r trajectoires, le nombre de simulations nécessaire est de $r \times (k+1)$.
L'effet élementaire de la variable X_i sur la trajectoire j est défini par :

$$EE_i^j = \frac{Y(X_1^j, ..., X_{i-1}^j, X_i^j + \Delta, ..., X_k^j) - Y(X_1^j, ..., X_{i-1}^j, X_i^j, ..., X_k^j)}{\Delta} \tag{3.35}$$

où Δ est une valeur dans $\left\{\frac{1}{p-1}, ..., 1 - \frac{1}{p-1}\right\}$ et $X = (X_1^j, ..., X_k^j)$ est une valeur dans Ω telle que $(X + e_i\Delta)$ appartient aussi à Ω pour $i = 1, ..., k$ avec e_i un vecteur composé de 0 sauf la $i^{\text{ème}}$ composante qui est égale à 1 ou -1.

Initialement, deux mesures de sensibilité étaient calculées : μ qui détermine l'influence globale d'une variable d'entrée sur la variable de sortie, et σ, qui détermine les effets non linéaires et les interactions. Cependant, Campolongo et al. (2007) ont proposé une nouvelle mesure, μ^*, qui est une mesure amélirorée de μ car une influence considérable d'une variable peut échapper à μ. La mesure μ^* permet de déterminer la sensibilité de la variable de sortie aux variables d'entrée et de classifier ces derniers par rapport à leur importance sur la variable de sortie. Les formules de ces trois mesures relatives à la variable X_i sont les suivantes :

$$\mu_i = \frac{1}{r} \sum_{j=1}^{r} EE_i^j \tag{3.36}$$

$$\mu_i^* = \frac{1}{r} \sum_{j=1}^{r} |EE_i^j| \tag{3.37}$$

$$\sigma_i = \sqrt{\frac{1}{r-1} \sum_{j=1}^{r} (EE_i^j - \mu_i)^2} \tag{3.38}$$

Pour une variable X_i, les mesures μ_i^* et σ_i sont utilisées afin de tirer des conclusions. Plus μ_i^* est élevée et plus X_i a une influence importante sur la sortie du modèle. Une mesure σ_i élevée implique que la variable X_i a des effets non linéaires ou qu'elle est impliquée dans une interaction avec au moins une autre variable.

3.2.2 Méthode de Sobol

La méthode de Sobol est basée sur une analyse fonctionnelle de la variance. Sobol (1993) propose de décomposer la fonction f du modèle en somme de fonctions de dimension croissante :

$$Y = f(X_1, ..., X_d) = f_0 + \sum_{i=1}^{k} f_i(X_i) + \sum_{1 \leq i < j \leq k} f_{ij}(X_i, X_j) + ... + f_{1,...,k}(X_1, ..., X_k) \tag{3.39}$$

où f_0 est une constante et les fonctions $f_{i_1,...,i_s}, \forall \{i_1, ..., i_s\} \subseteq \{1, ..., k\}$ sont mutuellement orthogonales.

La variance de Y peut alors se décomposer en :

$$Var(Y) = V = \sum_{i=1}^{k} V_i + \sum_{1 \leq i < j \leq k} V_{ij} + ... + V_{1,...,k} \tag{3.40}$$

avec

$$V_i = \mathbb{E}[Y|X_i]$$

$$V_{ij} = \mathbb{E}[Y|X_i, X_j] - V_i - V_j$$

$$V_{ijh} = \mathbb{E}[Y|X_i, X_j, X_h] - V_i - V_j - V_h - V_{ij} - V_{ih} - V_{jh}$$

$$...$$

$$V_{1...k} = V - \sum_{i=1}^{k} V_i - \sum_{1 \leq i < j \leq k} V_{ij} - ... - \sum_{1 \leq i_1 < ... < i_{k-1} \leq k} V_{i_1...i_{p-1}}$$

Ainsi, on peut définir des indices de sensibilité d'ordre 1 et 2 :

$$S_i = \frac{V_i}{V} \tag{3.41}$$

$$S_{ij} = \frac{V_{ij}}{V} \tag{3.42}$$

L'indice S_i représente l'indice de sensibilité de premier ordre de la variable X_i. Il quantifie la sensibilité de la sortie Y à la variable d'entrée X_i seule.

L'indice S_{ij} est l'indice de sensibilité de deuxième ordre des variables X_i et X_j. Il permet d'exprimer la sensibilité de la variance de Y à l'interaction des variables X_i et X_j, c'est-à-dire la sensibilité de Y aux variables X_i et X_j qui n'est pas prise en compte dans l'effet des variables seules.
On peut calculer de la même façon les indices de sensibilité d'ordre plus élevés.

Introduit par Homma et Saltelli (1996), l'indice de sensibilité total d'une variable X_i prend en compte l'influence venant de la variable X_i sous toutes ses formes (la variable seule + toutes les interactions avec les autres variables) :

$$S_{T_i} = \sum_{u \in \#i} S_u \tag{3.43}$$

où $\#i$ représente tous les ensembles d'indices contenant l'indice i.

Pour estimer les indices de sensibilité, la méthode de Monte Carlo est souvent utilisée.

3.2.3 Méthode FAST

La méthode FAST (Fourier Amplitude Sensitivity Testing) est basée sur les principes de l'analyse de Fourier. Similaire à la méthode Sobol, la méthode FAST décompose la variance de Y et utilise des indices de sensibilité. L'idée de cette méthode est d'utiliser la transformée de Fourier pour effectuer la décomposition. Pour $i = 1, ...n$, les variables d'entrée sont transformées en (Chan et al., 1997) :

$$X_i(s) = G_i(sin(\omega_i s)) \tag{3.44}$$

où G_i sont des fonctions à déterminer, et $\omega_i \in \mathbb{N}^*$ correspond à la fréquence de répétition de X_i.

En utilisant les propriétés des séries de Fourier, la variance de Y peut être approchée :

$$Var(Y) \approx 2 \sum_{j=1}^{\infty} ({A_j}^2 + {B_j}^2) \tag{3.45}$$

avec

$$A_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(X(s)) cos(js) ds$$

$$B_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(X(s)) sin(js) ds$$

La part de variance due à la variable X_i est égale à :

$$V_i = 2 \sum_{p=1}^{\infty} ({A_{p\omega_i}}^2 + {B_{p\omega_i}}^2) \tag{3.46}$$

Les valeurs $p\omega_i$ sont appelées des harmoniques.

On définit alors l'indice de sensibilité de X_i par :

$$S_i = \frac{V_i}{Var(Y)} = \frac{\sum_{p=1}^{\infty}({A_{p\omega_i}}^2 + {B_{p\omega_i}}^2)}{\sum_{j=1}^{\infty}({A_j}^2 + {B_j}^2)} \tag{3.47}$$

L'indice de sensibilité total est donné par :

$$S_{T_i} = 1 - \frac{\sum_{p=1}^{\infty}({A_{p\omega_{(-i)}}}^2 + {B_{p\omega_{(-i)}}}^2)}{\sum_{j=1}^{\infty}({A_j}^2 + {B_j}^2)} \tag{3.48}$$

où $\omega_{(-i)}$ sont toutes les fréquences autre que ω_i et ses harmoniques.

Saltelli et Bolado (1998) ont montré que les indices de sensibilité d'ordre 1 des méthodes Sobol et FAST sont équivalents.

Sachant que les sommes dans les indices de sensibilité vont jusqu'à l'infini, lors de l'estimation de ces indices, il convient de choisir une borne supérieure M pour les sommes. Généralement, on prend M égale à 4 ou 6 (Jacques, 2005).

3.3 Estimation

3.3.1 Algorithme à évolution différentielle

L'objectif principal est d'estimer les paramètres du sous-modèle conductance pour différents génotypes. Pour ceci, nous utiliserons l'algorithme à évolution différentielle qui est conçu pour des problèmes d'optimisation non linéaire et avec contraintes. Lorsqu'un paramètre a une valeur identique pour tous les génotypes, il est considéré comme une contrainte. Proposé par Storn et Price (1997), c'est un algorithme d'optimisation basé sur la population, et stochastique. En partant d'une solution initiale, il essaye d'améliorer la solution itérativement.

Généralement dans les problèmes d'optimisation, une fonction objective est désignée. Elle sert de critère pour déterminer la meilleure solution, son choix est donc importante. Nous allons utiliser une fonction qui calcule une somme pondérée de la valeur absolue maximale entre les valeurs prédites à partir des données expérimentales et les valeurs simulées par le sous-modèle conductance. Donc le but est de trouver la matrice M qui minimise :

$$\sum_{i=1}^{N_{gen}} \omega_i max|\widehat{Rho_i} - Rho_i| \tag{3.49}$$

où $M \in \mathbb{R}^{(N_{gen} \times N_p)}$ est la matrice des paramètres estimés, avec N_{gen} le nombre de génotypes et N_p le nombre de paramètres à estimer, ω_i est le poids correspondant au génotype i , Rho_i est le vecteur des valeurs prédites de la conductance pour le génotype i et $\widehat{Rho_i}$ est le vecteur des valeurs simulées de la conductance à partir du sous-modèle pour le génotype i.

Les principaux avantages de cet algorithme sont : la possibilité d'être utilisé pour des modèles non linéaires, peu de temps de calcul comparé aux autres algorithmes, simple utilisation et puissant pour converger vers le minimum global (Storn et Price, 1997).

L'évolution différentielle utilise une population de N vecteurs de dimension d pour chaque génération G, où d correspond aux nombre de paramètres :

$$X_{i,G} = (X_{1,i,G}, X_{2,i,G}, ..., X_{d,i,G}), \qquad i = 1, ..., N. \tag{3.50}$$

Après avoir calculer une population initiale, l'algorithme applique successivement trois opérations (mutation,croisement,sélection) sur chaque vecteur. Les détails de l'algorithme sont données ci-dessous :

— **initialisation**
Pour chaque paramètre, une borne inférieure et supérieure doivent être spécifiées. Puis, une population initiale est générée par tirage aléatoire uniforme sur l'ensemble des valeurs possibles de chaque paramètre.

— **mutation**
Pour chaque vecteur $X_{i,G}, i = 1, ..., N$, on génére un vecteur mutant :

$$V_{i,G+1} = X_{r_1,G} + F \cdot (X_{r_2,G} - X_{r_3,G}) \tag{3.51}$$

avec $F \in [0, 2]$, $r_1, r_2, r_3 \in \{1, ..., N\}$ des entiers distincts et différents de l'indice i. A cause de cette condition, la taille de la population N doit être au moins égale à 4. F est une constante réelle qui contrôle l'amplification de la variation différentielle $(X_{r_2,G} - X_{r_3,G})$ (El Dor, 2012).

— **croisement**
Cette étape permet d'augmenter la diversité des vecteurs. Un vecteur d'essai $U_{i,G+1}$ est créé à partir des éléments des vecteurs $X_{i,G}$ et $V_{i,G+1}$, pour $j = 1, ..., d$:

$$U_{j,i,G+1} = \begin{cases} V_{j,i,G+1} & \text{si} \quad (rand(j) \leq CR) \quad \text{ou} \quad (j = rnbr(i)) \\ X_{j,i,G} & \text{si} \quad (rand(j) > CR) \quad \text{et} \quad (j \neq rnbr(i)) \end{cases} \tag{3.52}$$

où $rand(j)$ est la $j^{\text{ème}}$ valeur d'un générateur de nombre aléatoire uniforme appartenant à $[0, 1]$, $rnbr(i) \in \{1, ..., d\}$ est un indice choisi aléatoirement qui permet de garantir qu'au moins un élément du vecteur mutant sera conservé, et $CR \in [0, 1]$ est le coefficient de croisement déterminé par l'utilisateur.

— **sélection**
Le vecteur $X_{i,G}$ est comparé au vecteur $U_{i,G+1}$ et celui qui a la valeur de fonction objective la plus petite sera conservé pour la prochaine génération :

$$X_{i,G+1} = \begin{cases} U_{i,G+1} & \text{si} \quad f(U_{i,G+1}) < f(X_{i,G}) \\ X_{i,G} & \text{sinon} \end{cases} \tag{3.53}$$

Ces quatre étapes sont réitérées jusqu'à ce qu'un nombre maximal d'itérations soit atteint.
L'algorithme DE utilisé est celui développé par Conceicao et Maechler (2016) et disponible dans le cadre du package R « DEoptimR ».

3.3.2 Mesures de performance

Pour mesurer la qualité de l'estimation, plusieurs indicateurs peuvent être utilisés en comparant les valeurs observées et simulées (Moriasi et al., 2007). Quelques uns sont présentés ci-dessous :

— Erreur absolue moyenne :

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i| \tag{3.54}$$

— Erreur quadratique moyenne :

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{3.55}$$

— Racine carré de l'erreur quadratique moyenne :

$$REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2} \tag{3.56}$$

— Racine carré relative de l'erreur quadratique moyenne :

$$RREQM = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}}{\bar{Y}} \tag{3.57}$$

— Coefficient de corrélation Pearson :

$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)(\widehat{Y}_i - \bar{\widehat{Y}}_i)}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)\sum_{i=1}^{n}(\widehat{Y}_i - \bar{\widehat{Y}}_i)}} \tag{3.58}$$

— Indice d'accord :

$$d = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(|\widehat{Y}_i - \bar{Y}| + |Y_i - \bar{Y}|)^2} \tag{3.59}$$

— Critère de Nash-Sutcliffe :

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3.60}$$

Chapitre 4

Résultats

4.1 Résultats des ajustements de courbes

4.1.1 Résultats avec les paramètres optimaux

Comme mentionné dans le chapitre précédent, le but de cette partie est de trouver une relation permettant de calculer la masse fraîche pour un certain temps t et la conductance totale pour une masse fraîche donnée sur la base de données observées pour chacun des génotypes de la population. Le temps t, exprimé en jours, correspond au DAFB (Day After Full Bloom), soit le nombre de jours après la pleine floraison. Nous allons utiliser différentes méthodes de régression, sur les valeurs observées des 156 génotypes, afin de parvenir au but. Pour chaque génotype et pour chaque méthode, deux régressions sont effectuées : les valeurs observées de la masse fraîche en fonction des valeurs observées du DAFB et les valeurs observées de la conductance totale en fonction des valeurs observées de la masse fraîche.

L'objectif est de trouver une méthode similaire pour tous les génotypes, malgrè que les nuages de points de ces derniers soient très différents (Annexe A), ce qui rend la tâche difficile. Une deuxième complexité est le faible nombre d'observations pour certains génotypes. Certains n'ont en effet que 5 observations. Néanmoins, dans une première étape, tous les génotypes ont été inclus.

Toutes les méthodes décrites dans le Chapitre 3 ont été utilisées grâce au logiciel R. Les fonctions « loess() », « ksmooth() », « smooth.spline() », « bs() », « pb() », « supsmu() » du logiciel permettent d'utiliser repectivement la méthode loess, la méthode kernel, la méthode splines de lissage, la méthode basis spline, la méthode P-spline et la méthode Friedman's supersmoother. Sachant que pour toutes ces méthodes, un paramètre de lissage doit être spécifié afin que le lissage soit effectué, et n'ayant aucune idée à priori sur la valeur à choisir, nous avons décidé d'utiliser la méthode de la validation croisée pour obtenir le paramètre optimal.

Les résultats pour le génotype Zéphir (le génotype ayant le plus grand nombre d'observations) figurent ci dessous. La figure (4.1) montre les résultats de la conductance en fonction de la masse fraîche tandis que la figure (4.2) montre les résultats de la masse fraîche en fonction du DAFB. Seules les fonctions « smooth.spline() » et « supsmu() » calculent la validation croisée par défaut, pour les autres fonctions elle doit être calculée manuellement avec la formule définie dans la section (3.1.7).

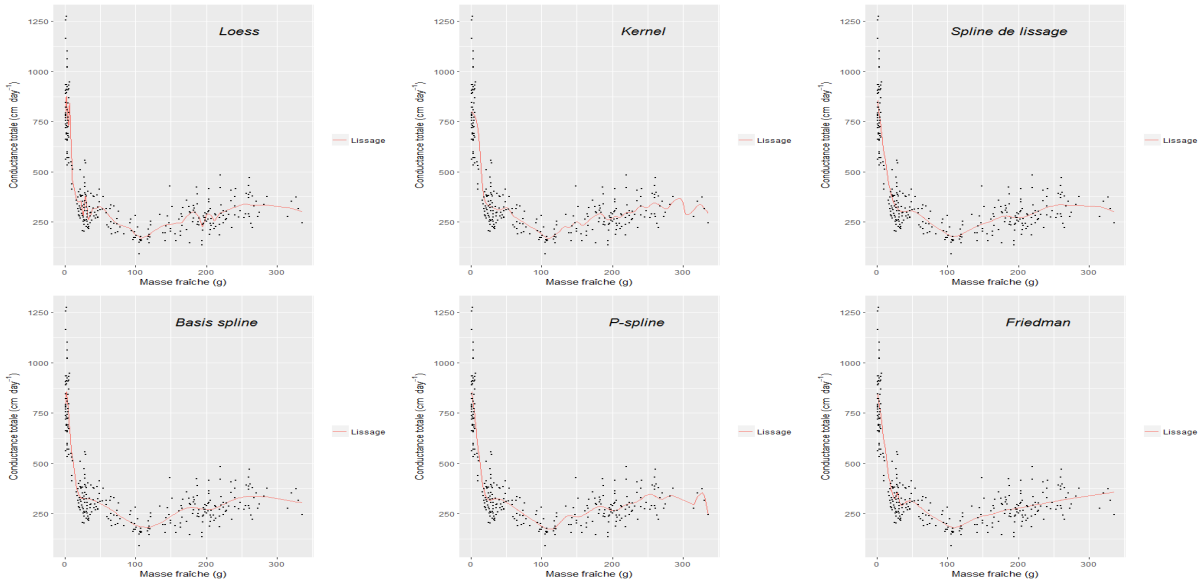

FIGURE 4.1 – Valeurs observées de la conductance totale en fonction de la masse fraîche et lissage avec chaque méthode pour le génotype Zéphir

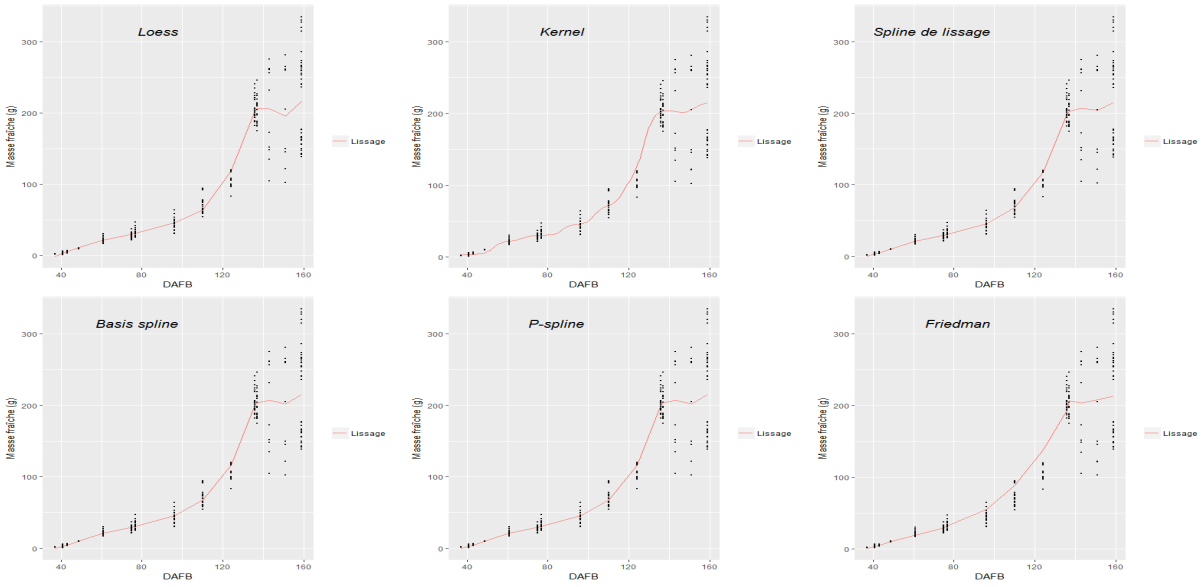

FIGURE 4.2 – Valeurs observées de la conductance totale en fonction de la masse fraîche et lissage avec chaque méthode pour le génotype Zéphir

Pour chaque méthode, les résultats sont différents pour les 156 génotypes. La variabilité du paramètre optimal pour chaque méthode est flagrante (4.3), ce qui nous empêche de fixer une même valeur de paramètre de lissage pour tous les génotypes. C'est pourquoi, nous avons mis en place une procédure afin de trouver une méthode avec un paramètre de lissage fixe qui pourrait convenir à l'ensemble des génotypes.

4.1.2 Sélection d'une méthode

Notre intention étant de trouver une méthode similaire pour tous les génotypes, nous avons défini deux critères de sélection : l'allure de la courbe, qui ne doit pas avoir de

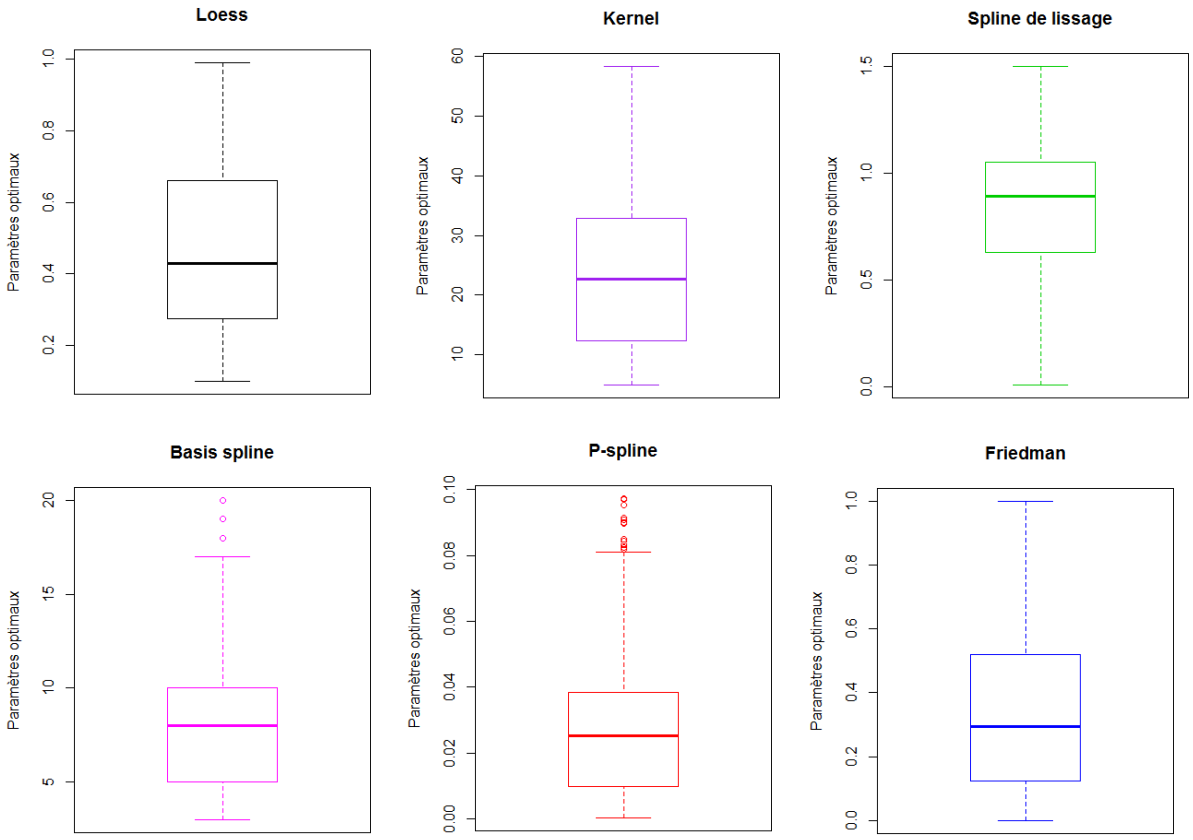

FIGURE 4.3 – Boxplots des paramètres optimaux des 156 génotypes pour chaque méthode

fluctuations, et la qualité de la prédiction. Cette dernière peut être mesurée avec un indicateur statistique très utilisé, le coefficient de détermination (R^2) :

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{4.1}$$

avec Y_i les observations, $\widehat{Y}_i$ les valeurs prédites et $\bar{Y}$ la moyenne des observations.

Ces deux critères vont nous permettre de comparer toutes les méthodes et de choisir la méthode qui convient le mieux à l'ensemble des génotypes. Mis à part le paramètre de lissage de la méthode basis spline qui appartient à l'ensemble des entiers naturels ($\mathbb{N}$), les paramètres de lissage appartiennent à l'ensemble des nombres réels positifs ($\mathbb{R}^+$). Donc pour définir avec quelles valeurs de paramètre de lissage les R^2 seront calculés, nous déterminons un intervalle commun à l'ensemble des génotypes pour chaque méthode et fixons des valeurs par pas réguliers de la borne inférieure à la borne supérieure de l'intevalle. Ces bornes correspondent aux valeurs extrêmes observées pour les paramètres optimaux. Pour plus de simplicité, ces valeurs sont arrondies. Concernant le pas, c'est un compromis entre précision et nombre de calculs de R^2. Le tableau (4.1) présente un récapitulatif des bornes et des pas choisis.

Méthode \ Intervalle	Borne inférieure	Borne supérieure	Pas
Loess	0.1	1	0.01
Kernel	5	59	0.5
Spline de lissage	0	1.5	0.01
Basis spline	3	20	1
P-spline	0	0.1	0.001
Friedman	0	1	0.01

TABLE 4.1 – Intervalles dans lesquels les paramètres de lissage sont fixés

Une fois les valeurs des paramètres de lissage fixées, nous procédons aux calculs du R^2. Sachant que pour une méthode et pour un paramètre de lissage donnés, il y a 156 valeurs de R^2 calculées (un par génotype), le R^2 moyen, valeur moyenne des 156 R^2, est calculé. Les résultats sont présents sur la figure (4.4).

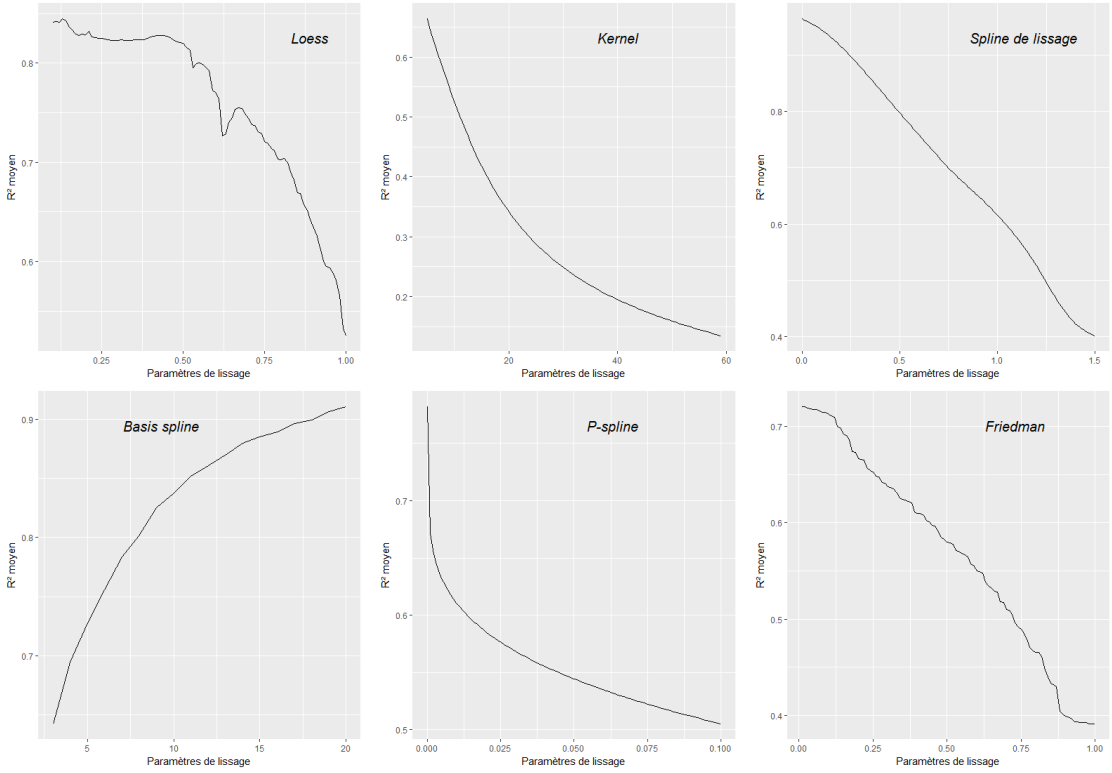

FIGURE 4.4 – R^2 moyen en fonction des paramètres de lissage pour toutes les méthodes

Excepté celle de la méthode basis spline, toutes les courbes sont décroissantes, ceci vient du fait que plus le paramètre de lissage augmente plus la courbe est lisse et donc plus la qualité de la prédiction est faible. C'est aussi valable pour la méthode basis spline mais dans le sens inverse, c'est-à-dire plus le paramètre est petit plus la courbe est lisse. Au contraire, pour de petites valeurs de paramètres (pour de grandes valeurs pour le basis spline), les courbes sont oscillantes, c'est pourquoi l'idée de prendre le paramètre correspondant au $R^2{}_{max}$ est écartée. Un compromis doit être trouvé. L'observation des ajustements nous a mené à fixer un seuil de R^2 de 0.7 au dessous duquel les ajustements

n'étaient pas satisfaisants. Ainsi, seuls les méthodes et paramètres aboutissant sur un R^2 moyen supérieur à 0.7 ont été conservés pour l'étape suivante.

Les méthodes loess, spline de lissage et basis spline ont beaucoup de paramètres supérieurs au seuil $R^2 = 0.7$, alors que les méthodes P-spline et Friedman ont seulement quelques paramètres supérieurs au seuil. Comme la méthode kernel n'en a aucun, elle est mise de côté après avoir vérifié que les plus petits paramètres (qui auront donc un grand R^2) ne conviennent pas.

Les paramètres correspondants au critère de qualité de l'ajustement étant établis, la deuxième phase est d'éliminer les paramètres pour lesquels des oscillations apparaissent. Nous avons essayé d'écrire un algorithme qui détecterait les oscillations pour une courbe donnée. Sachant que pendant l'ajustement, pour certaines valeurs X_i, on calcule des valeurs prédites $\widehat{Y}_i$, l'idée était de prendre la valeur absolue de la différence entre 2 pas de ces valeurs prédites, c'est-à-dire $\delta(i) = |\widehat{Y_{i+1}} - \widehat{Y}_i|$ pour $i = 1, ..., n-1$ et de trouver un algorithme à partir des $\delta(i)$ afin d'identifier les courbes oscillantes. Malheureusement cette idée n'a pas abouti.

Nous avons donc été contraints de classer les ajustements à l'oeil nu. Tout d'abord pour chaque méthode, une limite a été définie. Cette dernière est une valeur de paramètre de lissage telle que seules les valeurs supérieures (inférieures pour le basis spline) sont tolérées. Les autres valeurs résultaient sur des courbes oscillantes.

Pour résumer, nous avons défini deux limites pour chaque méthode : la première est obtenue avec le R^2, les paramètres dont la régression a un $R^2 \geq 0.7$ sont gardés, alors que la deuxième est obtenue grâce au critère d'oscillation : les paramètres, dont la courbe d'ajustement est satisfaisante, sont retenus.

Après croisement des deux limites, les méthodes P-spline et Friedman's supermoother sont écartées. Il reste les méthodes loess, spline et basis spline avec environ 10 paramètres chacune. En comparant les trois méthodes, nous avons décidé d'éliminer 53 génotypes du fait de leur faible nombre d'observations et de leurs profils atypiques. Donc il ne reste plus que 103 génotypes.

Enfin, c'est la méthode basis spline avec un paramètre de lissage égal à 4 qui a été conservé pour les deux régressions de la masse fraîche en fonction du DAFB et de la conductance en fonction de la masse fraîche. Néanmoins, pour certains génotypes avec beaucoup d'observations, nous avons décidé de prendre un paramètre de lissage égal à 5 afin d'obtenir une courbe plus lisse pour la régression de la conductance en fonction de la masse fraîche.

Les résultats pour le génotype Zéphir sont illustrés sur la figure (4.5). Les résultats pour quelques autres génotypes sont présents en Annexe B.

4.2 Résultats de l'analyse de sensibilité

La sensibilité du modèle Fruit Virtuel aux variations des paramètres a été analysée en utilisant les trois méthodes décrites au chapitre 3. L'objectif de cette étape est de déterminer quels sont les paramètres influents/non influents sur la conductance totale. Le modèle Fruit Virtuel compte une soixantaine de paramètres mais seuls les 15 paramètres

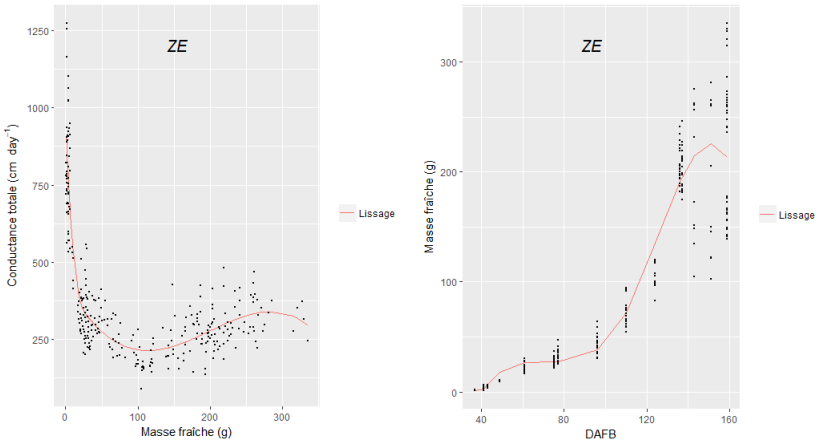

FIGURE 4.5 – Courbes de lissage obtenues avec la méthode choisie pour le génotype Zéphir

qui interviennent dans le sous-modèle conductance sont étudiés ici (Table 4.2). Le modèle simule une valeur de la sortie (conductance totale) pour un pas de temps quotidien. Par conséquent, l'analyse de sensibilité a été réalisée d'une manière dynamique à 8 dates (DAFB) au cours du développement du fruit, pour prendre en compte cette cinétique.

Le package « sensitivity » du logiciel R (Pujol, 2017) a été utilisé pour réaliser l'analyse de sensibilité avec trois méthodes : Morris, Sobol et FAST.

Paramètre	Valeur nominale	Unité
g'_{sto}	0.01599555	$cm.day^{-1}$
gc_1	860.1175	$cm.day^{-1}$
gc_2	0.01786	$cm.day^{-1}$
gc_3	692.5507	day^{-1}
gc_4	0.159	day^{-1}
gc_5	86.7248	day
g'_{ck}	3838.86	$cm.day^{-1}$
δ_{h1}	2	cm^2
δ_{h2}	5.22	sans dimension
cut_1	-0.0035	g^{-1}
cut_2	1.330164	sans dimension
η	0.601	sans dimension
γ	6.049	sans dimension
n	70925	sans dimension
$Asto$	0.15247872	sans dimension

TABLE 4.2 – Valeurs nominales des paramètres du sous modèle conductance

4.2.1 Morris

Etant donné qu'un intervalle doit être spécifié pour chaque paramètre, nous décidons de prendre comme bornes 15% de leurs valeurs nominales. La figure (4.6) montre les résultats obtenus en termes d'influence/non influence des 15 paramètres sur la conductance.

L'influence d'un paramètre peut évoluer au cours du temps. En effet, le paramètre cut_2 a par exemple une petite influence dans les premiers jours alors que dans les derniers jours son influence est la plus importante. Globalement, le paramètre gc_5 semble être le paramètre le plus influent sur la conductance. Toutefois, dans les derniers jours de simulation, son influence diminue. Les paramètres gc_2 et gc_1 ont une influence moyenne sur la sortie. Les autres paramètres ont peu d'influence, voire aucune pour les paramètres δ_{h1} et δ_{h2}.

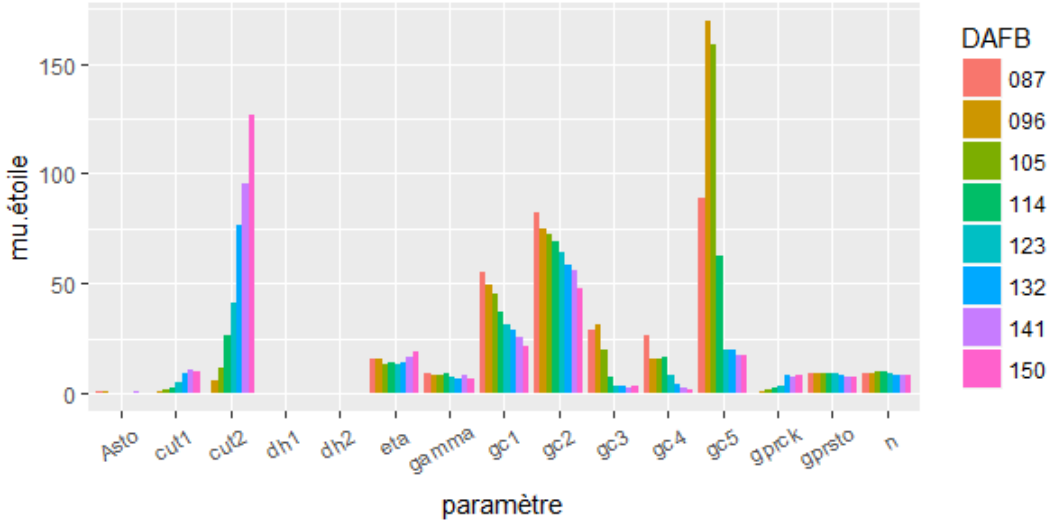

FIGURE 4.6 – Résultats des valeurs de μ^* des paramètres pour différentes valeurs de temps

Les effets non linéaires et les interactions sont fournis sur la figure (4.7). On trouve quasiment les mêmes résultats que sur la figure (4.6). Au début de la simulation, le paramètre gc_5 semble avoir les effets non linéaires ou les interactions les plus importants. A la fin de la simulation, c'est le paramètre cut_2 qui prend le relais. Les paramètres δ_{h1} et δ_{h2} ont des valeurs nulles.

4.2.2 Sobol

Pour cette méthode, nous prenons à nouveau pour chaque paramètre un intervalle avec des bornes de 15% des valeurs nominales. Nous fixons à 100 le nombre d'échantillons pour la technique de bootstrap.

La figure (4.8) montre les estimations des indices de sensibilité totale des paramètres pour plusieurs pas de temps. La figure (4.9) nous donne un aperçu de la contribution des paramètres à la variance de la sortie pour quatre pas de temps. Le paramètre gc_5 est le paramètre le plus influent au début de la simulation. En effet, il contribue à 57.2% de la variance de la sortie lors du premier jour de la simulation. Mais son influence diminue au cours du temps, il ne contribue plus à la fin de la simulation. Au contraire, le paramètre cut_2 n'est pas important au début de la simulation, alors qu'à la fin il contribue à la majorité (86.9%) de la variance de la sortie. Le paramètre gc_2 reste influent tout au long de la simulation, il est même le paramètre qui contribue le plus au $114^{\text{ème}}$ DAFB. Le

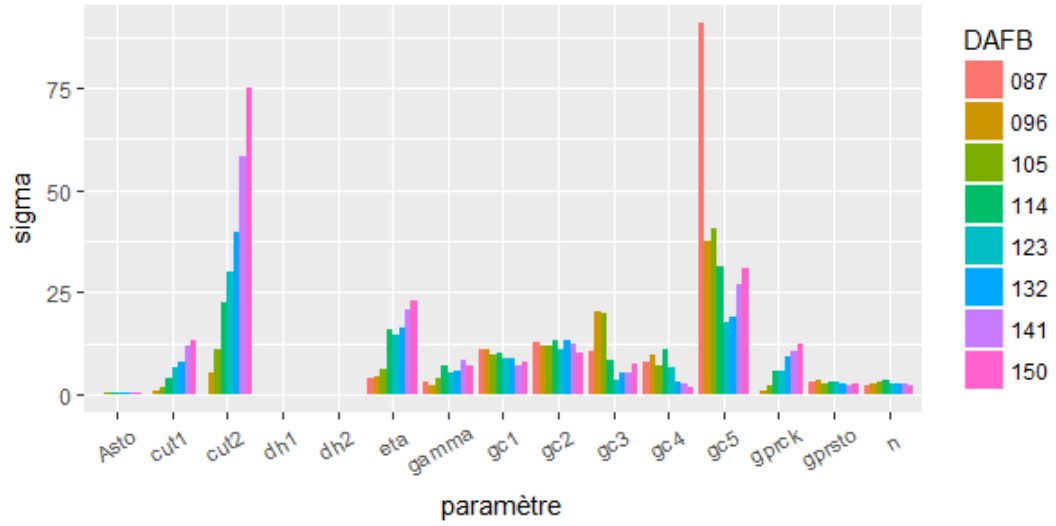

FIGURE 4.7 – Résultats des valeurs de σ des paramètres pour différentes valeurs de temps

paramètre gc_1 a une influence moins importante. Les paramètres $Asto$, δ_{h1} et δ_{h2} n'ont aucune influence sur la sortie.

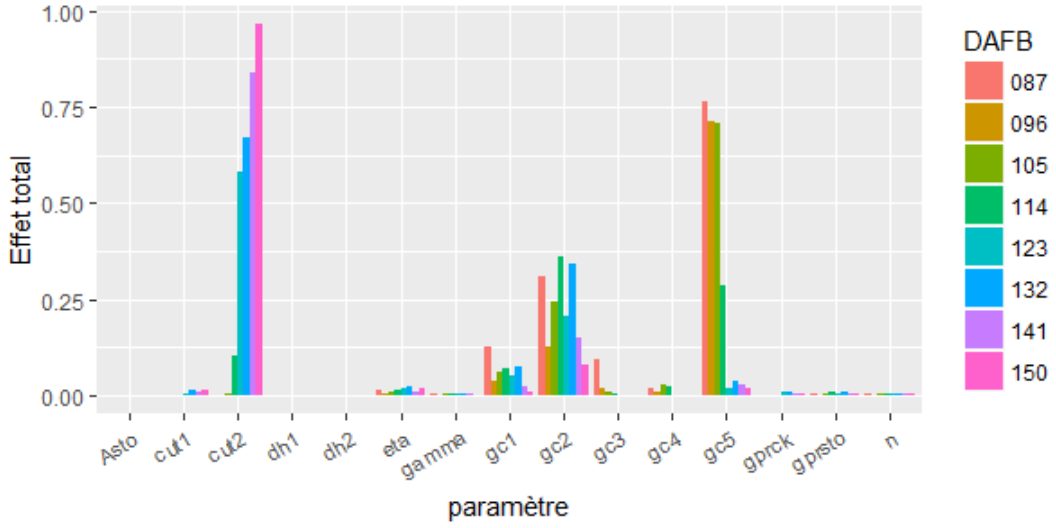

FIGURE 4.8 – Effets totaux des paramètres pour différents pas de temps

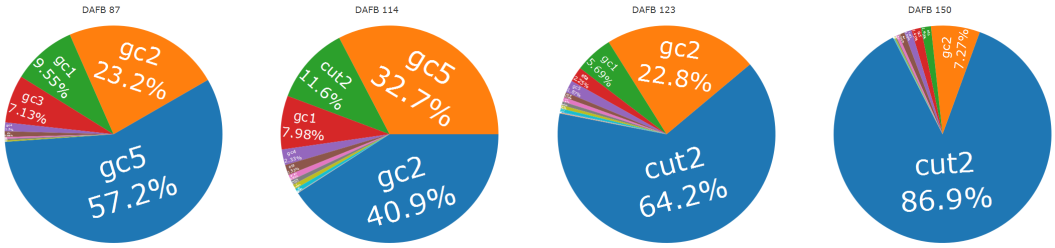

FIGURE 4.9 – Contribution des paramètres à la variance de la sortie pour quatre pas de temps

4.2.3 Fast

Nous avons fixé M, la borne supérieure des sommes des formules (3.47) et (3.48), à 4.

Les résultats sont similaires aux 2 précédentes méthodes (Figure 4.10). Le paramètre gc_5 est très influent dans la première partie de la simulation, puis devient moins influent. Contrairement à gc_5, le paramètre cut_2 n'influe pas la sortie au début de la simulation, mais devient influent à la fin. Le paramètre gc_2 est influent tout au long de la simulation. Le paramètre gc_1 a peu d'influence sur la sortie.

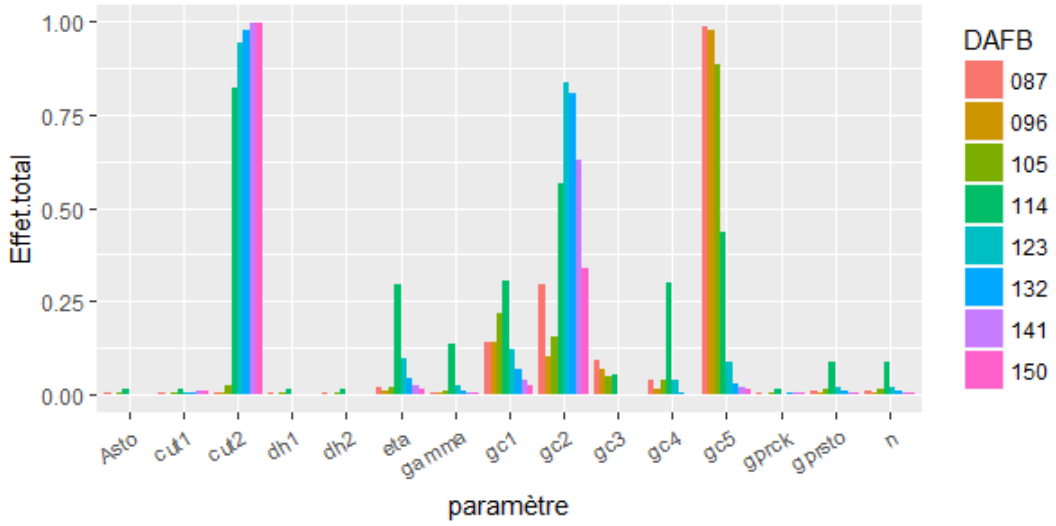

FIGURE 4.10 – Effets totaux des paramètres pour différents pas de temps

4.2.4 Synthèse

Trois méthodes d'analyse de sensibilité globale ont été utilisées afin de comparer les résultats. Ces derniers sont similaires, les paramètres gc_5 et cut_2 sont les plus influents respectivement au début et à la fin de la simulation. Les paramètres gc_2 et gc_1 ont une influence moyenne stable tout au long de la simulation. Tandis que les paramètres δ_{h1} et δ_{h2} ont une influence très petite sur la sortie (voire aucune d'après les méthodes Morris et Sobol).

4.3 Résultats des estimations

Nous avons utilisé un algorithme à évolution différentielle incrémenté via le package « DEoptimR » du logiciel R (Conceicao et al., 2016) pour calibrer le sous-modèle conductance. Il s'agissait, à partir des résultats expérimentaux, de déterminer la valeur des paramètres du modèle pour chaque génotype. De façon à simplifier le modèle, nous avons ensuite identifié les paramètres indépendants des génotypes qui sont alors fixés à leur valeur nominale.

Les paramètres η, γ, n et $Asto$ sont déterminés expérimentalement pour chaque génotype, par conséquent ils ne sont pas pris en compte pour l'estimation. Donc l'optimisation est basée sur les 11 paramètres restants. Comme 53 génotypes ont été mis de côté lors de la phase de lissage, les 103 génotypes restants sont inclus pour l'estimation.

Dans la première étape, tous les paramètres sont considérés dépendants des génotypes. Donc pour chaque génotype, nous estimons les 11 paramètres indépendamment des autres génotypes. Ensuite, la variabilité de tous ces paramètres estimés est examinée afin de fixer ceux qui ne présentent pas de variabilité entre différents génotypes. Si aucun paramètre n'est fixé, nous analysons la corrélation entre les paramètres. Si 2 paramètres ont une corrélation forte, nous fixons le paramètre le moins influent sur la sortie (déterminé grâce à l'analyse de sensibilité).

Une configuration de la fonction « JDEoptim » du package doit être effectuée afin de lancer les calculs d'estimation. Les bornes des intervalles que l'algorithme peut explorer pour chaque paramètre sont définies par plus ou moins 80% de la valeur nominale de chaque paramètre. La taille de la population ainsi que le nombre maximal de générations sont fixés à 1 000. Comme c'est un algorithme stochastique, on répéte cette optimisation 10 fois dans le but d'avoir un ensemble de 10 estimations des paramètres pour chaque génotype. Grâce aux mesures de performance décrites à la section (3.3.2), le meilleur ensemble de paramètres est sélectionné (dite 'meilleure solution').

4.3.1 Etape 1

Comme tous les paramètres sont considérés dépendants des génotypes, le problème d'optimisation est sans contrainte. Le but est de trouver la matrice M qui minimise :

$$\sum_{i=1}^{N_{gen}} \frac{N_i}{N_{tot}} max|\widehat{Rho_i} - Rho_i| \tag{4.2}$$

où $M \in \mathbb{R}^{(N_{gen} \times N_p)}$ avec N_{gen} le nombre de génotypes et N_p le nombre de paramètres à estimer, N_i est le nombre d'observations pour le génotype i, N_{tot} est le nombre d'observations total, Rho_i est le vecteur des valeurs prédites de la conductance pour le génotype i et $\widehat{Rho_i}$ est le vecteur des valeurs simulées de la conductance à partir du modèle pour le génotype i.

Afin d'étudier la variabilité des paramètres entre les différents génotypes, une analyse de la variance (ANOVA) a été utilisée. Les p-valeurs correspondants à chaque paramètre figurent dans le tableau (4.3). Comme les paramètres δ_{h1} et δ_{h2} ont une p-valeur supérieure à 0.05, ils sont considérés génotype-indépendants. Tous les autres paramètres restent génotype-dépendants.

Paramètre	P-valeur
g'_{sto}	<2e-16
gc_1	<2e-16
gc_2	<2e-16
gc_3	<2e-16
gc_4	<2e-16
gc_5	<2e-16
g'_{ck}	<2e-16
δ_{h1}	0.916
δ_{h2}	0.988
cut_1	<2e-16
cut_2	<2e-16

TABLE 4.3 – Les p-valeurs des paramètres de l'ANOVA

La variabilité des paramètres intra et inter génotype peut aussi être vérifier graphiquement (Figure 4.11). Seuls les paramètres *gprimesto*, δ_{h1} et δ_{h2} sont présents sur la figure (les autres paramètres sont en Annexe C). Les résultats du test de l'ANOVA sont confirmés, la variabilité des paramètres δ_{h1} et δ_{h2} au sein des génotypes est aussi grande que celle entre les génotypes (paramètres $dh1$ et $dh2$ sur la figure 4.11).

Sachant que nous avons un ensemble de 10 estimations des paramètres pour chaque génotype, nous devons seulement garder la meilleure estimation pour chaque génotype. Pour ceci, nous nous sommes basés sur les sept mesures de performance. Nous avons effectué un classement des estimations pour chaque mesure de performance puis nous avons sélectionné l'estimation ayant le premier rang sur la majorité des mesures de performance. La figure (4.12) montre les valeurs simulées de la conductance avec la meilleure estimation pour six génotypes.

Ainsi, comme les paramètres δ_{h1} et δ_{h2} sont considérés génotype-indépendants, nous avons choisi d'estimer une seule valeur pour l'ensemble des génotypes. Par conséquent, ces deux paramètres ont été définis comme des contraintes.

4.3.2 Etape 2

Comme il y a deux contraintes dans cette étape, le problème d'optimisation diffère un peu de l'étape précédente. Nous souhaitons trouver la matrice M qui minimise :

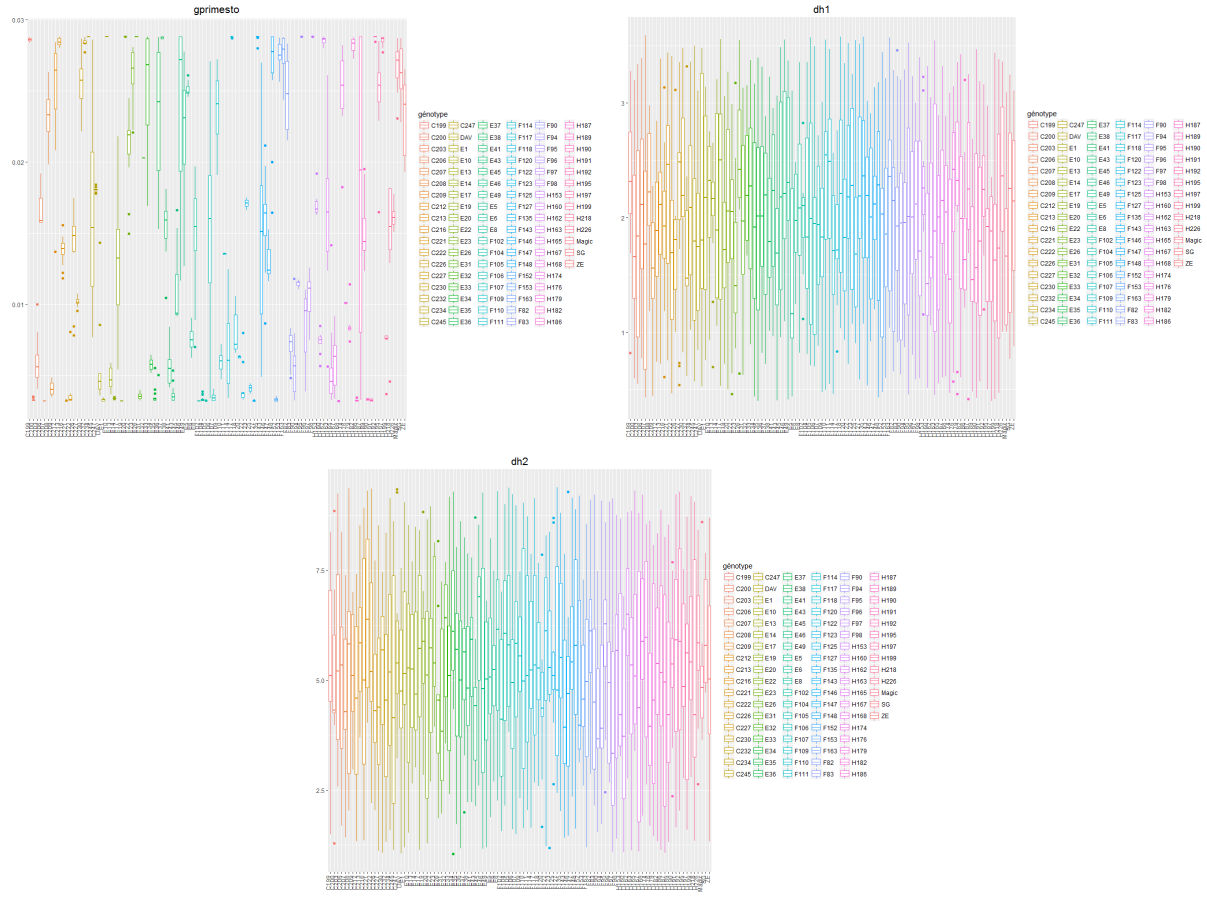

FIGURE 4.11 – Variabilité intra et inter génotype des paramètres

$$\begin{cases} \sum_{i=1}^{N_{gen}} \frac{N_i}{N_{tot}} max|\widehat{Rho_i} - Rho_i| \\ \delta_{h1_i} = \delta_{h1_j} \\ \delta_{h2_i} = \delta_{h2_j} \qquad \forall i \neq j \end{cases} \tag{4.3}$$

où δ_{h1_i} représente la valeur de δ_{h1} pour le génotype i, $M \in \mathbb{R}^{(N_{gen} \times N_p)}$ avec N_{gen} le nombre de génotypes et N_p le nombre de paramètres à estimer, N_i est le nombre d'observations pour le génotype i, N_{tot} est le nombre d'observations total, Rho_i est le vecteur des valeurs prédites de la conductance pour le génotype i et $\widehat{Rho_i}$ est le vecteur des valeurs simulées de la conductance à partir du modèle pour le génotype i.

On débute de la même façon que l'étape précédente, à savoir le test de l'ANOVA. D'après le tableau (4.4), aucun paramètre est indépendant des génotypes.

La matrice de corrélation (Table 4.5) fait ressortir une corrélation forte et significative entre les paramètres gc_1 et gc_2. Une ACP (Analyse en composantes principales) normée a été réalisée afin de vérifier cette corrélation (Figure 4.13). Etant données que les flèches correspondants aux deux paramètres sont proches du cercle de l'unité et que l'angle de ces 2 flèches est faible, on en conclut que l'ACP confirme la corrélation forte entre ces deux paramètres. D'après les résultats de l'analyse de sensiblité, le paramètre $gc1$ est

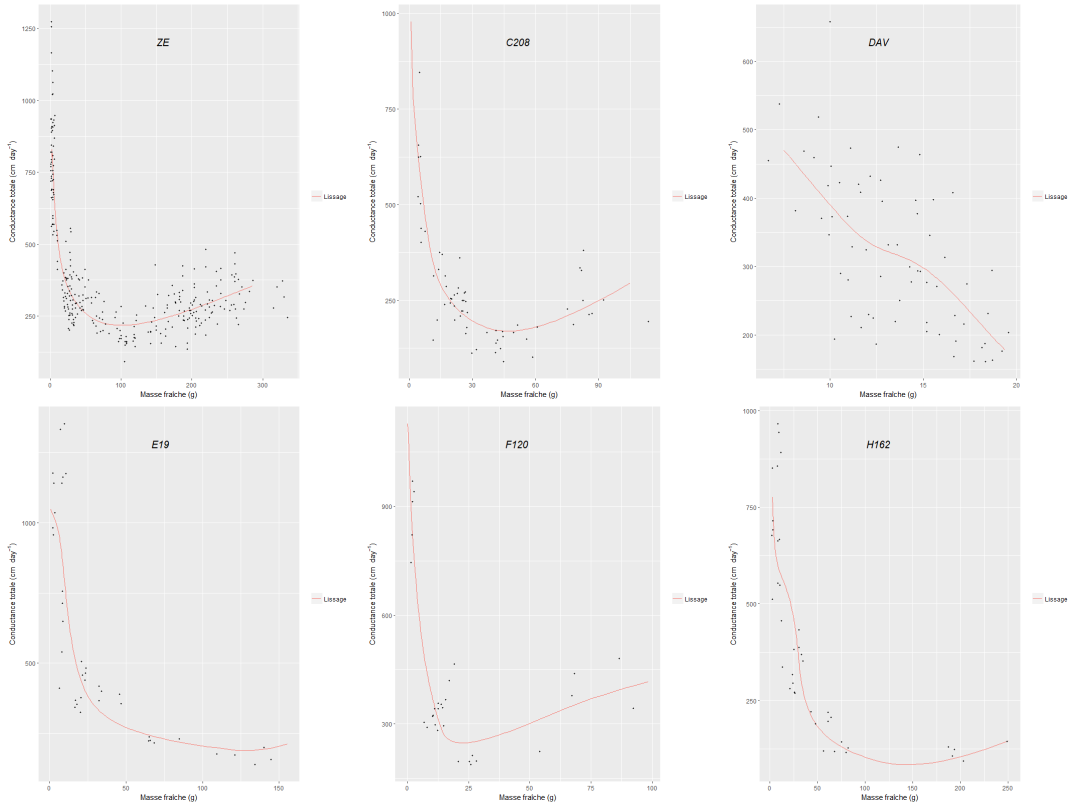

FIGURE 4.12 – Valeurs observées (points) et simulées (ligne) de la conductance pour six génotypes

moins influent que le paramètre $gc2$, ce qui implique que $gc1$ sera considéré comme une contrainte pour la prochaine étape.

4.3.3 Etape 3

Dans cette étape, trois contraintes sont définies : δ_{h1}, δ_{h2} et $gc1$. Le problème d'optimisation est similaire à celui de l'étape précédente, sauf qu'il faut rajouter une contrainte.

La variabilité des paramètres est à nouveau étudiée via le test de l'ANOVA (Table 4.6). D'après les p-valeurs, aucun paramètre ne semble être indépendant des génotypes.

Pour vérifier les corrélations entre les paramètres, nous calculons la matrice de corrélation (Table 4.7). Il n'y a pas de corrélation forte inter génotype entre deux paramètres.

Cependant, nous décidons de mettre le paramètre $cut1$ en contrainte car de fortes corrélations entre ce dernier et $cut2$ ont été détectées au sein de plusieurs génotypes. Comme $cut1$ est moins influent sur la sortie que $cut2$, c'est $cut1$ qui a été mis en contrainte.

4.3.4 Etape 4

Nous avons quatre contraintes à cette étape : δ_{h1}, δ_{h2}, $gc1$ et $cut1$.

Les mêmes démarches que les étapes précédentes ont été effectuées. D'après le test de l'ANOVA, il n'y a aucun paramètre indépendant des génotypes et d'après la matrice de corrélation il n'y aucune corrélation forte entre deux paramètres. De plus, aucune

Paramètre	P-valeur
g'_{sto}	<2e-16
gc_1	<2e-16
gc_2	<2e-16
gc_3	<2e-16
gc_4	<2e-16
gc_5	<2e-16
g'_{ck}	<2e-16
cut_1	<2e-16
cut_2	<2e-16

TABLE 4.4 – Les p-valeurs des paramètres du test de l'ANOVA

Paramètre	g'_{sto}	gc_1	gc_2	gc_3	gc_4	gc_5	g'_{ck}	cut_1	cut_2
g'_{sto}	1								
gc_1	0.29	1							
gc_2	0.27	0.60	1						
gc_3	0.04	0.14	0.10	1					
gc_4	0.22	-0.08	-0.10	0.08	1				
gc_5	-0.07	-0.16	-0.01	-0.25	0.09	1			
g'_{ck}	-0.03	-0.02	-0.10	-0.03	-0.16	0.17	1		
cut_1	-0.14	0.08	-0.06	0.17	-0.18	0.03	0.11	1	
cut_2	-0.07	-0.19	-0.04	-0.02	-0.10	0.02	-0.14	0.05	1

TABLE 4.5 – Matrice de corrélation

corrélation intra-génotype entre deux paramètres n'a été détéctée. Il n'y a donc plus de paramètre à mettre en contrainte.

Les graphiques des valeurs simulées de la conductance en fonction de la masse fraîche pour six génotypes sont présentés sur la figure (4.14).

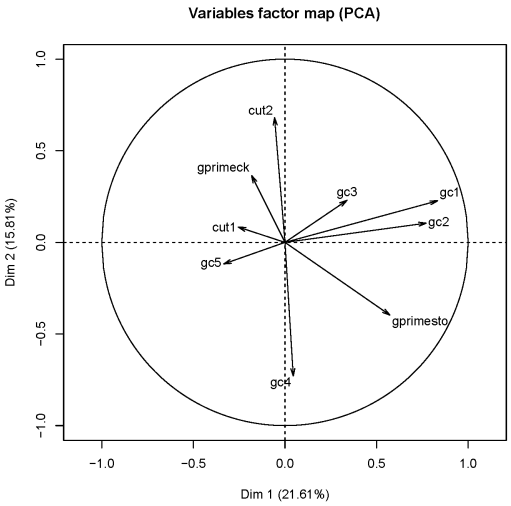

FIGURE 4.13 – ACP normée

Paramètre	P-valeur
g'_{sto}	<2e-16
gc_2	<2e-16
gc_3	<2e-16
gc_4	<2e-16
gc_5	<2e-16
g'_{ck}	<2e-16
cut_1	<2e-16
cut_2	<2e-16

TABLE 4.6 – Les p-valeurs des paramètres du test de l'ANOVA

4.3.5 Conclusion

En appliquant notre méthode d'estimation avec quatre différentes étapes, nous avons défini les paramètres considérés comme génotype-dépendants et ceux qui ne le sont pas. Sur 11 paramètres génotype-dépendants initialement, 7 sont conservés à l'issue des 4 étapes. Le tableau (4.8) récapitule les hypothèses de la nature des paramètres au cours du processus d'estimation.

Pour comparer les qualités de simulation entre différentes étapes, nous avons tracé la variabilité de la racine carrée relative de l'erreur quadratique moyenne (RREQM) de la meilleure solution de chaque génotype pour les 4 étapes (Figure 4.15). Les résultats sont moins bons au fur et à mesure que l'on avance dans notre processus d'estimation. Cela peut être dû au fait que l'algorithme n'arrive pas à converger vers une valeur constante pour les paramètres définis en contraintes. Néanmoins, la qualité de simulation reste très bonne et satisfaisante à l'issu de la dernière étape.

Paramètre	g'_{sto}	gc_2	gc_3	gc_4	gc_5	g'_{ck}	cut_1	cut_2
g'_{sto}	1							
gc_2	0.03	1						
gc_3	0.08	0.31	1					
gc_4	0.25	-0.23	-0.07	1				
gc_5	-0.05	0.09	-0.23	-0.08	1			
g'_{ck}	-0.07	0.18	0.25	-0.08	0.17	1		
cut_1	0.02	-0.04	0.21	-0.19	0.09	0.06	1	
cut_2	-0.08	0.01	-0.19	-0.09	-0.02	-0.04	0.31	1

TABLE 4.7 – Matrice de corrélation

Paramètre	Etape 1	Etape 2	Etape 3	Etape 4
g'_{sto}	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
gc_1	génotype-dépendant	génotype-dépendant	génotype-indépendant	génotype-indépendant
gc_2	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
gc_3	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
gc_4	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
gc_5	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
g'_{ck}	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant
δ_{h1}	génotype-dépendant	génotype-indépendant	génotype-indépendant	génotype-indépendant
δ_{h2}	génotype-dépendant	génotype-indépendant	génotype-indépendant	génotype-indépendant
cut_1	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-indépendant
cut_2	génotype-dépendant	génotype-dépendant	génotype-dépendant	génotype-dépendant

TABLE 4.8 – Résumé des étapes

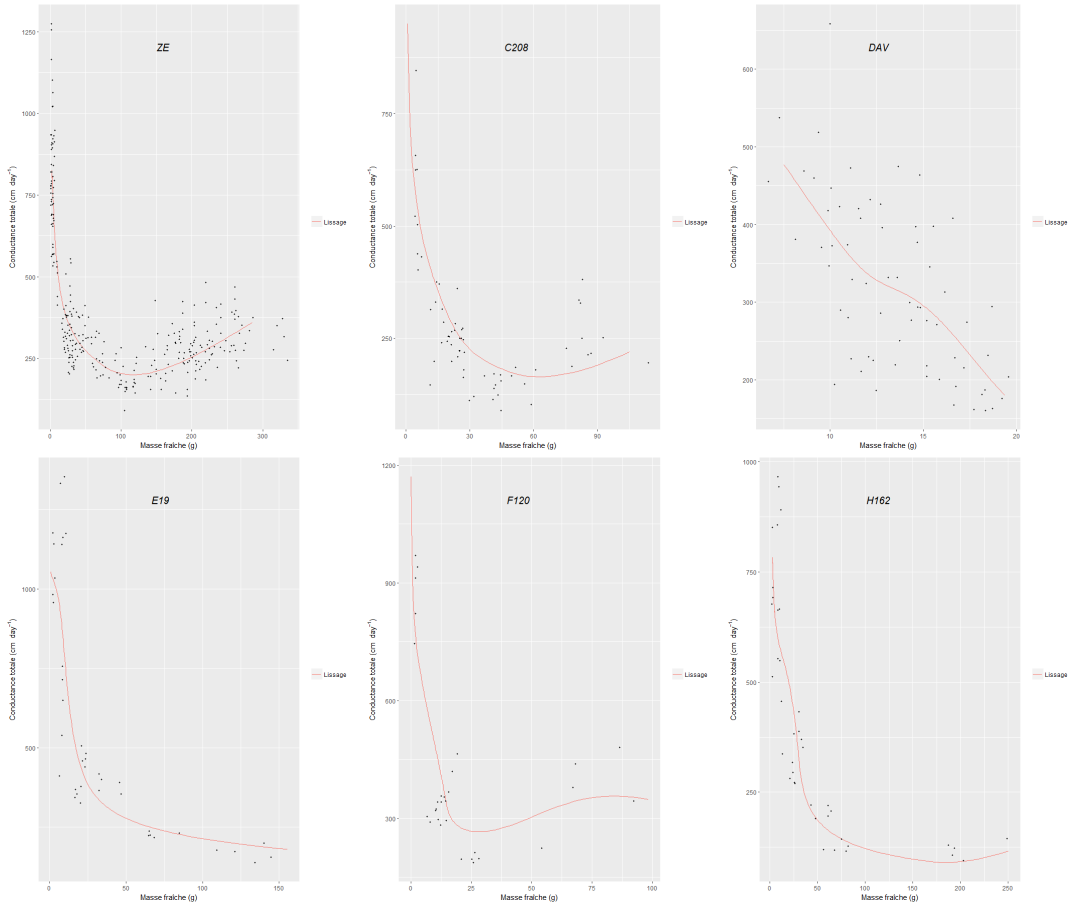

FIGURE 4.14 – Valeurs observées (points) et simulées (ligne) de la conductance pour six génotypes

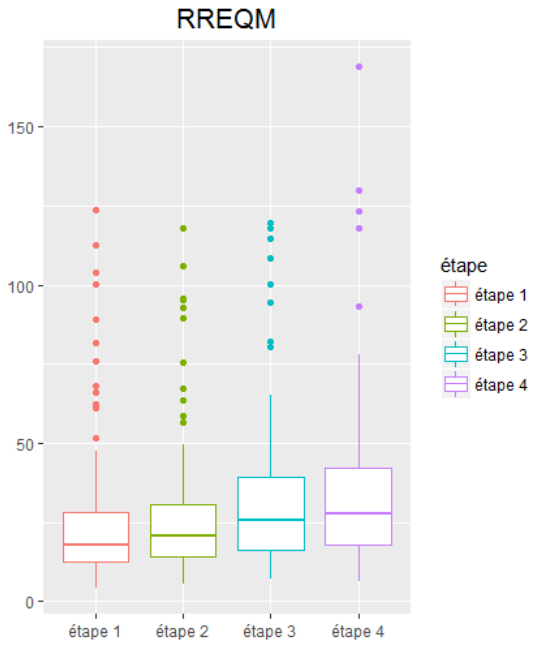

FIGURE 4.15 – Comparaison de la racine carrée relative de l'erreur quadratique moyenne de la meilleure solution de chaque génotype entre les différentes étapes

Chapitre 5

Conclusion

Ce rapport présente une approche d'estimation des paramètres du sous-modèle conductance du modèle fruit virtuel en y intégrant la variabilité génétique de ces paramètres. Il s'agissait de déterminer les paramètres génotype dépendants et ceux qui sont génotype indépendants sur la base de données collectées pour 156 génotypes.

Avant de procéder à l'estimation, deux étapes étaient nécessaires :

— Un travail de lissage de nuages de points a été effectué dans le but d'ajuster des courbes liant le temps (DAFB) et la conductance du fruit. En effet, les données expérimentales correspondaient à des nuages de points liant d'une part la masse fraiche du fruit à la conductance et d'autre part le temps à la masse fraiche. Pour une même valeur de DAFB, nous avions plusieurs valeurs observées de la masse fraîche d'une part et d'autre part plusieurs valeurs observées de la conductance pouvaient correspondre à la même masse fraîche. Par conséquent, un ajustement a été effectué afin d'obtenir une courbe continue entre la masse fraîche et le DAFB. Le même travail a été réalisé entre la conductance totale et la masse fraîche. Pour obtenir les courbes d'ajustement, six méthodes de régression non paramétrique ont été utilisées et comparées pour en retenir la plus adaptée à l'ensemble des génotypes ayant suffisamment de données (103 génotypes). La méthode Basis spline a ainsi été retenue.

— Dans un deuxième temps, l'analyse de sensibilité a été effectuée dans le but d'étudier la sensibilité du modèle Fruit Virtuel aux variations de ses paramètres à des stades différents de la croissance du fruit. Trois méthodes d'analyse de sensibilité globale ont été appliquées et ont abouti quasiment à des résultats similaires. Ils ont révélé l'impact significatif de deux paramètres, gc5 et cut2 tandis que les autres paramètres ont un impact inférieur ou nul sur la conductance totale. Le choix de ces méthodes était basé sur leur adaptation aux modèles numériques non-linéaires.

Quatre des 15 paramètres avaient des valeurs observées et donc ne nécessitaient pas d'être estimées. Pour l'estimation des 11 paramètres restants, nous avons eu recours à l'optimisation en formulant un problème de minimisation des erreurs de prédiction du modèle FV. Pour cela, un processus itératif a été mis en place : les 11 paramètres sont

considérés génotype-dépendants, puis au fur et à mesure nous considérons certains paramètres génotype-indépendants via les analyses de variabilité, de sensibilité et de corrélation. Le processus fait ainsi évoluer le problème d'optimisation d'une configuration sans contrainte au début et intègre au fur et à mesure des contraintes liées aux paramètres génotype indépendants. Pour pouvoir comparer la qualité d'estimation obtenue à chaque étape, des mesures de performance ont été calculées. Ces dernières ont permis d'identifier les estimations des paramètres les plus précises.

Il en résulte que 7 paramètres se sont révélés génotype-dépendants et 4 génotype-indépendants.

Dans un avenir proche et sur la base des expérimentations en cours, il serait intéressant de généraliser l'approche proposée aux 53 génotypes écartés dans ce travail.
Les valeurs estimées des paramètres pourraient faire l'objet d'une recherche de QTL (Quantitative Trait Loci) à l'aide des cartes génétiques disponibles pour la population étudiée. Une analyse de sensibilité pourrait être réalisée par la suite pour déterminer les paramètres du modèle les plus influents sur d'autres sorties d'intérêt (masse fraiche du fruit, sweetness, apparition de fissures) dans le but de concevoir des idéotypes variétaux innovants.
Une possibilité pour améliorer l'estimation des paramètres serait de reconfigurer l'algorithme à évolution différentielle voire d'utiliser un autre algorithme.

Bibliographie

[1] Besse, P., Thomas-Agnan, C. (1989), Robust locally weighted regression and smoothing scatterplots, *Statistique et analyse des données*, **14**, 55-84.

[2] Bollaerts, K., Eilers, P. (2006), Simple and multiple P-splines regression with shape constraints, *British Journal of Mathematical and Statistical Psychology*, **59**, 451-469.

[3] Campolongo, F., Cariboni, J. (2007), An effective screening design for sensitivity analysis of large models, *Environmental Modelling & Software*, **22**, 1509-1518.

[4] Chan, K., Saltelli, A. (1997), Sensitivity Analysis Of Model Output : Variance-based Methods Make The Difference, *Winter Simulation Conference Proceedings*, 261-268.

[5] Cleveland, W. (1979), Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829-836.

[6] Cleveland, W., Devlin, S. (1988), Locally weighted regression : an approach to regression analysis by local Fitting, *Journal of the American Statistical Association*, **83**, 596-610.

[7] Conceicao, E., Maechler, M. (2016), DEoptimR, package R.

[8] Eilers, P., Durban, M. (2016), Twenty years of P-splines, *Statistics and Operations Research Transactions*, **39**, 149-186.

[9] Eilers, P., Marx, B. (1996), Flexible Smoothing with B-splines and penalties, *Statistical Science*, **11**, 89-121.

[10] El Dor, A.. (2012), Perfectionnement des algorithmes d'optimisation par essaim particulaire : applications en segmentation d'images et en électronique, thèse de doctorat à l'Université Paris-Est.

[11] Eubank, R. (1999), *Nonparametric regression and spline smoothing*, Marcel Dekker, Inc.

[12] Friedman, J. (1984). A variable span smoother, *Laboratory for Computational Statistics*, Department of Statistics, Stanford University.

[13] Génard, M., Bertin, N. (2007), Towards a virtual fruit focusing on quality : modelling features and potential uses, *Journal of Experimental Botany*, **58**, 917-928.

[14] Génard, M., Robin, C. (2010), Elaboration de la qualité du fruit : composition en métabolites primaires et secondaires, *Innovations Agronomiques*, **9**, 47-57.

[15] Gibert, C., Génard, M. (2009), Modelling the effect of cuticular crack surface area and inoculum density on the probability of nectarine fruit infection by Monilinia laxa, *Plant Pathology*, **58**, 1021-1031.

[16] Gibert, C., Lescourret, F. (2005), Modelling the effect of fruit growth on surface conductance to water vapour diffusion, *Annals of Botany*, **95**, 673-683.

[17] Hastie, T. J., Chambers, J. M. (1991), *Statistical Model in S*, Chapman and Hall, CRC.

[18] Hastie, T. J., Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer.

[19] Hastie, T. J., Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, CRC.

[20] Homma, T., Saltelli, A. (1996), Importance measures in global sensitivity analysis of nonlinear models , *Reliability Engineering and System SaJety* , **52**, 1-17.

[21] Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante*, thèse de doctorat de l'Université Joseph Fourier.

[22] Memmah, M., Quilot-Turion, B. (2014), The relationship between metaheuristics stopping criteria and performances : cases of NSGA-II and MOPSO-CD for sustainable peach fruit design, *International Journal of Applied Metaheuistic Computing*, **5**, 47-74.

[23] Moriasi, D., Arnold, J. (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the ASABE*, **50**, 885-900.

[24] Morris, M. (1991), Factorial sampling plans for preliminary computational experiments, *Technometrics*, **33**, 161-174.

[25] Nadaraya, E. (1964), On Estimating Regression, *Theory of Probability and its Applications*, **9**, 141-142.

[26] O'Sullivan, F. (1986), A statistical perspective on III-posed inverse problems, *Statistical Science*, **1**, 502-518.

[27] Parzen, E. (1962), On estimation of a probability density function and moden, *Annals of Mathematical Statistics*, **33**, 1065-1076.

[28] Pujol, G. (2017), sensitivity, package R.

[29] Quilot-Turion, B., Ould-Sidi, M. (2012), Optimization of parameters of the 'Virtual Fruit' model to design peach genotype for sustainable production systems, *European Journal of Agronomy*, **42**, 34-48.

[30] Reinsch, C. (1967), Smoothing by spline functions, *Numerische Mathematik*, **10**, 177-183.

[31] Rosenblatt, M. (1956), Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, **27**, 832-837.

[32] Saltelli, A., Bolado, R. (1998), An alternative way to compute Fourier amplitude sensitivity test (FAST), *Computational Statistics & Data Analysis*, **26**, 445-460.

[33] Saltelli, A., Ratto, M. (2008), *Global Sensitivity Analysis : The Primer*, Wiley.

[34] Sobol, I. (1993), Sensitivity analysis for nonlinear mathematical models, *Mathematical Modeling & Computational Experiment*, **1**, 407-414.

[35] Storn, R., Price, K. (1997), Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, *Journal of Global Optimization*, **11**, 341-359.

[36] Watson, G. (1964), Smooth regression analysis, *Sankhya : The Indian Journal of Statistics, Series A*, **26**, 359-372.

Annexe A

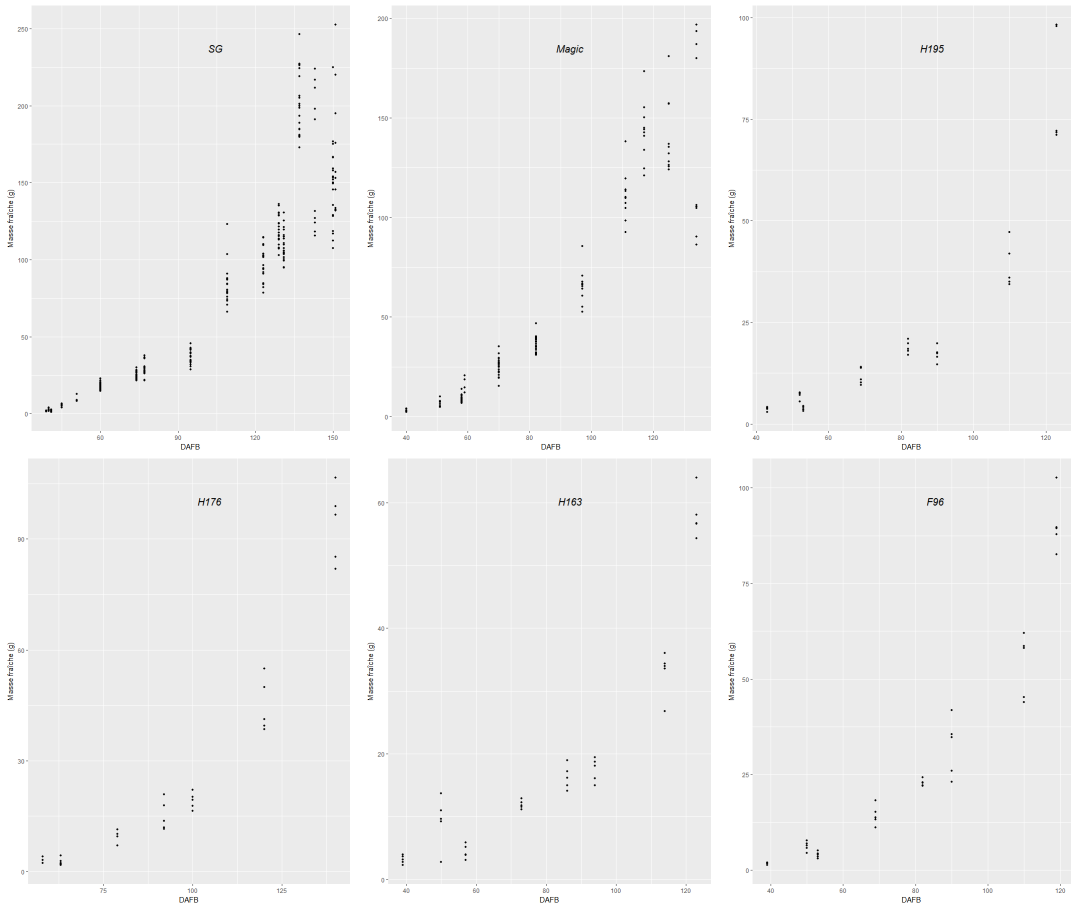

FIGURE A.1 – Données observées de la masse fraîche en fonction du DAFB pour 6 génotypes

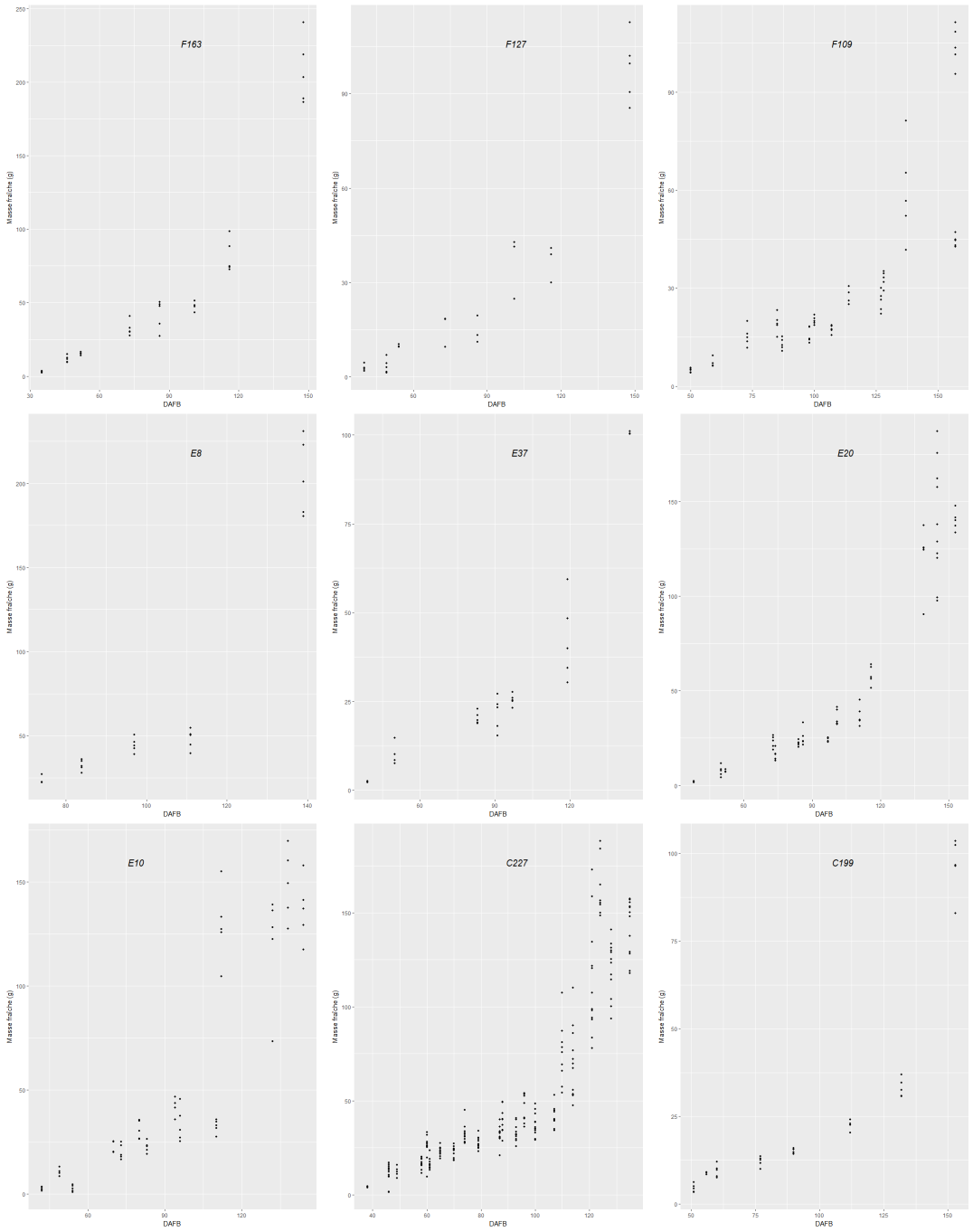

FIGURE A.2 – Données observées de la masse fraîche en fonction du DAFB pour 9 génotypes

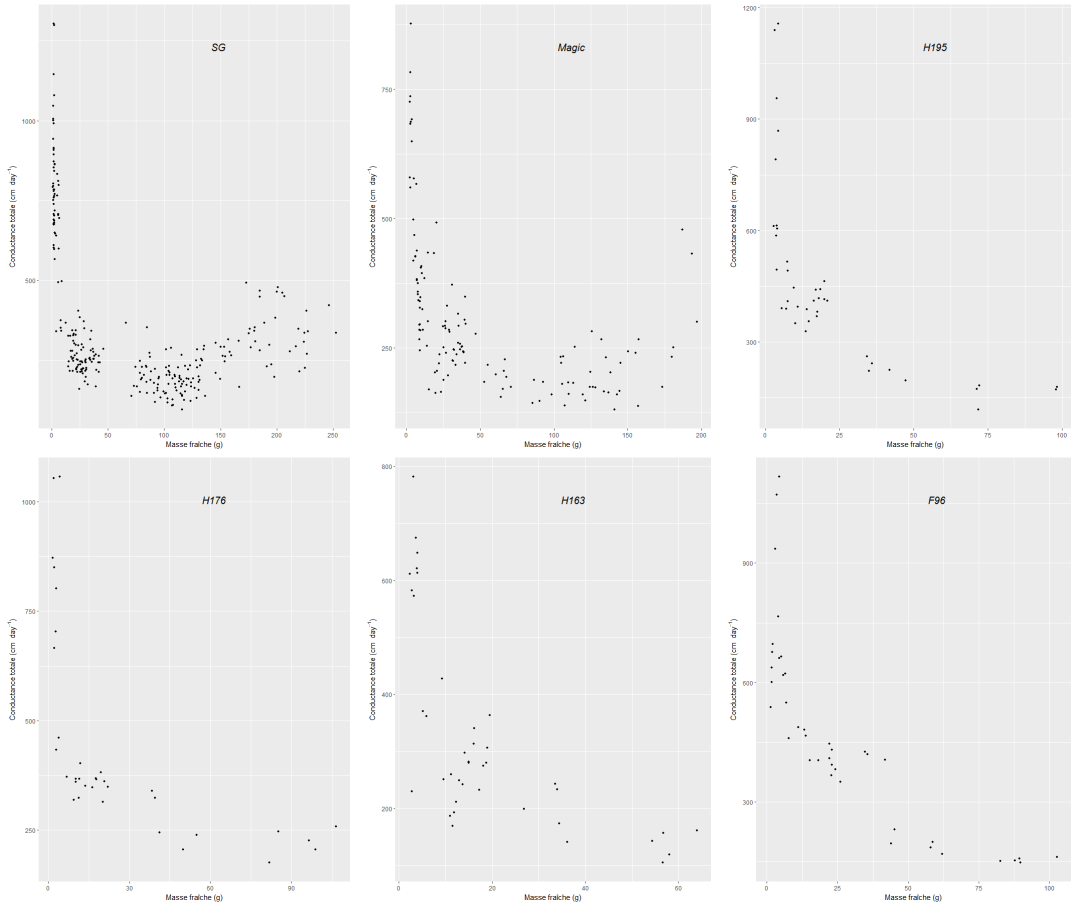

FIGURE A.3 – Données observées de la conductance totale en fonction de la masse fraîche pour 6 génotypes

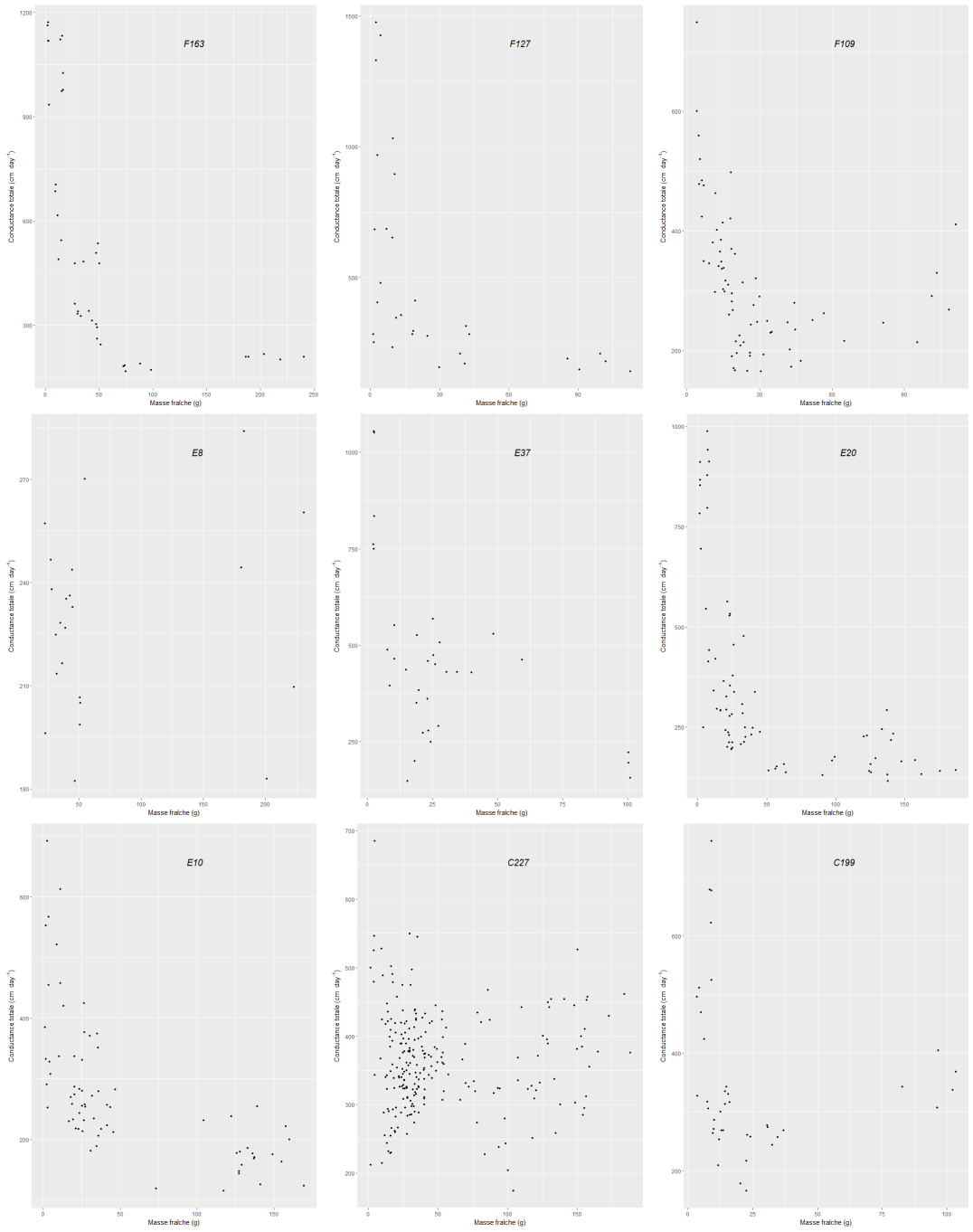

FIGURE A.4 – Données observées de la conductance totale en fonction de la masse fraîche pour 9 génotypes

Annexe B

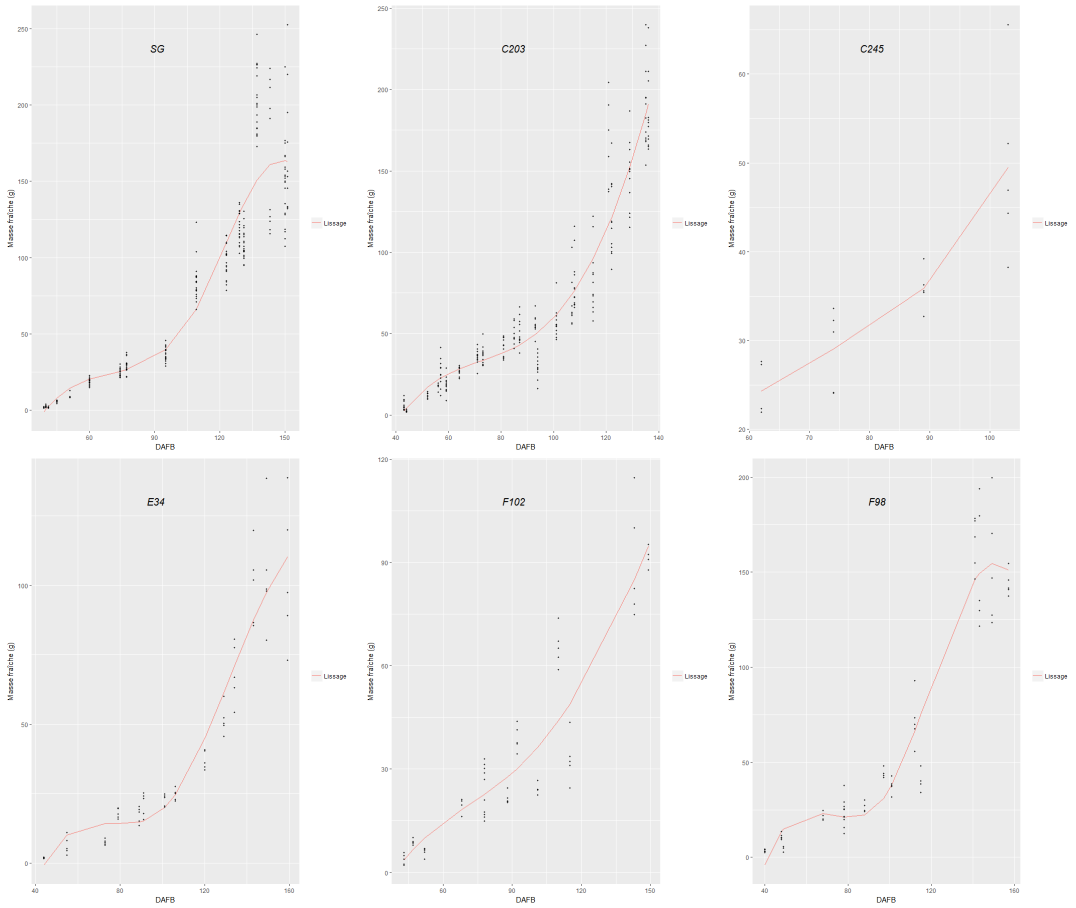

FIGURE B.1 – Lissage obtenue avec la méthode basis spline pour la masse fraîche en fonction du DAFB pour 6 génotypes

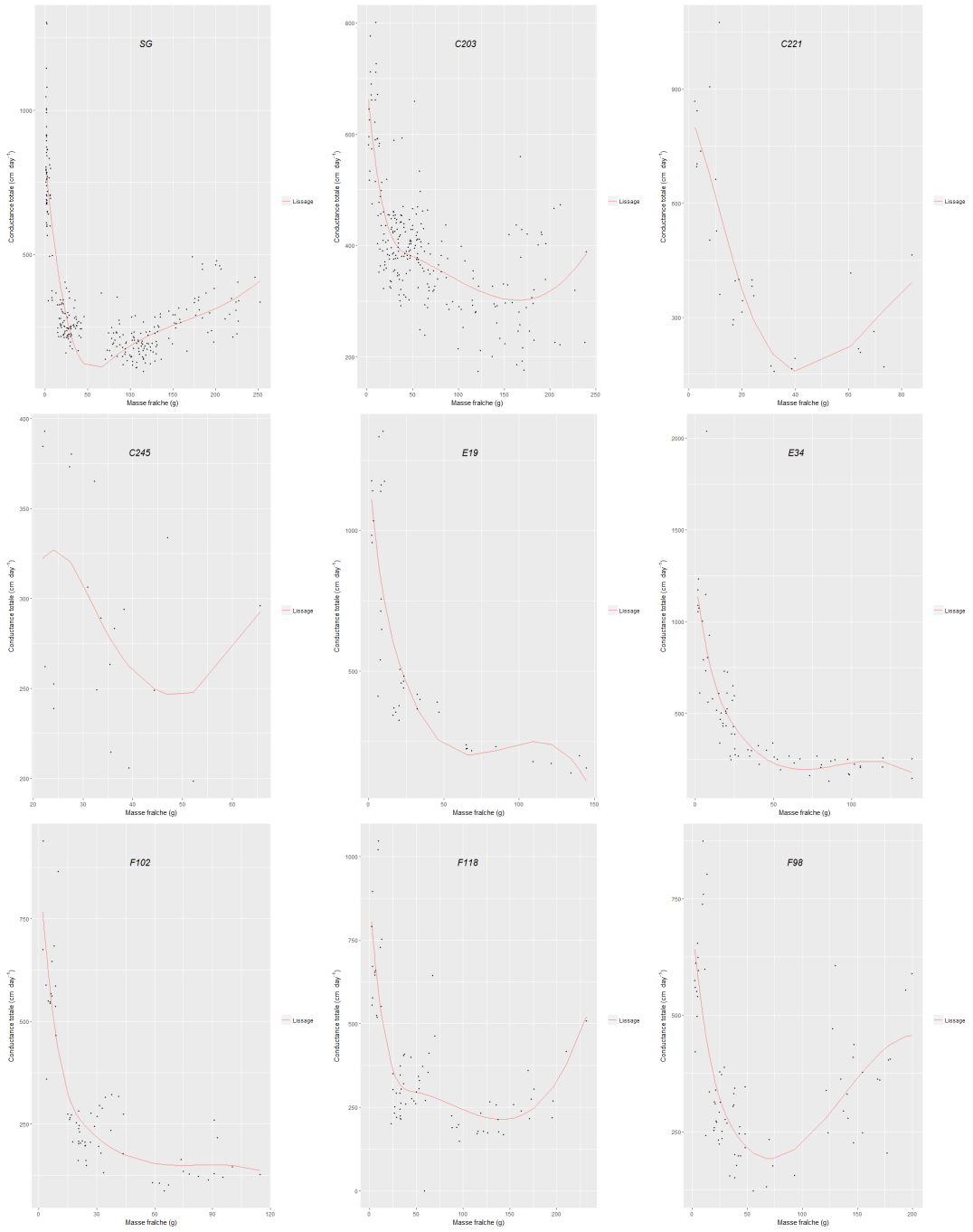

FIGURE B.2 – Lissage obtenue avec la méthode basis spline pour la conductance totale en fonction de la masse fraîche pour 9 génotypes

Annexe C

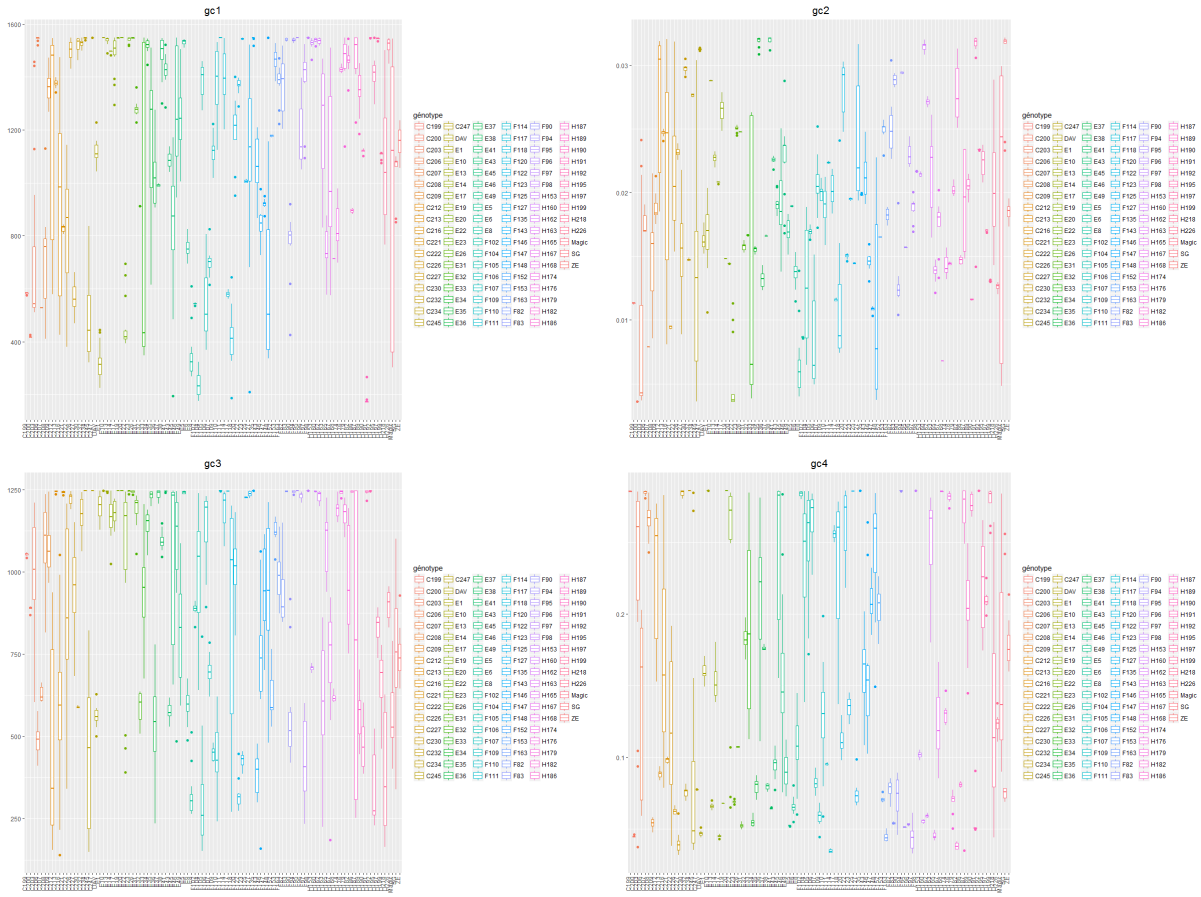

FIGURE C.1 – Variabilité intra et inter génotype de 4 paramètres

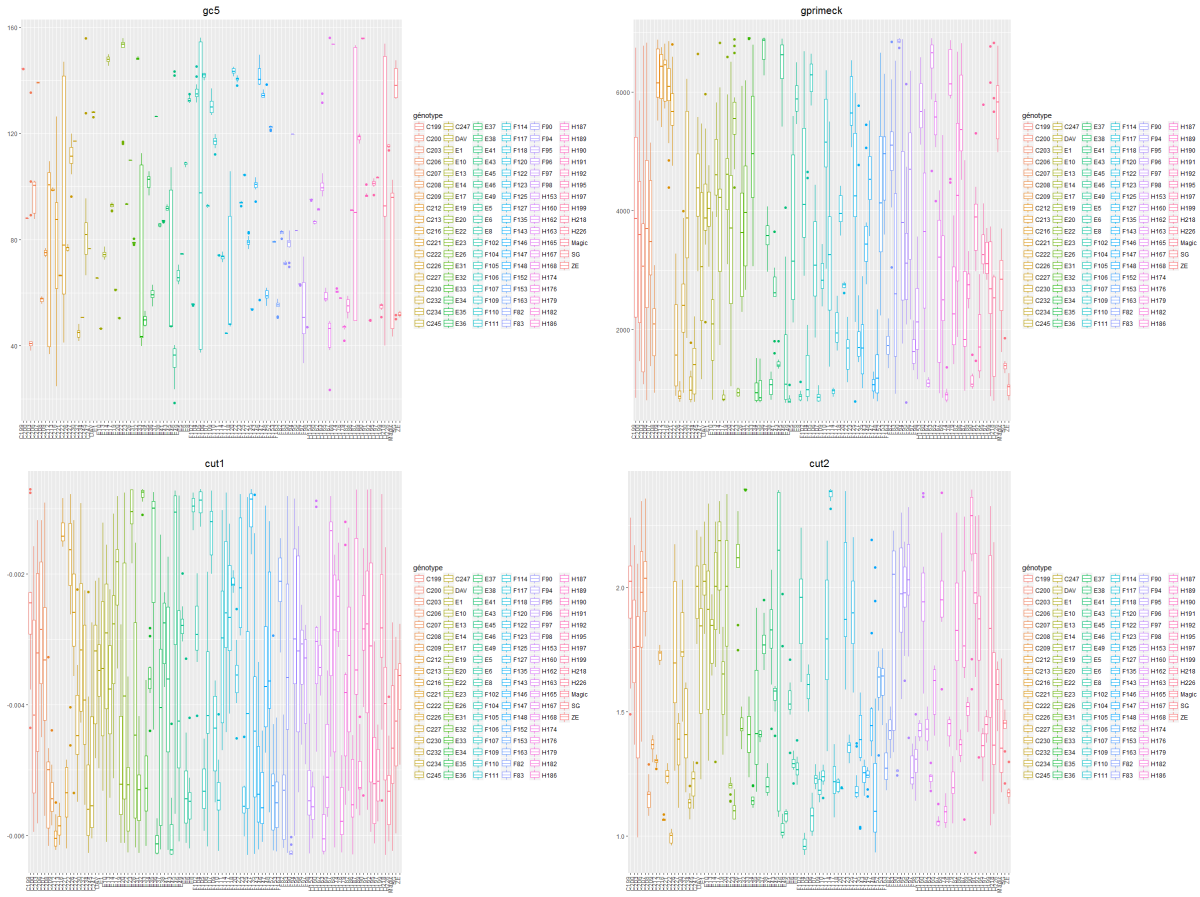

FIGURE C.2 – Variabilité intra et inter génotype de 4 paramètres

Annexe D

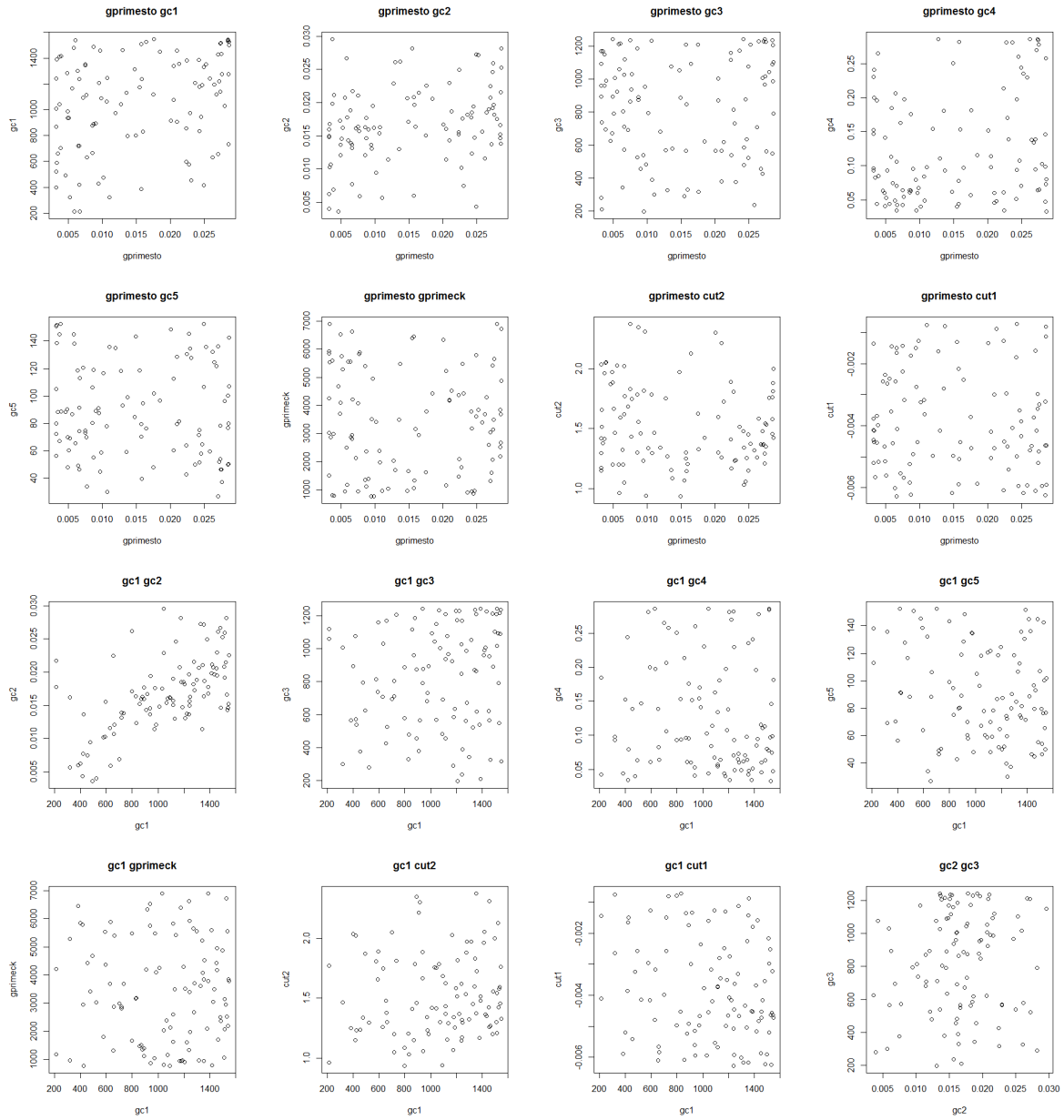

FIGURE D.1 – Corrélations entre les paramètres à l'étape 2 du processus d'estimation des paramètres

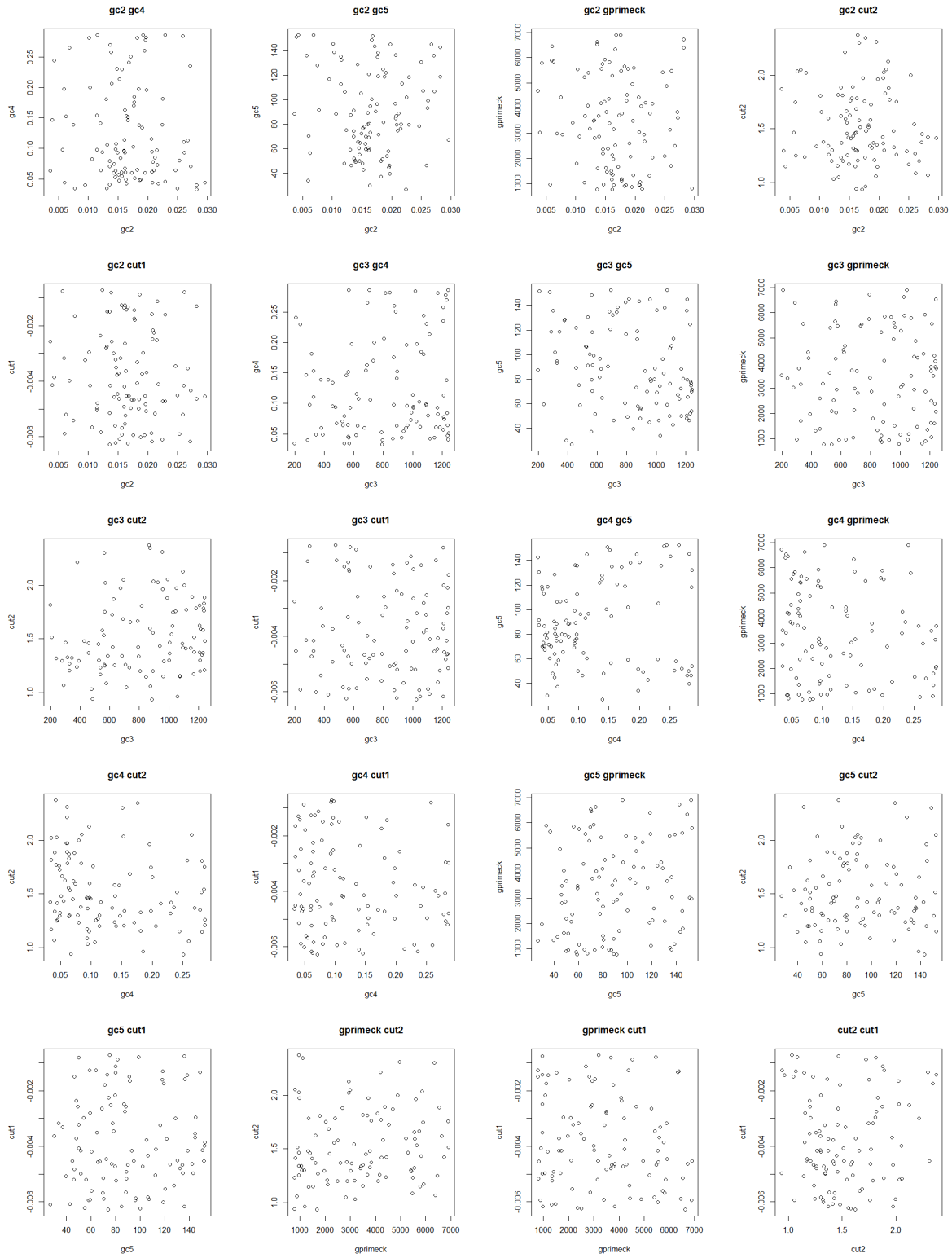

FIGURE D.2 – Suite de la figure précédente

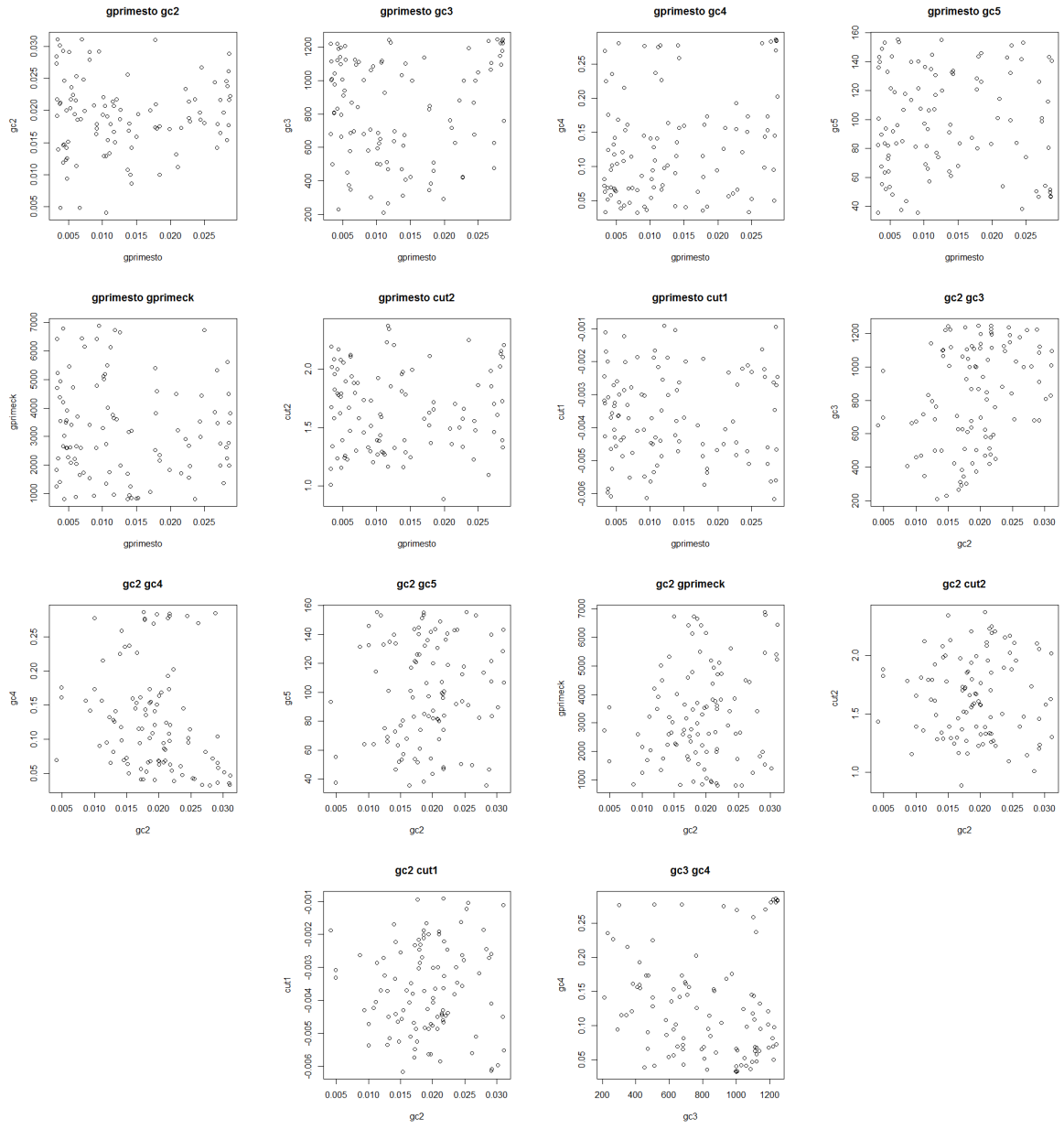

FIGURE D.3 – Corrélations entre les paramètres à l'étape 3 du processus d'estimation des paramètres

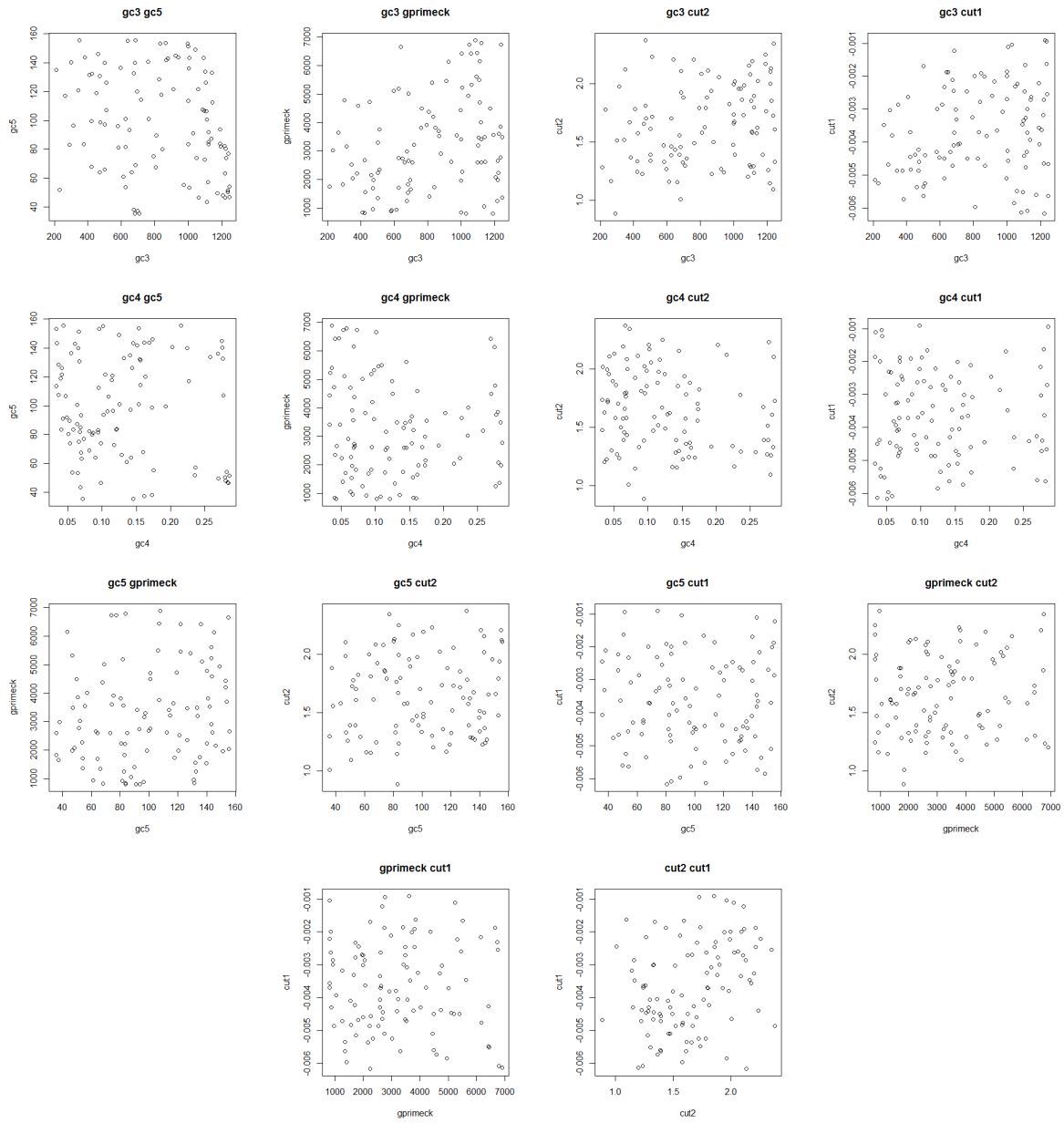

FIGURE D.4 – Suite de la figure précédente

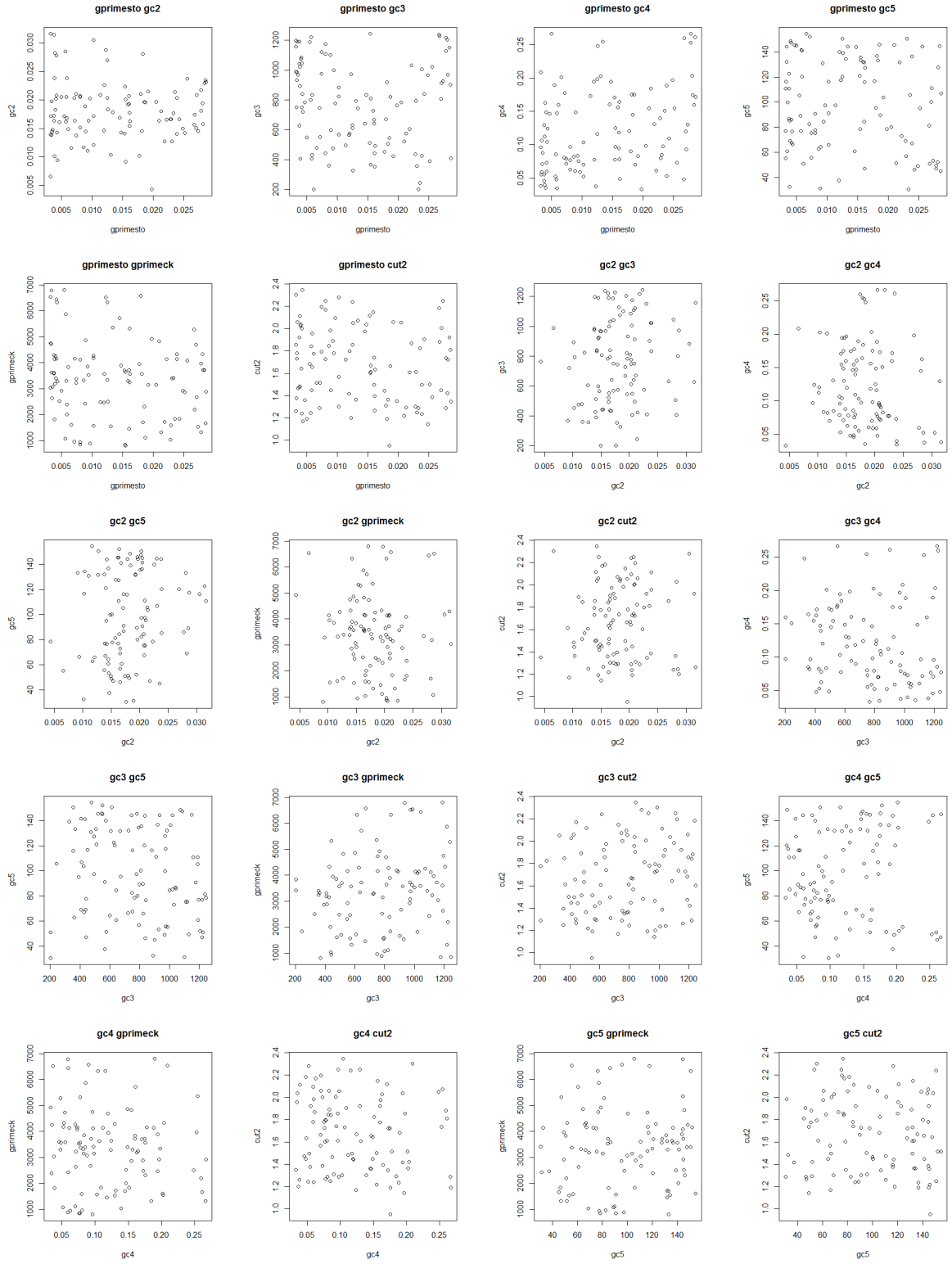

FIGURE D.5 – Corrélations entre les paramètres à l'étape 4 du processus d'estimation des paramètres