Chapitre G

Ajustement et validation d'un modèle

Soit une fonction de croissance déterministe $y(t, \mathbf{P})$, $\mathbf{P}$ étant l'ensemble des paramètres supposés constants sur le temps. Son ajustement à une série d'observations peut concerner en pratique deux situations expérimentales très différentes :

- mesures non destructrices : les observations peuvent être réalisées sur les mêmes «individus» durant toute la durée du processus (dimensions macroscopiques d'une plante ou d'un organe, densité optique d'une même culture cellulaire *in vitro*, par exemple), offrant la possibilité d'ajustements soit individu par individu, soit sur les moyennes temps par temps ;
- mesures destructrices : l'acquisition des données nécessite une suite d'échantillons différents (par exemple poids, teneurs, diverses mesures physiques fines).

Examinons dans ses grandes lignes les problèmes importants qui se posent dans ce travail d'adéquation, ceci étant un préalable à la comparaison ultérieure de différents modèles mis en concurrence pour tel phénomène précis de croissance. D'une part quelques notions de base seront rappelées, concernant les propriétés statistiques des échantillons de mesures et l'essentiel des tests statistiques nécessaires pour une bonne conduite de l'analyse. Il s'agit en définitive pour le modélisateur de s'assurer correctement de la qualité et de la validité de l'ajustement obtenu avant d'en déduire des considérations d'ordre biologique sur ce que peut lui apporter le modèle. La question générale de l'estimation des paramètres sera présentée sous deux aspects pratiques. Examinée d'abord dans le cadre statistique d'une régression, elle sera posée ensuite d'une manière plus générale, hors toute hypothèse probabiliste *a priori*, avec le recours à un algorithme itératif.

Ce sont là des problèmes usuels de portée générale qui sont abondamment traités dans divers ouvrages de statistiques ou de mathématiques appliquées auxquels le lecteur peut se référer, dont nous donnons quelques références. Il nous semble néanmoins utile d'en donner ici un aperçu servant de memento pouvant guider et éclairer la démarche de modélisation au regard de l'utilisation parfois un peu aveugle de procédures informatiques dont peut disposer aisément le praticien biologiste alors que s'impose toujours la nécessité de connaître et respecter les conditions d'emploi de toute méthodologie analytique.

G.1. Analyse statistique

G.1.1. Propriétés des échantillons

Rappelons les hypothèses usuelles suivantes (hors statistique non paramétrique) :

- *indépendance* des différentes mesures (nature aléatoire simple des échantillons) ;
- *loi de probabilité* des données : à chaque instant t_j de mesure, les observations y_{ji} suivent une loi normale de Laplace-Gauss ;
- *homogénéité des variances* d'échantillon (homoscédasticité).

Disposant d'échantillons de taille n (supposée ici identique pour tous les échantillons) nous nous référons pour chaque temps t_j d'une série chronologique de longueur p, au modèle stochastique classique de base :

$$Y_{ji} = \mu_j + \varepsilon_{ji} \quad ; \quad j = 1, ..., p \quad ; \quad i = 1, ..., n$$

μ_j est la valeur théorique (espérance mathématique $E(Y_j)$ de la population parente à l'instant t_j) et ε_{ji} les variables aléatoires associées. Les termes ε_{ji} sont des variables gaussiennes centrées de même variance σ^2 indépendante du temps : $\mathcal{N}(0, \sigma^2)$.

Ces hypothèses sont à soumettre à divers tests statistiques classiques (par exemple test d'adéquation du χ^2, tests de normalité de Kolmogorov-Smirnov et de Shapiro-Wilks, test d'homoscédasticité).

Soulignons que ce sont des hypothèses fortes pouvant ne pas correspondre aux caractéristiques des observations empiriques. D'une part la loi de probabilité d'une variable en croissance varie souvent au cours du processus, sa distribution pouvant passer d'une dissymétrie gauche à une dissymétrie droite. D'autre part la variabilité des observations fluctue elle-même selon l'instant de croissance, *i.e.* selon l'état physiologique de l'objet mesuré. Nous avons donné des exemples sur ces questions souvent négligées (voir chap. 21 du livre *Biomathématiques de la croissance* de R. Buis, fig. 21.1).

G.1.2. Régression polynomiale

Hors toute référence à un modèle de croissance donné, *i.e.* sans hypothèses *a priori* sur l'équation de vitesse, une première étude de cinétique peut faire appel au modèle statistique univarié de *régression polynomiale* :

$$Y = P_k(X) + \varepsilon \qquad [1]$$

$P_k(X)$ étant un polynôme de degré k de la variable explicative X. Celle-ci est ici le temps t de chaque série de mesures (temps supposé connu « sans erreur »), soit :

$$Y = \sum_{k=0}^{q} \beta_k t^k + \varepsilon \quad ; \quad t = 1, \ldots, p \quad ; \quad q \leq p - 1 \qquad [2]$$

Par exemple avec deux instants de mesure ($p = 2$) nous retrouvons le schéma de la régression linéaire; pour $p = 3$ nous pouvons obtenir un ajustement parabolique…

▶ L'estimation des coefficients de régression β_j de [2] est basée sur le *principe des moindres carrés* qui fournit des estimateurs sans biais [1]. Cette méthode se propose de minimiser, pour tout temps t_j, la somme des carrés des résidus e_j ou écarts entre valeurs observées y_j et valeurs estimées

$$\hat{y}_j = \sum_{k=0}^{q} \hat{\beta}_k t^k$$

Les $\hat{\beta}_j$ sont solutions du système d'équations répondant à la condition :

$$\min S_j = \sum_j e_j^2 = \sum_j (y_j - \hat{y}_j)^2$$

c'est-à-dire :

$$\left\{ \frac{\partial S_j}{\partial \hat{\beta}_j} = 0 \right\}$$

▶ Le *test statistique* de signification globale d'une régression consiste en une *analyse de variance* classique comparant la variance entre échantillons (entre « temps de mesures ») (variance inter-groupes) à la variance de nature aléatoire intra-groupes (dite « erreur expérimentale »). La statistique de test $F = \sigma^2_{\text{inter}} / \sigma^2_{\text{intra}}$ calculée est à référer à la loi de Fisher-Snedecor en hypothèse nulle $\sigma^2_{\text{inter}} = \sigma^2_{\text{intra}}$. L'analyse est conduite en test unilatéral de l'hypothèse alternative simple $\sigma^2_{\text{inter}} > \sigma^2_{\text{intra}}$.

▶ En pratique il convient d'*optimiser* le modèle [2] en recherchant le degré q minimal qui permet une détermination statistiquement significative de la variation inter-groupes (à telle probabilité, ou risque

1 On peut considérer comme « pratiquement équivalente » l'estimation par la méthode du *maximum de vraisemblance* (*maximum likelihood*) dont le développement implique le recours à des bases probabilistes, alors que la méthode des moindres carrés est de portée plus générale.

d'erreur). Divers logiciels de statistiques proposent des algorithmes de *régression pas à pas* (*stepwise regression*). Nous noterons ci-après une méthode simple de régression basée sur l'estimation de polynômes orthogonaux.

Régression sur polynômes orthogonaux

Le principe est de décomposer la variation inter-groupes à $p - 1$ degrés de liberté en $p - 1$ composantes orthogonales. L'indépendance de celles-ci permet de les soumettre chacune à un test en analyse de variance par rapport à la variation résiduelle. En ne retenant que les composantes significatives, il est ainsi possible d'optimiser la représentation par une régression polynomiale de degré égal au degré le plus élevé des composantes significatives.

Résumons cette méthode en écrivant le modèle sous une forme générale avec x pour variable explicative (en cinétique ce sera le temps de mesure t). Au lieu du modèle classique [1] en x nous posons :

$$Y = \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_2 + \ldots = \sum_{k=0}^{p-1} \xi_k$$

les composantes ξ_k étant des polynômes orthogonaux de degré successif de la forme :

$$\begin{aligned} \xi_1 &= \alpha_1 + x \\ \xi_2 &= \alpha_2 + \beta_2 x + x^2 \\ \xi_3 &= \alpha_3 + \beta_3 x + \gamma_3 x^2 + x^3 \\ &\ldots \end{aligned} \qquad [3]$$

On déduit ensuite les sommes des carrés d'écarts sur les n individus et les p valeurs de la variable explicative. Pour le détail des calculs pas à pas voir par exemple Mather (1965).

Schématisons la méthode de calcul à partir d'un exemple simple à 3 valeurs de la variable explicative x en progression arithmétique (0, 1, 2). Soient k_{1l} les valeurs de la composante ξ_1 pour les différentes valeurs de x

$$k_{11} = \alpha_1 + x_1 \; ; \; k_{12} = \alpha_1 + x_2 \; ; \; k_{13} = \alpha_1 + x_3$$

En simplifiant par centrage $\sum_l k_{1l} = 0$, nous avons $\alpha_1 = -\bar{x} = -1$, d'où la série des coefficients polynômiaux de ξ_1 : $\{-1, 0, 1\}$. De la même manière et en ajoutant une condition d'orthogonalité nous avons pour la composante ξ_2, à un coefficient ⅓ près, la série $\{1, -2, 1\}$. Dans ce cas simple on voit l'interprétation de cette décomposition. Le polynôme ξ_1 correspond à la tendance linéaire définie par le contraste entre les deux valeurs extrêmes, alors que ξ_2 oppose la valeur médiane aux extrêmes (écart à la tendance linéaire).

En pratique, avec des contraintes expérimentales (les x choisis en progression arithmétique), on dispose de séries de coefficients polynômiaux publiées dans certains ouvrages de tables statistiques. Par exemple, Fisher R.A. & Yates F. (1963).

Par le calcul des composantes ξ_k l'analyse statistique revient ainsi à expliciter $(p - 1)$ comparaisons indépendantes appelées *contrastes*, chacun étant basé sur une suite de ces coefficients polynômiaux.

Remarques

1 - Cette méthode est une application particulière de recherche d'une *courbe ou d'une surface significative de réponse* du type effet (concentration d'un ou plusieurs facteurs) (voir ouvrages usuels de statistiques).

2 - Il existe plusieurs types de décomposition en polynômes orthogonaux, sujet bien étudié en mathématiques (tels les polynômes de Legendre, les polynômes trigonométriques de Tchebychev, les séries de Fourier). Ces derniers sont proposés par certains logiciels d'ajustement (*curve fitting*) (par exemple *TableCurve 2D* ® de *Jandel Scientific*).

Exemple - Reprenons les données publiées dans Mather (1965) sur la croissance pondérale de la Tomate (sur 6 instants de mesure, $p = 6$). L'analyse porte sur le log des mesures (pour normaliser la distribution). Les temps de mesure étant équidistants et notés de 1 à 6 (semaines), nous avons les séries de coefficients polynômiaux suivants (composantes en ligne, variable explicative en colonne) :

Composante	Variable explicative					
	1	2	3	4	5	6
ξ_1	−5	−3	−1	1	3	5
ξ_2	5	−1	−4	−4	−1	5
ξ_3	−5	7	4	−4	−7	5
ξ_4	1	−3	2	2	−3	1
ξ_5	−1	5	−10	10	−5	1

La figure G1 montre les ajustements fournis par les composantes linéaire, quadratique et cubique. Les composantes de degré supérieur (4 et 5) n'apportent aucune détermination significative. On peut donc retenir une régression polynômiale de degré 3.

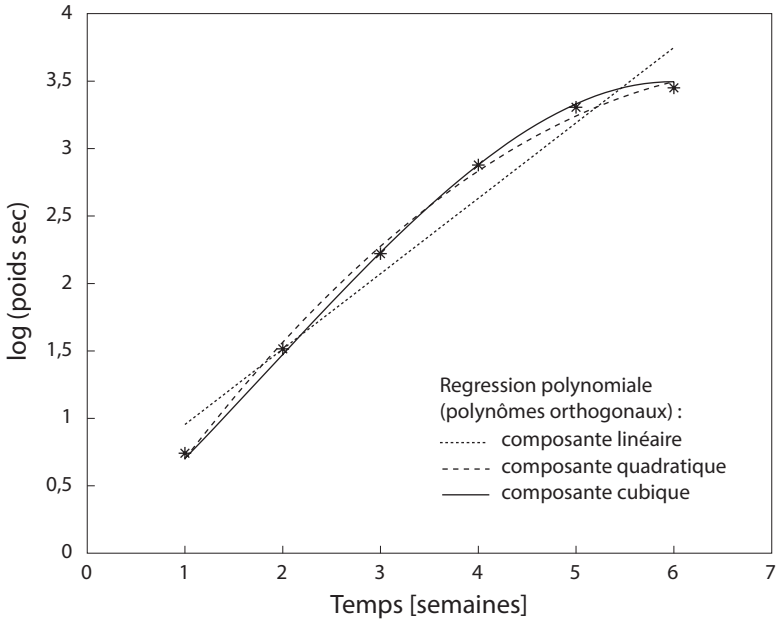

Figure G1 - Croissance pondérale de la Tomate [d'après Mather (1965)]
∗ : données (moyennes)

G.1.3. Qualité de l'ajustement : validation d'un modèle

▶ *Coefficient de détermination ou de corrélation*

Ce coefficient exprime la proportion de la variance expliquée par rapport à la variance totale des observations. Plutôt que d'utiliser les sommes des carrés des écarts (on dit alors « coefficient de détermination observé »), il est préférable de le calculer avec les variances estimées (*i.e.* tenant compte des degrés de liberté en analyse de la variance). On précise « *coefficient de détermination ajusté* », soit :

$$\hat{\rho}^2 = \frac{\hat{\rho}^2_{\text{totale}} - \hat{\rho}^2_{\textit{résiduelle}}}{\hat{\rho}^2_{\text{totale}}}$$

(en réalité cette estimation n'est pas dépourvue de biais). Considérant que la régression polynômiale a la nature d'une régression multiple, on parle également de *coefficient de détermination ou de corrélation multiple*.

Ce coefficient ne fournit qu'une indication globale qui peut se révéler insuffisante. Une valeur élevée peut en effet masquer des inadéquations locales que seule peut révéler une analyse des résidus.

▶ ***Analyse des résidus***

En rapport avec les hypothèses préliminaires il importe de vérifier *a posteriori* les conditions de validité de l'analyse, ce que l'on fait par une *analyse statistique des résidus* $e_j = \hat{\varepsilon}_j$. Résumons les points essentiels.

Une première approche de la vérification des hypothèses initiales peut être engagée d'abord avec le tracé de l'*histogramme des résidus*, pouvant montrer au moins la symétrie de leur distribution. Ensuite le tracé graphique des résidus en fonction du temps peut déceler si leur distribution (amplitude, répartition autour de leur moyenne de groupe, valeurs extrêmes) semble conforme aux hypothèses d'homoscédasticité et d'indépendance.

Nous avons donné des exemples de distributions de résidus (voir par exemple la fig. 6.10 du livre *Biomathématiques de la croissance* de R. Buis pour la croissance logistique de *Lupinus*, obtenue avec le logiciel *Table Curve 2D* qui affiche à la fois les données, le lissage et les résidus). Autre précision, on peut recourir à la représentation conjointe des observations, de l'ajustement et des limites de son *intervalle de confiance* (voir fig. 19.2 du livre *Biomathématiques de la croissance* de R. Buis sur un exemple de croissance cellulaire). Mais il reste généralement nécessaire de réaliser les tests adéquats sur les résidus portant sur les points suivants :

- normalité de la distribution des résidus : voir les tests de normalité précédemment cités,
- homoscédasticité des résidus (test de Hartley),
- indépendance ou non-corrélation des résidus.

Notons en particulier le test de Durbin-Watson qui vérifie l'existence d'une autocorrélation entre résidus successifs en posant le modèle linéaire d'autorégression :

$$\hat{\varepsilon}_j = \rho \hat{\varepsilon}_{j-1} + \nu_i \quad , \text{ avec } \nu_i \sim \mathcal{N}(0, \sigma_\nu)$$

Avec la statistique de test

$$\frac{\sum_j (\hat{\varepsilon}_j - \hat{\varepsilon}_{j-1})^2}{\sum_j \hat{\varepsilon}_j^2}$$

on teste l'hypothèse nulle $\rho = 0$. À noter que, comme pour toute analyse d'autorégression, l'efficacité de ce test dépend de la longueur de la série.

En complément de ces tests il convient de vérifier si les résidus moyens des différents groupes présentent des valeurs relatives d'un même ordre de grandeur. Ceci est particulièrement important en début de croissance où les résidus, faibles en valeur absolue, peuvent être importants en valeur relative. Ceci montrerait une inadéquation locale du modèle, alors qu'un test global d'analyse de variance pourrait conclure à une détermination significative. A titre d'exercice on peut vérifier que divers exemples publiés en détail dans la littérature illustrent cette distorsion entre l'interprétation résultant d'un jugement statistique global et l'existence de défauts locaux qu'il convient de ne pas négliger.

G.2. Estimation des paramètres d'un modèle déterministe

Revenons à nos modèles déterministes autonomes de croissance et résumons les bases mathématiques d'estimation de leurs paramètres selon le principe des *moindres carrés* que nous avons mentionné

précédemment pour le modèle stochastique de régression [1]. Résumons la question en nous référant à la méthode de Marquardt d'utilisation classique pour les modèles non linéaires.

Algorithme de Marquardt

Nous ne détaillerons pas cet algorithme classique (Marquardt, 1963) dont on trouvera un exposé simple, illustré par des données biologiques, dans Conway *et al.* (1970). Donnons-en seulement le principe qui au départ se réfère à la *méthode de Newton* (ou de *Newton-Raphson*) de résolution d'une équation $f(x) = 0$ (sur cette méthode de base voir par exemple Grivet, 2013).

Soit x^0 une première approximation *a priori* de la racine (numérotation de l'itération en exposant). On considère la tangente à la courbe $y = f(x)$ en ce point x^0. Soit x^1 l'intersection de cette tangente avec l'axe des abscisses. On répète le processus avec cette valeur, d'où une nouvelle approximation x^2, et ainsi de suite jusqu'à convergence en se fixant un seuil donné. Nous utilisons donc le processus itératif :

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)} \text{ jusqu'à} \left|\frac{f}{f'}\right| < \varepsilon$$

On voit que ce processus peut ne pas converger si f' est faible ou nulle.

Avec p paramètres (vecteur $\mathbf{a}$) nous avons :

$$\mathbf{a}^{k+1} = \mathbf{a}^k + \Delta_g^k$$

où la correction par gradient $\Delta_g^k \propto \mathbf{D}^k$, $\mathbf{D}^k$ étant la direction du vecteur au k$^{\text{ième}}$ point considéré.

Le principe de l'algorithme de Marquardt consiste en une combinaison de cette technique du gradient avec l'approximation fournie par une série de Taylor. Débutant avec la première jusqu'à l'obtention d'un minimum, il utilise ensuite le développement en série de Taylor. Selon la formulation de Marquardt-Levenberg, $\mathbf{J}$ étant la matrice jacobienne de f (dérivées partielles $\partial f/\partial a_i$), chaque itération modifie l'estimation précédente, passant de $\mathbf{p}$ à $\mathbf{p}+\mathbf{q}$, par approximation linéaire :

$$f(\mathbf{p}+\mathbf{q}) = f(\mathbf{p}) + \mathbf{J}\mathbf{q}$$

Minimisant la somme des carrés des écarts S par $\nabla_q S = 0$, nous avons :

$$(\mathbf{J}^T\mathbf{J})\mathbf{q} = \mathbf{J}^T(f(\mathbf{p}+\mathbf{q}) - f(\mathbf{q}))$$

d'où on tire $\mathbf{q}$ déterminant la nouvelle estimation.

On trouvera par exemple dans Conway *et al.*(1970) une illustration de la méthode sur des données de croissance logistique.

Dans ce modèle (logistique simple de Verrhulst) les valeurs *a priori* des paramètres sont indiquées au départ à l'aide du tracé graphique des observations (au moins pour la valeur asymptotique et pour la position du point d'inflexion). Pour la logistique généralisée de Richards on peut consulter Causton et Venus qui étudient en détail l'estimation des paramètres de ce modèle de cinétique avec divers exemples de croissance végétale. On y trouvera des compléments dans le cas d'hétéroscédasticité dans la série des observations (pondération des données).

G.3. Analyse factorielle des courbes de croissance

Cette méthode d'analyse de *séries temporelles de croissance* (séries longitudinales constituées de la suite des mensurations) a été exposée au chapitre 21 du livre *Biomathématiques de la croissance* de R. Buis dans le cas où les observations portent sur un même ensemble d'individus suivis tout au long de la croissance. Nous avons vu que les *structures factorielles des variables* sont une image géométrique du processus présentant des singularités cinétiques intéressantes.

Autre point de vue, ce type de méthodologie peut s'intéresser préférentiellement, non aux variables, mais à la *représentation des individus*. Dans ce cas il s'agit de rechercher une *typologie des sujets* en fonction des caractéristiques des courbes individuelles de croissance. Les observations de chaque sujet sont soumises à un ajustement par une fonction de croissance donnée. Les variables soumises à une analyse en composantes principales sont les paramètres du modèle. Nous avons noté l'exemple de la logistique de Richards appliquée à la croissance radiale (circonférence du tronc) sur une population de Peupliers. Le but est de visualiser l'existence de groupes d'arbres, pouvant correspondre par exemple à des clones différents ou selon les conditions environnementales (Houllier, 1987 ; Caussinus et Ferré, 1989).

Cette question a été reprise par Abidi *et al.* (1995) où l'on trouve d'une part une discussion sur le choix de la métrique, et d'autre part le traitement de *séries individuelles incomplètes*. Cette dernière question est importante en pratique lorsque, en raison de contraintes expérimentales, tous les individus ne peuvent être mesurés aux mêmes âges. Un exemple est donné avec l'analyse de la croissance staturale de l'espèce humaine (de 6 à 19 ans) à partir du modèle de Preece et Baines I, modèle à 5 paramètres bien adapté à des courbes non sigmoïdes (voir chap. 10 du livre *Biomathématiques de la croissance* de R. Buis). Cette analyse précise les propriétés statistiques des estimations individuelles des paramètres. Elle explicite les deux représentations distinctes, celle des paramètres du modèle et celle des sujets. Cette analyse factorielle est conduite en vue de la recherche d'une *typologie des individus*, sans développer de conclusions sur la cinétique du processus lui-même (sur ce dernier point voir le chap. 21 du livre *Biomathématiques de la croissance* de R. Buis).

G.4. Modèle déterministe versus modèle aléatoire ?

Hormis les problèmes d'ajustement et de validation que nous venons de résumer, se pose pour le modélisateur la question du choix du type de modèle, strictement déterministe ou au contraire de nature stochastique. On note parfois qu'il convient d'assortir le premier d'un terme aléatoire afin de tenir compte de la variabilité des mesures. En dépit de l'importance de celle-ci, souvent considérée comme irréductible aux données biologiques, il reste essentiel de s'attacher à la mise au point de modèles dynamiques déterministes, dûment étudiés hors tout aléa, ainsi que nous l'avons expliqué dans la partie Introduction du livre *Biomathématiques de la croissance* de R. Buis.

Conclusion

En conclusion de ce chapitre, il convient surtout de bien distinguer ce qui relève de la *variabilité* intrinsèque des données et ce qui ressort de la *stabilité* du modèle lui-même, indépendamment du fait que tout modèle reste une construction dotée d'une certaine approximation en raison de processus ignorés ou mal connus qu'il ne prend pas en compte. Hors tout débat épistémologique sur le statut de la modélisation, il s'agit de souligner l'importance de la notion de *stabilité structurelle*, que nous avons notée par ailleurs. Divers exemples mettent en relief non seulement son importance théorique mais aussi les conséquences qui en découlent quant à l'interprétation et l'utilisation du modèle.

En connotation avec la formalisation des cinétiques de croissance, la dynamique des populations en fournit de bonnes illustrations (voir chap. D de ce site web compagnon). Ainsi tel modèle, descripteur phénoménologique des équilibres effectifs/ressources, peut conduire à la prédiction d'une émergence d'instabilités majeures, conséquence d'une forte sensibilité de la dynamique à la structure même du modèle. Il s'agit alors d'interpréter cette grande sensibilité structurelle en les rapportant non seulement à l'estimation numérique des paramètres, mais aussi, bien entendu, à la pertinence des composantes élémentaires du modèle telles que celui-ci les choisit et les formalise mathématiquement. Divers exemples donnés en plusieurs chapitres ont montré comment ont pu être proposées certaines modifications ou extensions de modèles classiques de base, visant, fût-ce empiriquement,

à corriger leur éventuelle inéquation. Plus généralement on conviendra que l'ajout d'aléas ne saurait éviter expérimentation et simulation sur le modèle lui-même par une série de va et vient permettant de diverses manières une amélioration de la formalisation.

Références

Sont privilégiées les références comportant à la fois un minimum de bases théoriques et de nombreux exemples diversifiés.

Abibi H., Pontier J., Borms J., Duquet W., 1995, *Rev. Stat. appl*., **43**, 55-72

Buis R., 2016, *Biomathématiques de la croissance, Le cas des végétaux*, 608 p., Coll. Grenoble Sciences, EDP Sciences

Causton D.R, Venus J.C., 1981, *The Biometry of plant growth*, 307 p., éd. Arnold

Dagnelie P., 2011, *Statistique théorique et appliquée*, 2 : *Inférence statistique,* 3e éd., 736 p., de Boeck

Fisher R.A. & Yates F., *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed., 1963, 146 p., Oliver and Boyd

Grivet J.-P., 2013, *Méthodes numériques appliquées pour le scientifique et l'ingénieur*, 391 p., Coll. Grenoble Sciences, EDP Sciences

Mather K., 1965, *Analyse statistique en Biologie*, trad. fr., 327 p., Gauthier-Villars

Tomassone R., Dervin C., Masson J.-P., 1993, *Biométrie*, *Modélisation de phénomènes biologiques*, 553 p., Masson

Sur des points particuliers d'estimation

Conway G.R., Glass N.R., Wilcox J.C., 1970, *Ecology*, **51**, 503-507

Marquardt D.W., 1963, *J. Soc. Ind. Appl. Math*., **11**, 431-441