Modélisation du processus d'apparition des feuilles par des durées successives -alternative aux modèles de régression

Sandra Plancade, Elodie Marchadier, Sylvie Huet, Adrienne Ressayre, Camille Noûs, Christine Dillmann

Modélisation du processus d'apparition des feuilles par des durées successives - alternative aux modèles de régression.

Sandra Plancade [1] & Elodie Marchadier [2] & Sylvie Huet [3] & Adrienne Ressayre[2] & Camille Noûs[4] & Christine Dillmann [2]

[1] *UR MIAT, INRAE, Université de Toulouse, France, sandra.plancade@inrae.fr*
[2] *GQE - Le Moulon, Université Paris-Saclay, INRAE, CNRS, AgroParisTech, France, elodie.marchadier@universite-paris-saclay.fr, adrienne.ressayre@universite-paris-saclay.fr, christine.dillmann@inrae.fr*
[3] *UR MaIAGE, INRAE, Université Paris-Saclay, France, shuet.inra@gmail.com*
[5] *Cogitamus Laboratory, France, camille.nous@cogitamus.fr* *

Résumé. Le processus d'apparition des feuilles ou *phyllochrone* caractérise le rythme de développement global des plantes annuelles. Dans le cadre de tests d'hypothèses (comparaison entre des conditions), le phyllochrone est usuellement analysé par régression linéaire en ignorant l'autocorrélation et les variations temporelles. Plus généralement, cette approche est classique pour tester les effets *genotype* × *environnement* sur des traits phénotypiques. Ses limites sont soulevées depuis une dizaine d'années mais seules quelques approches alternatives ont été développées, notamment des modèles de survie pour le taux de germination, probablement du fait d'une complexité d'implémentation accrue. Dans cette veine, nous proposons un modèle de durées successives pour le phyllochrone, à la fois plus réaliste et plus flexible que les approches de régression linéaire, qui permet d'une part de comparer le phyllochrone entre différentes des conditions et d'autre part d'étudier la dynamique du phyllochrone. Nous discuterons des limites et avantages des deux approches.

Mots-clés. Développement des plantes, phyllochrone, modélisation, modèles de durées successives, modèles de semi-markov.

Abstract. The leaf appearance process or *phyllochron* characterises the global development of annual plants. In the hypothesis testing context (comparison between condition), phyllochron is usually analysed by linear regression, ignoring autocorrelation and temporal variations. More generally, this approach is classic to test the *genotype*×*environment* effects on phenotypic traits. Concerns have been raised for a dozen years, but only a few alternative models have been proposed, notably survival analysis for germination, probably due to a higher complexity. In this spirit, we propose a successive time-to-event model for phyllochron, which is both more realistic and more flexible than linear regression, and enables us to compare phyllochron between conditions and to analyse phyllochron dynamics. We will discuss the limits and benefits of both models.

*Camille Noûs incarne la nature collégiale de nos travaux, se voulant un rappel de ce que la science doit à la *disputatio*, ainsi que du caractère intrinsèquement désintéressé, collaboratif et ouvert de la construction comme de la dissémination des savoirs.

Keywords. Plant development, phyllochron, modelisation, successive time-to-event models, semi-markov models.

1 Le processus d'apparition des feuilles ou phyllochron

La croissance et le développement des plantes consistent en un ensemble de processus temporels coordonnés, continus ou par états. Ces processus concernent des traits phénotypiques (croissance d'organes, etc) et des phénomènes biochimiques complexes. Dans le cadre de plantes annuelles, la mesure du processus d'apparition des feuilles ou *phyllochrone* est une méthode souvent utilisée pour déterminer le rythme de développement global de la plante. Le taux de croissance des plantes étant grandement affecté par la temperature, le temps est usuellement converti en *temps thermique* correspondant à un cumulé de température.

Les modèles de développement des plantes peuvent être regroupés en deux classes : les modèles prédictifs et les modèles de tests d'hypothèses. Les premiers incluent un ensemble de phénomènes complexes, à l'échelle d'une parcelle (He et al, 2012) ou de la plante (Functional Structural Plant Models, Vidal and Andrieu (2020)), pour prédire par exemple la production de biomasse. Ces modèles combinent des modélisations - le plus souvent déterministes - des phénomènes sous-jacents. Les modèles de tests d'hypothèses comparent usuellement une quantité (notamment un trait phénotypique), résumée par un paramètre, entre des classes ou conditions.

Dans le cas du phyllochrone, l'approche de tests d'hypothèses classique consiste à résumer le processus d'apparition des feuilles en un unique coefficient obtenu par régression linéaire du nombre de feuilles observées sur le temps, puis à tester l'effet de variables sur ce coefficient par des tests univariés ou des modèles mixtes (Padilla and Otegui, 2005; Correia et al, 2016). Mais le modèle sous-jacent à cette approche comporte deux limitations (i) l'hypothèse d'un taux d'apparition des feuilles constant, et (ii) les hypothèses liées au modèle de régression. L'utilisation en agriculture de modèles de régression (sans autocorrélation) pour modéliser des phénomènes autocorrélés est critiquée depuis une dizaine d'années (McNair et al, 2012; Onofri et al, 2019). Quelques approches alternatives basées sur des modèles de survie ont été développés, notamment dans le cadre de la germination (Humplík et al, 2020; Romano and Stevanato, 2020), mais demeurent marginales. Le modèle de phyllochrone que nous proposons se situe dans cette veine, en se basant sur une modélisation statistique plus pertinente que le modèle de régression.

2 Modèle de durées successives pour le phyllochrone

2.1 Hypothèses du modèle

Le processus d'apparition des feuilles ou phyllochrone d'une plante p, illustré en fig.1, est caractérisé par le vecteur $(Y_{p,f})_{f=1,\dots,F^{max}}$, où $Y_{p,f}$ est la durée entre l'apparition des feuilles

$(f-1)$ et f (ou entre le semis et l'apparition de la première feuille). On note également $H_{p,f} = \sum_{f'=1}^{f} Y_{p,f'}$ le temps d'apparition de la feuille f depuis le semis. On dispose de mesures répétées du nombre de feuilles $(t_{p,j}, X_{p,j})_{j=1,\dots,N_p}$ où $X_{p,j}$ est le nombre de feuilles de la plante p au temps $t_{p,j}$, mais les temps d'apparition des feuilles sont inconnus. On fait les hypothèses suivantes :

(H1) Les durées entre l'apparition des feuilles successives $(Y_{p,f})_f$ d'une même plante sont indépendantes.

(H2) La distribution de $Y_{p,f}$ dépend du rang de la feuille f, du génotype (groupe de plantes issues d'une même graine plusieurs générations auparavant) et de l'année. On note $\mu_{y,g,f} = \mathbb{E}[Y_{p,f}]$ pour une plante p issue du genotype g pour l'année y.

L'hypothèse (H1) implique que l'accélération ou le ralentissement du processus d'apparition des feuilles au cours de la saison résulte uniquement d'un processus déterministe à l'échelle du génotype. L'hypothèse (H2) stipule que les paramètres dépendent du temps "interne" de la plante via le rang de la feuille, considéré comme un stade de développement.

2.2 Inférence

Pour chaque plante, les observations répétées du nombre de feuilles peuvent être transformées en des intervalles (éventuellement infinis) dans lesquels les feuilles apparaissent : $H_{p,f} \in [\nu_{p,f}, \tau_{p,f})$. Or, les temps d'apparition des feuilles $(H_{p,f})_f$ ne sont pas indépendants (seules les durées entre l'apparition des feuilles le sont), donc la vraisemblance des observations pour une plante s'écrit comme une intégrale de dimension $F^{\max}$ et une maximisation directe n'est pas envisageable. Néanmoins, la maximisation de la vraisemblance conditionnellement aux variables latentes $(H_{p,f})_f$ est très simple : il s'agit d'inférer les paramètres de distributions unidimensionnelles à partir d'échantillons i.i.d. Ce contexte appelle donc naturellement à des algorithmes de type Expectation Maximization (EM).

Algorithme Monte Carlo EM sous hypothèse gaussienne.

L'algorithme Monte-Carlo EM consiste à générer les variables latentes $(H_{p,f})_f$ en partant d'une valeur initiale des paramètres, puis à réestimer les paramètres en maximisant la vraisemblance des données complètes (observées et latentes). Après une période de burn-in, les variables latentes générées à toutes les itérations sont utilisées, avec un poids décroissant (variante SAEM, Delyon et al (1999)). Dans notre contexte, les variables latentes conditionnellement aux observations suivent une distribution multivariée tronquée, et les algorithmes de rejet classiques deviennent inopérants dès que la dimension (le nombre total de feuilles) augmente. Mais sous hypothèse de normalité, des algorithmes efficaces ont été développés (package `R TruncatedNormal` , Botev and Belzile (2020)). Ainsi, dans une première version de notre modèle, nous avons supposé les durées $(Y_{p,f})_f$ gaussiennes.

Algorithme EM pour des modèles de semi-markov unidirectionnels

Sous les hypothèses (H1) et (H2), notre modèle correspond à un modèle de semi-markov (SMM) unidirectionnel, observé avec censure par intervalle. Les SMM sont des processus stochastiques multi-états qui généralisent les modèles de markov en combinant des transitions markoviennes et des durées d'états explicites (non nécessairement exponentielles, contrairement aux markov). Dans un contexte de temps discret, des équations forward-backward similaires aux algorithmes pour les SMM cachés permettent d'implémenter un algorithme EM pour des distributions quelconques des durées des états. A notre connaissance, aucune implémentation n'est publiquement disponible, et nous avons mis en oeuvre l'algorithme sous $\widehat{\mathrm{R}}$. La complexité de l'algorithme est quadratique en la durée totale d'observation qui est inversement proportionnelle au pas de discrétisation.

3 Application du modèle

On dispose d'un jeu de données sur le maïs (projet ITEMAIZE, Durand et al (2010)) comprenant des mesures du phyllochrone pour plusieurs centaines de plantes appartenant à neuf génotypes issus de deux variétés, réparties sur trois années.

3.1 Comparaison du phyllochrone entre groupes genotypiques

La différence de phyllochrone entre des classes est analysée par un test du χ^2 du rapport de vraisemblance, qui compare les modèles où les paramètres dépendent ou non de la classe. Ce test suppose l'indépendance entre les plantes, qui peut être invalidée par le design experimental. Ainsi, dans notre étude, les semis sont réalisés en rangées, créant une corrélation entre les plantes soumises à des conditions similaires. Ce biais peut être contourné par un test de permutation dans lequel la distribution de la statistique de test sous l'hypothèse nulle est obtenue en générant des classes aléatoires qui préservent le regroupement par rangée. Néanmoins, cette approche requiert un nombre suffisant de rangées car le nombre total de permutations possibles conditionne la précision de la p-value.

3.2 Etude de la dynamique du phyllochrone et test de l'hypothèse de phyllochrone constant

Notre modèle flexible permet d'explorer la structure de la dynamique du phyllochrone. La représentation du phyllochrone moyen $(\widehat{\mu}_{g,f})_f$ estimé pour chaque genotype fait apparaitre des variations à tous les rangs de feuilles (fig.2, première ligne). Afin d'extraire les variations significatives, des sous-modèles paramétriques sont considérés, sous forme de contraintes sur la fonction $f \mapsto \mu_{g,f}$: constante, linéaire, constante par morceaux (deux phases), bilinéaire; et le meilleur modèle est sélectionné par une combinaison du critère AIC et du test du χ^2 du rapport de vraisemblance.

Le phyllochrone montre un fort effet année avec d'importantes variations au cours de la saison en 2015, mais le modèle constant reste peu sélectionné y compris en 2014 et 2016. De plus, le modèle complet est souvent préféré aux sous-modèles paramétriques (fig.2, seconde ligne), confirmant la pertinence d'un modèle flexible. Par ailleurs, les modèles paramétriques facilitent l'interprétation en soulignant notamment les différences entre les variétés (fig.2, troisième ligne) et pourraient fournir une paramétrisation plus parcimonieuse du phyllochrone pour des modèles globaux (FSPM, modèles de culture).

4 Développements et perspectives

L'objectif de ce travail est d'explorer une alternative au modèle classique de phyllochrone par régression linéaire, en combinant une modélisation pertinente et la possibilité de tester l'effet de conditions. Mais cette première version comporte des limites plus ou moins difficilement dépassables.

4.1 Discussions et développements envisageables

Modélisation

Hypothèse gaussienne. La paramétrisation d'une durée par une distribution gaussienne dans le cadre du premier algorithme d'estimation requiert une déviation standart très grande devant la moyenne, ce qui est le cas pour une majorité mais pas tous les génotypes de notre jeu de données. Néanmoins, l'implémentation d'autres distributions (normale asymétrique discrétisée et beta-binomiale avec offset) avec l'approche par SMM montre des valeurs du phyllochrone $(\widehat{\mu}_{g,f})_f$ et des résultats des tests entre groupes génotypiques très semblables, ce qui indique une faible influence de l'hypothèse de normalité.

Hypothèse d'indépendance. L'hypothèse d'indépendance des temps $(Y_{p,f})_f$ entre l'apparition de feuilles successives est sujet à discussion parmi les biologistes. Une modélisation plus générale pourrait être considérée, notamment un modèle auto-régressif (AR). L'estimation sous hypothèse gaussienne reviendrait simplement à considérer une paramétrisation plus complexe de la matrice de covariance, mais au prix d'une augmentation du nombre de paramètres. Cependant, un modèle AR sortirait du cadre des SMM et ne permettrait plus d'utiliser les algorithmes dédiés.

Validation du modèle. Dans le cadre de données censurées par intervalle, les hypothèses portant sur les variables non-observées ne peuvent pas être directement vérifiées, mais la comparaison des observations et des quantités équivalentes prédites par le modèle (fig.4) montre une bonne reconstitution.

Procédure de test

L'approche par regression linéaire permet de prendre en considération des designs complexes, concernant les effets d'intérêt (e.g. design hiérarchique) ou les artefacts techniques, via des modèles mixtes. Dans le cadre de notre modèle, l'utilisation d'un test de permutation décrit en Section 3.1 permet de prendre en compte un effet batch sans l'hypothèse (arbitraire) d'additivité des effets présente dans les modèles mixtes, mais cette approche est limitée à des designs peu complexes.

Une alternative serait d'inclure les effets des variables d'intérêt dans le modèle. L'inclusion d'effets fixes constants ne pose pas de problèmes algorithmiques majeurs, autant pour l'approche gaussienne que par SMM, car l'étape de simulation réalisée séparemment pour chaque plante, qui représente la principale difficulté des deux algorithmes, serait inchangée. L'algorithme de SMM permet en outre l'inclusion d'une covariable longitudinale (climatique). Néanmoins, le passage au test requerrait des développements mathématiques conséquents, incluant des estimations approchés dans des algorithme de type EM. Etant donné la complexité de notre approche par rapport aux modèles linéaires classiques, il nous semble prioritaire d'évaluer quantitativement les bénéfices potentiels avant de poursuivre.

4.2 Modélisation par processus stochastique versus régression - quantification de l'impact relatif des deux approches

Modélisation plus réaliste. La différence conceptuelle entre les modèles de durées successives et régression (sans autocorrélation) est illustrée fig.3. Avec les modèles de régression, les observations générées pour une plante ne sont pas croissantes au cours du temps, en désaccord avec la réalité biologique. De plus, le modèle de durées successives conduit à une divergence du processus entre les plantes, qui correspond aux observations expérimentales.

Quantification du biais de l'approche par régression linéaire. Les tests classiques basés sur le modèle linéaire peuvent présenter un biais par l'absence de prise en compte de la variance de l'estimateur du taux d'apparition des feuilles, ainsi qu'un biais et/ou une perte de puissance si le phyllochrone dévie du modèle constant. Néanmoins, ces modèles demeurant beaucoup plus simples à implémenter et quasi-exclusivement utilisés, il serait intéressant d'évaluer quantitativement ces biais, par une étude de simulations, ou éventuellement par l'analyse de la distribution de la statistique de test. Dans ce cadre, notre modèle peut constituer le modèle de référence pour générer les données.

Modèles de régression non-linéaires. L'approche par régression linéaire ne permet pas d'étudier la dynamique du phyllochrone, mais quelques modèles de régression non-linéaires ont été développés (Clerget and Bueno, 2013; Baumont et al, 2019). Le premier article propose un test de sélection de modèles, sans inclure d'effet plante. La procédure de test considère le nombre d'observations (et non de plantes) comme la taille d'échantillon, conduisant notamment à des conclusions aberrantes dans le cas asymptotique de mesure en temps continu. Plus généralement, ces approches sont suceptibles d'être biaisées plus fortement que les tests basés sur le phyllochrone linéaire, puisque l'hypothèse d'indépendance des résidus du modèle de régression est directement utilisée dans la procédure de test.

Perspective globale : analyse de phénomènes dynamiques par des modèles de régression. Au delà du phyllochrone, dans le cadre d'analyse *genotype* × *environnement*, le phénotype ou la dynamique d'intérêt sont classiquement réduits à un coefficient auquel on applique un modèle mixte. L'exemple du phyllochrone pourrait permettre de poser un cadre de réflexion sur ce sujet, et évaluer dans quels contextes et pour quelles questions biologiques il serait pertinent de développer des modèles basés sur des processus stochastiques, plus réalistes en termes de modélisation mais plus complexes en termes d'implémentation.

References

Baumont M, Parent B, Manceau L, Brown HE, Driever SM, Muller B, Martre P (2019) Experimental and modeling evidence of carbon limitation of leaf appearance rate for spring and winter wheat. *Journal of Experimental Botany* 70(9):2449–2462

Botev Z, Belzile L (2020) Truncatednormal: Truncated multivariate normal and student distributions URL https://CRAN.R-project.org/package=TruncatedNormal, r package version 2.2

Clerget B, Bueno CS (2013) The effect of aerobic soil conditions, soil volume and sowing date on the development of four tropical rice varieties grown in the greenhouse. *Functional Plant Biology* 40:79–88

Correia LE, Matsunaga F, Alvim CA, Rakocevic M (2016) Phyllochron, leaf expansion and life span in adult coffea arabica l. plants: Impact of axis order, growth intensity period and emitted leaf position. *2016 IEEE International Conference on Functional-Structural Plant Growth Modeling, Simulation, Visualization and Applications (FSPMA)* pp 38–43

Delyon B, Lavielle M, Moulines E (1999) Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics* 27(1):94–128

Durand E, Tenaillon MI, Ridel C, Coubriche D, Jamin P, Jouanne S, Ressayre A, Charcosset A, Dillmann C (2010) Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. *BMC evolutionary biology* 10

He J, Le Gouis J, Stratonovitch P, Allard V, Gaju O, Heumez E, Orford S, Griffiths S, Snape JW, Foulkes MJ, Semenov MA, Martre P (2012) Simulation of environmental and genotypic variations of final leaf number and anthesis date for wheat. *European Journal of Agronomy* 42:22–33

Humplík JF, Dostál J, Ugena L, Spíchal L, De Diego N, Vencálek O, Furst T (2020) Bayesian approach for analysis of time-to-event data in plant biology. *Plant Methods* 16(1):14

McNair JN, Sunkara A, Frobish D (2012) How to analyse seed germination data using statistical time-to-event analysis: non-parametric and semi-parametric methods. *Seed Science Research* 22(2):77–95

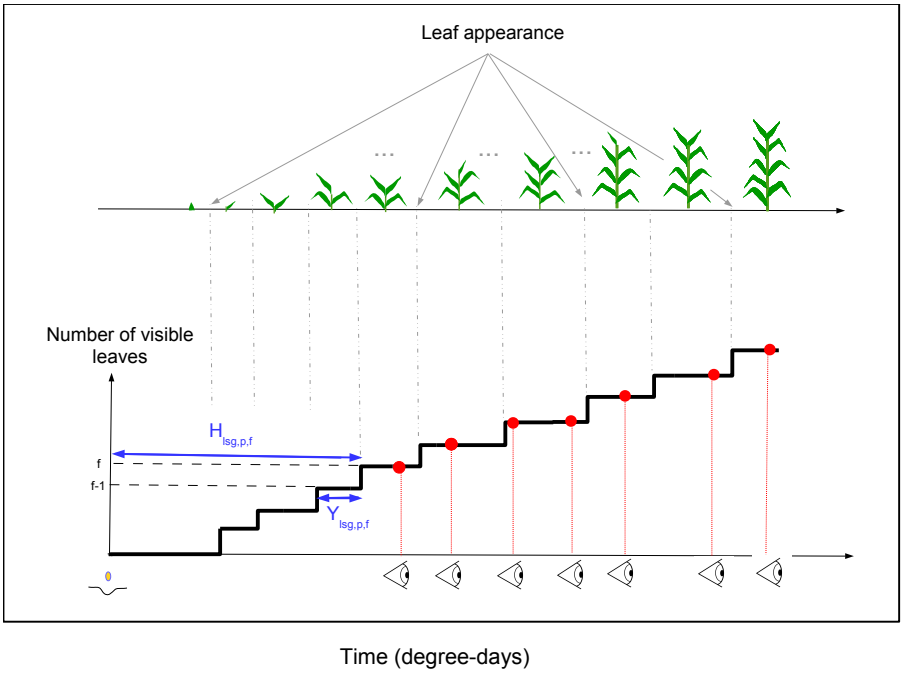

Figure 1: **Illustration du processus d'apparition des feuilles ou** *phyllochrone* **pour un plant de mas**

Onofri A, Piepho HP, Kozak M (2019) Analysing censored data in agricultural research: A review with examples and software tips. *Annals of Applied Biology* 174(1):3–13

Padilla JM, Otegui ME (2005) Co-ordination between leaf initiation and leaf appearance in field-grown maize (Zea mays): genotypic differences in response of rates to temperature. *Annals of Botany* 96(6):997–1007

Romano A, Stevanato P (2020) Germination data analysis by time-to-event approaches. *Plants* 9(5)

Vidal T, Andrieu B (2020) Contrasting phenotypes emerging from stable rules: A model based on self-regulated control loops captures the dynamics of shoot extension in contrasting maize phenotypes. *Annals of Botany* 126(4):615–633

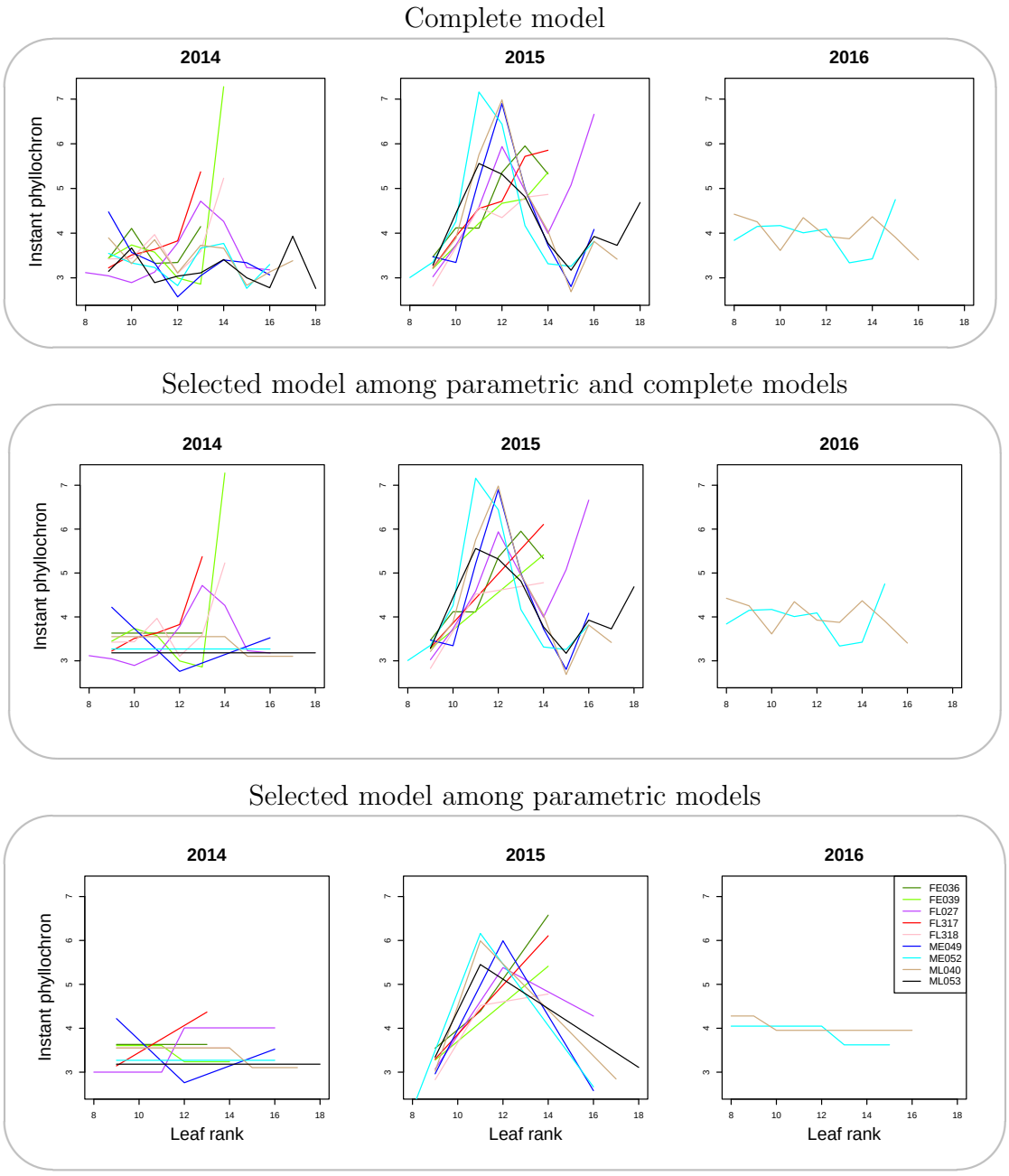

Figure 2: **Valeurs estimées du phyllochrone $(\mu_f)_f$ pour chaque genotype et chaque année.** Première ligne : modèle complet; deuxième ligne : modèle sélectionné parmi les modèles paramétriques et le modèle complet; troisième ligne : modèle sélectionné parmi les modèles paramétriques. Pour chaque combinaison genotype-année, chaque modèle est comparé au modèle constant par le test du χ^2 du rapport de vraisemblance, et parmi ceux dont la p-value est inférieure à 0.01, le modèle sélectionné est celui présentant le plus petit AIC. Si aucun modèle ne présente de p-value inférieure à 0.01, le modèle constant est sélectionné.

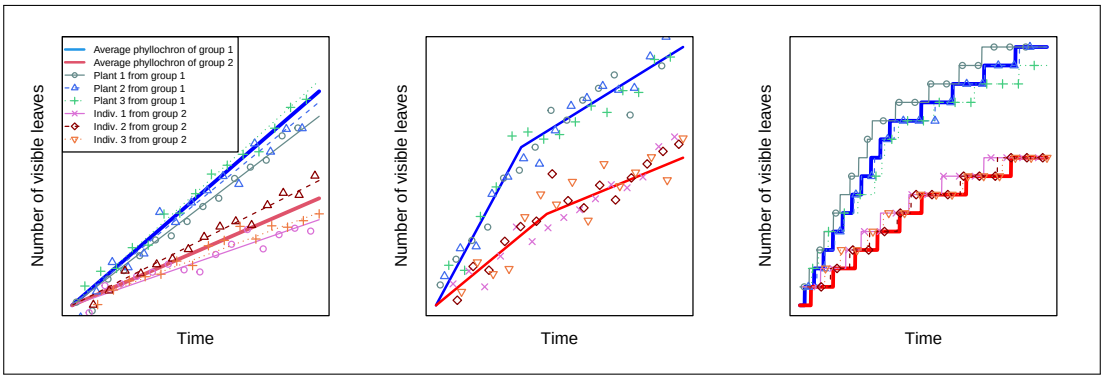

Figure 3: **Différence conceptuelle entre les modèles de régression et de durées successives.** Trois types de modèles du phyllochrones (deux issus de la littérature, et notre modèle) sont représentés. Gauche : modèle de régression linéaire avec effet individuel; centre : modèle bilinéaire sans effet individuel; droite : modèle de durées successives avec deux phases sous-jacentes dans le phyllochrone. Pour chaque modèle, le phyllochrone de six plantes appartenant à deux groupes est simulé. Les traits pleins représentent le phyllochrone moyen à l'échelle du groupe, les lignes fines représentent le phyllochrone de chaque plante, et les points des observations à des temps discrets. Pour les deux modèles de régression, les valeurs du nombre de feuilles générées ne sont pas croissantes au cours du temps, contrairement au modèle de durées successives. Pour ce modèle, les segments en trait plein correspondent aux $(\mu_f)_f$ et les segments en pointillés correspondent aux $(Y_{p,f})_f$.

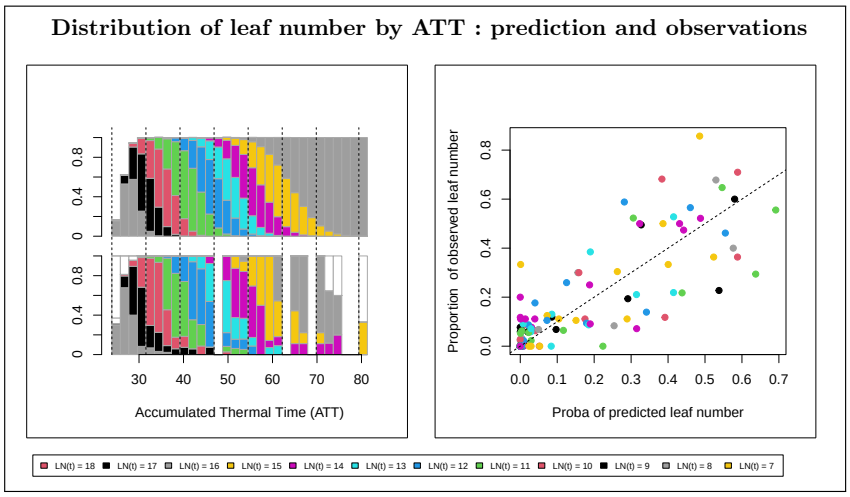

Figure 4: **Validation du modèle : nombres de feuilles observés et prédits pour un génotype** . Gauche : le graphe du haut montre la distribution du nombre de feuilles visibles $LN(t)$ prédit par le modèle dans chaque intervalle de temps thermique. Le graphe du bas présente la distribution des $LN(t)$ observés sur l'ensemble des plantes, pour les mesures réalisées pendant chaque intervalle de temps thermique. Droite : chaque point représente la proportion des observations pour un rang folliaire donné et dans un intervalle de temps thermique donné en fonction de la même quantité prédite par le modèle; chaque couleur correspond à un rang folliaire et le blanc correspond aux rangs folliaires non modélisés.