Détection de motifs disruptifs au sein de plantes : une approche de quotientement/classification d'arborescences

Pierre Fernique [1] & Jean-Baptiste Durand [2] & Yann Guédon [3]

[1] *CIRAD, AGAP et Inria, Virtual Plants F-34095 Montpellier, France ; pierre.fernique@inria.fr*

[2] *Univ. Grenoble Alpes, Laboratoire Jean Kuntzmann et Inria, Mistis F-38041 Grenoble, France ; jean-baptiste.durand@imag.fr*

[3] *CIRAD, AGAP et Inria, Virtual Plants F-34095 Montpellier, France ; guedon@cirad.fr*

Résumé. Les modèles de détection de ruptures multiples pour séquences sont transposés aux arborescences. L'objectif est de quotienter une arborescence en sous-arborescences homogènes. Comme les algorithmes optimaux de segmentation de séquences ne peuvent être transposés aux arborescences, nous proposons ici une méthode heuristique permettant de segmenter efficacement une arborescence. Les sous-arborescences obtenues sont ensuite groupées dans une phase de post-traitement car des sous-arborescences disjointes relativement similaires sont observées dans les canopées d'arbre. Ces modèles sont illustrés par le cas du manguier où les collections de sous-arborescences permettent d'identifier les motifs disruptifs (juxtaposition de sous-arborescences végétatives, florifères ou en pause) observés dans les canopées.

Mots-clés. Architecture des plantes ; détection de ruptures multiples ; manguier ; motif arborescent ; quotientement d'arborescence ; regroupement d'arborescences ; segmentation d'arborescences

Abstract. Multiple change-point models for path-indexed data are transposed to tree-indexed data. The objective of multiple change-point models is to partition a heterogeneous tree into homogeneous subtrees. Since optimal algorithms for segmenting sequences cannot be transposed to trees, we propose here an efficient heuristic for tree segmentation. Segmented subtrees are grouped together in a post-processing phase since similar disjoint patches are often observed in tree canopy . Application of such models is illustrated on mango tree where subtrees are assimilated to plant patches and clusters of patches to patch types (e.g. vegetative, flowering or resting patch).

Keywords. multiple change-point detection ; mango tree ; plant architecture ; quotient tree ; tree clustering ; tree pattern ; tree segmentation

1 Introduction

Comme d'autres arbres tropicaux, le manguier est caractérisé par de forts asynchronismes phénologiques [1] entre arbres et au sein des arbres, entraînant des motifs disruptifs au niveau de la canopée [2] (Chacko, 1986) caractérisés par la présence de taches de couleur représentant des amas homogènes d'entités végétatives ou reproductives. Ces asynchronismes concernant de plus ou moins grands systèmes ramifiés entraînent divers problèmes agronomiques, tels que l'utilisation répétée des pesticides pour protéger les stades phénologiques sensibles aux parasites, ou une période de maturité des fruits trop étendue, conduisant à des difficultés d'organisation de la récolte des fruits. L'objectif est de définir une méthodologie statistique permettant de détecter ces motifs disruptifs. Cette approche est particulièrement intéressante car elle permettrait de quantifier ce phénomène et, plus généralement, de mettre en évidence des motifs disruptifs pour des espèces où ces motifs ne sont pas directement apparents car exprimés à une échelle plus fine.

Les données indexées par des arborescences sont utilisées comme des représentations de l'architecture de la plante et, il est supposé que ces motifs peuvent être assimilés à une partition de l'arborescence en sous-arborescences. On suppose donc qu'il existe des sous-arborescences à l'intérieur desquelles les caractéristiques des entités botaniques suivent la même loi et qu'entre ces sous-arborescences, ces caractéristiques suivent des lois différentes. Bien que la présence de ces motifs disruptifs sur les plantes soit un phénomène spatio-temporel, nous nous concentrons ici sur son aspect spatial à une date donnée. Un tel point de vue entraîne de nombreuses valeurs manquantes au sein des arborescences car à une date donnée ce sont principalement les sommets correspondant aux feuilles de l'arborescence qui sont observés. Les modèles statistiques classiques pour les donnes indexées par des arborescences basés sur des hypothèses de Markov (voir Durand et al, 2005 par exemple) ne sont donc plus pertinents car les sommets internes, et donc les transitions, ne sont pas observés. La stratégie choisie est la recherche de changements brusques dans les proportions des types d'entités au sein de l'arborescence. Ceci est l'analogue du problème de segmentation de séquence (Picard et al, 2007) mené sur des arborescences. Il est à noter que les méthodes exactes pour déterminer la segmentation la plus probable d'une séquence ne peuvent pas être transposées à des données arborescentes. Nous proposons donc ici d'utiliser un algorithme de type glouton pour quotienter les arborescences. Comme l'ont souligné Picard et al (2007), la sortie de la procédure de quotientement est une collection de sous-arborescences où chaque élément est considéré comme différent des autres alors que deux sous-arborescences non-adjacentes peuvent être très similaires. Suite à l'étape de quotientement nous proposons donc une étape de classification d'arborescences basée sur des modèles de mélange afin d'identifier les sous-arborescences similaires.

1. Au niveau de la floraison et/ou de la croissance des pousses
2. l'étendue du couvert de la plante

2 Méthodologie

La représentation en arborescence des plantes. Compte tenu de la méthodologie décrite dans Durand et al (2005), une structure arborescente peut être utilisée comme une représentation de l'architecture de la plante. Les données notées $\bar{x} = (x_t)_{t\in\mathcal{T}}$ sont donc indexées par une arborescence : $\mathcal{T} \subset \mathbb{N}$ est l'ensemble de sommets de l'arborescence $\tau = (\mathcal{T}, \mathcal{E})$ et $\mathcal{E} \subset \mathcal{T} \times \mathcal{T} \setminus \mathcal{R}$ est l'ensemble des arêtes dirigées. $\mathcal{R}$ est l'ensemble des racines et $\mathcal{L}$ l'ensemble des feuilles de τ. Nous considérons ici que τ est *stricto sensu* une arborescence et que la seule racine de τ est notée r. $\mathrm{pa}(.)$ représente le parent, $\mathrm{ch}(.)$ l'ensemble des enfants, $\mathrm{de}(.)$ l'ensemble des descendants et $\mathrm{nd}(.)$ l'ensemble des non-descendants d'un sommet. Ces notations s'appliquent également à un ensemble de sommets. Les versions en majuscules indiquent la fermeture de l'ensemble correspondant, par exemple,

$$\forall t \in \mathcal{T},\ \mathrm{De}(t) = \mathrm{de}(t) \cup \{t\}.$$

Pour tout ensemble $\mathcal{A} \subseteq \mathcal{T}$, $\tau_{\mathcal{A}}$ est le sous-arborescence induite par $\mathcal{A}$. Le degré entrant d'un sommet t d'un arbre τ est noté $\deg_{\tau}^{-}(t)$. Il est égal à 0 si le sommet est une racine et à 1 sinon. Dans la suite nous considérerons que $\bar{x}$ est la réalisation d'un processus stochastique $\bar{X} = (X_t)_{t\in\mathcal{T}}$ à valeurs dans l'espace d'observation $\mathcal{X}$ tel que $\mathcal{X} \subset \mathbb{N}$.

Les modèles de quotientement. Un modèle de quotientement est défini par un quotientement de sommets, noté Π, de sorte que chaque quotient induit *stricto sensu* une arborescence. Les sommets d'un même quotient sont supposés indépendants et identiquement distribués. La paramétrisation d'un modèle de quotientement est donc définie par ces quotients et complétée par une loi d'observation par quotient. En conséquence, la log-vraisemblance $\mathcal{L}(\bar{x}; \Pi, \theta_{\Pi})$ du modèle se décompose comme suit :

$$\mathcal{L}(\bar{x}; \Pi, \theta_{\Pi}) = \sum_{\pi\in\Pi} \sum_{v\in\pi} \log f_{\pi}(x_v),$$

où $f_{\pi}(\cdot)$ désigne la loi d'observation du quotient $\pi \in \Pi$ et θ_{Π} l'ensemble des paramètres de ces lois d'observation. Les quotients de Π peuvent également être identifiés par l'ensemble des points de rupture, noté $\mathcal{P}$. Chacun de ces points correspond à la racine de l'arborescence induite par le quotient considéré

$$\forall \Pi \in \mathfrak{P}(\mathcal{T}),\ \mathcal{P} = \left\{ t \in \mathcal{T} \,\middle|\, \exists \pi \in \Pi,\ [t \in \pi] \wedge \left[\deg_{\tau_\pi}^{-}(t) = 0\right] \right\},$$

où $\mathfrak{P}(.)$ désigne l'ensemble des parties d'un ensemble . On note $\nu(\cdot)$ la fonction qui renvoi le quotientement associé à un ensemble de points de rupture :

$$\begin{array}{rccc} \nu & : & \mathfrak{P}(\mathcal{T}) & \to & \mathfrak{P}(\mathcal{T}) \\ & & \mathcal{P} & \mapsto & \Pi \end{array}.$$

Inférence des quotients Dans notre cas, étant donné un quotientement Π, l'estimation des lois d'observation est un simple problème d'estimation au sens du maximum de vraisemblance au sein de chaque quotient. Par contre, étant donné un nombre K de quotients, trouver le quotientement qui maximise la log-vraisemblance est complexe. Les méthodes pour déterminer le quotientement optimal pour les séquences ne peuvent être transposées au cas des arborescences. Nous proposons donc une approche heuristique pour trouver une solution localement optimale (voir Hawkins, 1976 pour une approche similaire sur les séquences). Soit $\mathcal{P}^{(k)}$ l'ensemble des points de rupture associé à $k+1$ quotients, correspondant à un optimum local de la log-vraisemblance. Par définition, $\mathcal{P}^{(0)}$ est l'ensemble des points de ruptures qui induit un quotient et contient donc uniquement la racine de l'arborescence, $\mathcal{P}^{(0)} = \{r\}$. Trouver le point de rupture supplémentaire de $\mathcal{P}^{(1)}$ qui maximise la log-vraisemblance du modèle de quotientement avec deux quotients est facilement obtenu en testant successivement tous les sommets non-racines comme point de rupture couplé à la racine :

$$\mathcal{P}^{(1)} = P^{(0)} \cup \left\{ \arg\max_{t \in \mathcal{T}} \left\{ \mathcal{L}\left(\bar{x}; \nu\left(\mathcal{P}^{(0)} \cup \{t\}\right), \theta_{\nu\left(\mathcal{P}^{(0)} \cup \{t\}\right)}\right)\right\}\right\}.$$

Le quotientement optimal d'une arborescence en 2 quotients est donc facilement obtenu. Le principe de notre heuristique est donc d'utiliser ce principe pour construire le quotientement de manière itérative. Dans le but de diminuer la probabilité d'être piégé dans un optimum local, pour chaque étape nouveau point de rupture trouvé, la suppression de points de ruptures est testée jusqu'à ce qu'une suppression n'augmente pas la log-vraisemblance de l'optimum local correspondant.

Sélection du nombre de quotients. Si le nombre des quotients est inconnu, il doit être sélectionné. Ce problème peut être rapporté au contexte plus général de la sélection de modèles, avec, comme dans le cas des séquences, la nécessité d'utiliser des critères adaptés au cas des modèles de quotientement. Dans notre contexte comprenant observations catégorielles et valeurs manquantes, les critères de vraisemblance pénalisée classiques sélectionnent des modèles sur-paramétrés et ne conviennent donc pas. Nous avons donc considéré les heuristiques de pente en utilisant la méthode d'estimation de la pente dirigée par les données proposée par Baudry et al (2012).

Classification des sous-arborescences. Les modèles de quotientement permettent de détecter les sous-arborescences telles que les observations ne changent pas substantiellement au sein de chaque sous-arborescence, mais changent fortement entre deux sous-arborescences adjacentes. La présence de sous-arborescences non-adjacentes similaires nous a amené à supposer :

— Qu'il existe un petit nombre de types de quotients et que tous les sommets d'un quotient sont du même type.

— Que les sommets dans un même quotient sont indépendants et identiquement distribués étant donné le type de ce quotient.

L'algorithme EM et l'affectation des quotients au sens du maximum *a posteriori* des modèles de mélange standards McLachlan and Peel (2000), sous la contrainte que les sommets appartenant à un même quotient sont affectés à la même composante, ont donc été utilisés afin de grouper les sous-arborescences.

3 Application à la détection des motifs disruptifs au sein de manguiers

Conception expérimentale. Le verger expérimental est situé à la station de recherche du Cirad [3] à Saint-Pierre, île de la Réunion. Cinq manguiers ont été décrits pour cinq variétés différentes Dambreville et al (2013) sur 3 ans. Les différentes dates considérées au cours de ces années amènent à un total de 181 arborescences à quotienter où chaque entité observée est florifère (F), végétative (V) ou au repos (R) :

$$\mathcal{X} = \{\mathrm{F}, \mathrm{V}, \mathrm{R}\}.$$

Quotientement des arborescences. La méthode d'heuristique de pente nécessitant le calcul de modèles sur-paramétrés, nous avons donc considéré jusqu'à 20 points de ruptures par arborescences. Sur les 181 arborescences, seulement 132 ont été quotientées avec succès. Ces échecs sont principalement dus à la présence d'arbres avec un niveau de bruit très faible, et donc les modèles sur-paramétrés ne pouvaient pas être estimés. Notons que même si nous n'avons pas considéré ces arborescences, ces échecs pourraient être considérés comme une indication d'homogénéité initiale des arborescences concernées. Suite à l'étape de quotientement nous étions en présence de 608 sous-arborescences de compositions et tailles différentes à classer. Seulement quelques sous-arborescences de hauteur 0 ont été détectées (6%), indiquant un sur-quotientement relativement faible.

Regroupement des arborescences. Pour le modèle de mélange, nous avons considéré trois états différents pour les sous-arborescences. D'après les lois d'observation

$$\begin{aligned} f_0(\mathrm{F}) &= 0.7, & f_0(\mathrm{V}) &= 0.04, & f_0(\mathrm{R}) &= 0.26, \\ f_1(\mathrm{F}) &= 0.08, & f_1(\mathrm{V}) &= 0.79, & f_1(\mathrm{R}) &= 0.13, \\ f_2(\mathrm{F}) &= 0.2, & f_2(\mathrm{V}) &= 0.08, & f_2(\mathrm{R}) &= 0.72, \end{aligned}$$

les types de quotients sont bien séparés et biologiquement interpretables : l'état 0 correspond aux quotients florifères, l'état 1 aux quotients végétatifs et l'état 2 correspond aux

3. Centre français de recherche agricole pour le développement international

quotients au repos. D'après les poids

$$\pi_0 = 0.22, \quad \pi_1 = 0.46, \quad \pi_2 = 0.32,$$

on remarque une légère dominance des quotients végétatifs, mais tous les types de quotients sont clairement présents. Cet excès de quotients végétatifs est biologiquement interprétable puisque les manguiers observées étaient jeunes et donc pas encore dans leur régime de production permanente. Bien qu'il y ait un certain degré d'opposition entre quotients végétatifs et florifères, chacun de ces quotients a tendance à contenir aussi des entités au repos.

Bibliographie

[1] J.-P. Baudry, C. Maugis, and B. Michel (2012), *Slope heuristics : overview and implementation*, Statistics and Computing, 22, 455–470.

[2] E. Chacko (1986), *Physiology of vegetative and reproductive growth in mango (*Mangifera indica *L.) trees*, In *Proceedings of the First Australian Mango Research Workshop*, volume 1, pages 54–70, CSIRO Australia, Melbourne.

[3] A. Dambreville, P. Fernique, C. Pradal, P.-E. Lauri, F. Normand, Y. Guédon, and J.-B. Durand (2013), *Deciphering mango tree asynchronisms using Markov tree and probabilistic graphical models*, In R. Sievänen, E. Nikinmaa, C. Godin, A. Lintunen, and P. Nygren, editors, FSPM2013 - 7th International Workshop on Functional-Structural Plant Models, pages 210–212, Saariselkä, Finlande.

[4] J.-B. Durand, Y. Guédon, Y. Caraglio, and E. Costes (2005), *Analysis of the plant architecture via tree–structured statistical models : the hidden Markov tree models*, New Phytologist, 166 (3), 813–825.

[5] D. M. Hawkins (1976), *Point estimation of the parameters of piecewise regression models*, Applied Statistics, 25 (1), 51–57.

[6] G. McLachlan and D. Peel (2000), *Finite mixture models*, Wiley New York.

[7] F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin (2007), *A segmentation/clustering model for the analysis of array CGH data*, Biometrics, 63 (3), 758–766.