Contributions to hidden Markov models and applications to plant structure analysis

Jean-Baptiste Durand

Université Grenoble Alpes

École doctorale "Mathématiques, Sciences et Technologies de l'Information, Informatique"

Spécialité doctorale "Mathématiques Appliquées et Informatique"

Mémoire en vue d'obtenir une habilitation à diriger des recherches

présentée par Jean-Baptiste Durand

Mars 2020

Contributions to hidden Markov models and applications to plant structure analysis

---

Reviewers:

Pr. G. Buck-Sorlin	IRHS, AgroCampus Ouest, France
Pr. H. Holzmann	Philipps-Universität, Marburg, Germany
Pr. G.J. McLachlan	University of Queensland, Brisbane, Australia

Contents

Abstract:
In this document, I present various contributions to hidden Markov models on graphs and more generally, to the statistical analysis of graphical data, with a particular focus on tree graphs. In a part following an introduction, three main types of problems in tree analysis are exposed: hidden Markov tree models to predict tree shapes and perform vertex segmentation, edit distances to perform clustering at whole-tree scale and multiple change-point detection on trees. Then some more detailed focus is given to multivariate count modelling, which is one of the main problem to be solved in hidden Markov tree estimation. This is addressed using the theory of probabilistic graphical models. A presentation of three specific contributions to hidden Markov chain modelling follows: quantifying state uncertainty, optimal timeout modelling and latent chain coupling. Lastly, an overview of different approaches applied to several plant growth modelling problems is exposed, preceding some conclusions and general perspectives.
Keywords:
Hidden Markov models, Probabilistic graphical models, Statistical analysis of tree-structured data, Applications to plant structure analysis.

Résumé :
Dans ce rapport, je présente diverses contributions aux modèles de Markov cachés sur graphes et plus généralement, à l'analyse statistique de données graphiques, avec une attention privilégiée portée aux arborescences. Après l'introduction, trois types de problèmes principaux en analyse d'arborescences sont exposés: les arbres de Markov cachés pour prédire des formes arborescentes et les segmenter à l'échelle des sommets, les distances d'édition pour en réaliser la classification non-supervisée à l'échelle d'arborescences entières et la détection de ruptures multiples sur arborescences. Une présentation plus détaillée est ensuite donnée des modèles pour comptages multivariés, qui est l'un des verrous méthodologiques essentiels à lever pour l'estimation d'arbres de Markov cachés. Le problème est abordé sous l'angle des modèles probabilistes graphiques. S'ensuit une présentation de trois contributions spécifiques à la modélisation par chaînes de Markov cachées: la quantification de l'incertitude sur les états, la modélisation du délai de mise en veille optimal et le couplage de chaînes latentes. L'avant-dernière partie présente différentes méthodes adaptées à divers problèmes de modélisation de la croissance et de la structure des plantes; elle est suivie de conclusions et perspectives générales.
Mots clés :
Modèles de Markov cachés, Modèles probabilistes graphiques, Analyse statistique d'arborescences, Applications à l'analyse de la structure des plantes.

List of acronyms

Scientific terms

AMP	alternative Markov property (on chain graphs) Andersson / Madigan / Perlman
AS	annual shoot
BBI	biennial bearing index
BIC	Bayesian information criterion
BLUP	best linear unbiased predictor
BNP	Bayesian non-parametric
BSE	bovine spongiform encephalopathy
CIC-HMT	conditionally independent children - hidden Markov tree
DAG	directed acyclic graph
DWT	discrete wavelet transform
EEG	electroencephalogram
EM	expectation maximisation
GHMM	graphical hidden Markov model
GLM	generalised linear model
GLMM	generalised linear mixed model
GU	growth unit
GS	Granny Smith
HMC	hidden Markov chain
HMM	hidden Markov model
HMOT	hidden Markov out-tree
HMT	hidden Markov tree
HSMC	hidden semi-Markov chain
ICL	integrated classification likelihood
LAD	leaf area density
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
LMM	linear mixed model
LWF	Lauritzen / Wermuth / Frydenberg (particular Markov property on chain graphs)
MANOVA	multivariate analysis of variance
MAP	maximum a posteriori
MCMC	Monte-Carlo Markov chain
MDP	Markov decision process
MLE	maximum likelihood estimation (or estimator)
MODWT	maximum overlap discrete wavelet transform
MRF	Markov random field
NN	neural network
P / NP	polynomial / nondeterministic polynomial (classes of computational problems in computational complexity theory)
PC (A)	principal component (analysis)

PDAG	partially directed acyclic graph
PGM	probabilistic graphical model
POMDP	partially observed Markov decision process
QTL	quantitative trait locus
SAM	shoot apical meristem
SCA	space colonization algorithm
STK	Starkrimson
UG	undirected graph
VBEM	variational Bayesian expectation maximisation
VOMC	variable-order Markov chain
WS	water-stressed
WW	well-watered

Names of institutions

ACI	Action concertée incitative (former call of ANR)
AGAP[1], AMAP[2]	Laboratories in plant science, Montpellier
ANR	French national research agency
CIRAD[3]	French agricultural research and international cooperation organization
Inria[4]	National institute for research in digital science and technology
LJK	Laboratoire Jean Kuntzmann in Grenoble (applied mathematics)

[1] https://umr-agap.cirad.fr/
[2] http://amap.cirad.fr/
[3] https://www.cirad.fr/
[4] https://www.inria.fr/

1. Introduction

In this document, several contributions to the analysis of tree-structured data are presented. Some extensions to the more general case of graphical data or oppositely, specific developments dedicated to sequence analysis are addressed. We mainly focus on statistical modelling problems, although these are occasionally complemented by other approaches issued from discrete optimization and combinatorics.

In the early 2000s, the existing methods for the statistical analysis of tree-structured data were rather rare. One possible reason for this is that such type of data were often long and costly to acquire. Remarkable exceptions could be found, in the one hand, in document analysis and categorisation and in the other hand, in wavelet analysis. In the former field, trees represent document organisation at several scales: columns, divided into text zones vs. images and then, paragraph at a finer scale. In such contexts, tree sizes are rather low and trees may have arbitrary shapes. In the latter, trees are induced by the wavelet decompositions of signals or images. Tree sizes are comparable with signal lengths and thus, are usually quite larger than in the case of documents. Wavelet trees have deterministic shapes (binary or quad trees). In both contexts however, trees share specific features. Vertices have external (i.e., non-topological) quantitative or qualitative properties, meaning that these are no deterministic functions of tree shape itself. Trees are so-called *piecewise homogeneous*, referring to the existence of underlying zones within which the statistical properties of vertices are comparable, whereas they change abruptly between zones. The second feature is dependencies, which in the end stochastically determines zone extents. To incorporate both features, hidden Markov models (HMMs) were proposed to either segment trees or predict their vertex properties (Crouse *et al.*, 1998; Diligenti *et al.*, 2001).

In contrast, in phylogenetics or cell division analysis, trees could be far more variable, but they were mostly unobserved, except for the leaves. In phylogenetics, building the tree is the main aim of models, whereas in the case of cell populations, this is inference of the tree growth dynamics. Quite a number of sophisticated methods emerged, mainly derived from branching processes (Haccou *et al.*, 2005), but they were mostly dedicated to assess the effect of different covariates, competition or interactions between species or types of cells, or to validate assumptions on the times between tree splittings. However, they did not aim directly at modelling tree shapes.

In some specific fields of science, however, statistical modelling of tree shapes, their variability or dependencies with respect to covariates was of uttermost importance. Among these fields were dendritic tree analysis (Polavaram, 2014) and plant growth modelling. In the latter field, considerable effort was dedicated to the acquisition of trees with their topologies and local features, but there was a lack of appropriate statistical models to answer some biological questions which, presumably, could find their answers in the collected data if adequately analysed. The requirements of the models firstly consisted, as in the examples in document categorisation and wavelet analysis, in accounting for piecewise homogeneity and local dependencies. This was not exactly sufficient, since there were also global dependencies and various other sources of variability in the structure and the features: environment and genotype mainly, together with their interactions.

In collaboration with different laboratories and teams (AMAP, AGAP, Inria Virtual Plants and Mosaic), we developed dedicated models to address the challenge of statistical analysis of tree structures. Our strategy and thinking were inspired by the methods developed in the context of analysis frameworks for biological sequences, particularly genomic data. Indeed, this field shared common problems with tree analysis: comparison of graphs relying on distances, clustering or segmentation of structured and dependent data (either at tree scale with such distances or at vertex scale with generative models), change-point detection, classification, pattern identification (either rare or frequent patterns), and assessing the effect of covariates. Each problem was addressed using specific methods, respectively efficient computation of edit distances, spectral clustering, mixture and hidden Markov models, greedy heuristics, neural networks, variable-order Markov models and Generalized Linear Mixed Models (GLMMs).

A first family of HMMs on trees, referred to as hidden Makov trees (HMTs), was studied from the viewpoint of estimation and model selection (Crouse *et al.*, 1998; Durand *et al.*, 2004 and 2005). One shortcoming of the model was that children states were assumed as independent given their parent state. This prevented, in particular, potential order of children to be taken into account by the model. This is why new models were introduced to enable modellers to test and compare different assumptions on the order of children (total, partial or no order). These models also offered new possibilities for specifying various kinds of dependencies, in coherence with orderings. In the case of unordered children, we showed an equivalence between invariance by permutation of ordered models and modelling multivariate count distributions. At the core of such models was the specification of parametric families of distributions for multivariate count data, representing the number of children vertices in each state given their parent state. For the sake of parsimony and versatility, conditional independence relationships between count vector components had to be identified. Statistical estimation of graphs and parameters, as well as model selection, were at the core of P. Fernique's PhD, co-supervised with Y. Guédon. In this work, three kinds of conditional independence assumptions were considered, relying on graphical Markov properties: directed, undirected and partially directed. We also addressed alternative local dependency assumptions between children vertices in Markov trees, depending on arcs being directed towards the root vertex, with some marginal independence assumptions between states (Markov in-trees), or being directed towards leaves, in absence of marginal independence properties (Markov out-trees).

HMMs proved useful for clustering (or segmentation of) vertices within a tree; however, they were not directly usable for clustering whole trees, considered as statistical units. To solve this problem, approaches based on distance matrices between trees were more adequate in the case where the variability to be accounted for was not only in external vertex features, but also in tree topology. Spectral clustering was a possibility, which however raised the issue of selecting the number of clusters. Indeed, criteria based on Euclidean distances in the space of spectral representation would not take into account the true topology of tree spaces. We thus developed clustering methods based on cluster representatives. As in iterative K-means or centroid-based clustering algorithms (Taillard, 2003), the principle was to represent a cluster by the closest tree, on average, to the trees in that cluster. In K-means, the closest point for squared Euclidean distances has a closed-form expression while in centroid-based clustering, the representative has to be chosen within the sample. However in the case of trees, there was no

known efficient algorithm to compute the closest tree representative, which has to be searched among every tree in the space. Not only are its height and size not bounded by those of the sample but even more limiting, each distance computation between a candidate representative and the sample points has time-complexity more than the product of tree sizes times the sum of their maximal degrees. It was conjectured that finding the optimal solution was an NP-Hard problem, so we focused on restricting the search space while keeping it dense, in some sense, using the notion of self-nested tree (Azaïs *et al.*, 2018).

In applications where vertex-scale external features are available, information brought by the structure alone is generally not sufficient. Incorporating external features directly, using distances in some feature space, raises the question of variable normalization with respect to the cost of elementary edit operations. As an alternative, HMMs can be used as denoising tools to summarize features using a small number of states. These may be ordered in our applications, which facilitates the definition of distances. However, replacing features by hidden states requires some high level of confidence in their restoration. More generally, assessing uncertainty in the state process is some useful diagnostic tool, which helps modellers to understand how models affect states to vertices and eventually, what interpretation can be associated to states. The first methodological contribution we proposed to hidden Markov chain (HMC) and HMT analysis was a complementary set of diagnostic tools to assess uncertainty on hidden state processes. Firstly, algorithms were introduced to compute joint state entropy given observations, while providing its additive decomposition along the structure. This aimed at quantifying the local contribution of data to global uncertainty. Secondly, restoration algorithms for state process exploration were extended from HMC to HMT models. These mainly aimed at providing alternative restorations to the Maximum A Posteriori (MAP) yielded by the Viterbi algorithm (Durand and Guédon, 2016).

In the domain of sequential decision, indirectly observed Markovian processes were introduced about the same period as HMCs. These were referred to as Partially Observed Markov Decision Processes (POMDPs); see Åström (1965). We applied HMCs for sequential decision in the problem of optimal timeout identification. This arises in situations where a device has sleep modes with low levels of consumption and a service mode with a high level of consumption. Using the device turns it into service mode, which induces some additional consumption. Determining whether it should be put to one of the sleep modes requires some modelling of future requests. In collaboration with S. Girard and L. Donini, my first co-supervised PhD student, we proposed a solution under the simple assumption of renewal processes. We then extended it to the case of HMCs (Durand *et al.*, 2013a). Eventually some perspectives were explored, mainly as extensions to timeouts for multiple printers with possible job redirections.

Our latest contribution to HMC modelling is related to B. Olivier's PhD thesis. It focuses on (semi-)Markov chains indirectly observed through multiple heterogeneous, asynchronous channels. More specifically, each channel has its own observation time step, and they may have random delays regarding regime switching as represented by underlying Markov chains. This led to original models for coupled hidden semi-Markov chains. Their main application was joint analysis of eye movements and electroencephalograms (EEGs); see Olivier *et al.* (2017).

A particular focus was given to various problems in plant structure modelling, involving different scales (from cells and tissues to orchads) and structures (sequences, trees and other

graphs). Among the most significant contributions was the definition of a global framework to model plant architecture at tree scale, accounting for ontogenetic, genetic, environmental and individual effects and their interactions, with some applications to quantification and prediction of flowering regularity, patchiness and resistance to water stress. We also contributed to research projects motivated by more cognitive and fundamental goals, on various species (*Symphonia globulifera*, *Acacia mangium*, *Arabidopsis thaliana*, rose, beech, mango and apple trees, Aleppo pines).

This manuscript is structured as follows: in Section 2, our methodological contributions to tree-structured data analysis are detailed, focusing on both probabilistic and combinatorial approaches. In Section 3, some more detailed focus is given to multivariate count modelling, based on probabilistic graphical models. In Section 4, our three main contributions to HMC modelling are depicted. Each of these three sections is organized according to the same canvas: generic statement of the problem, our contributions and perspectives. In Section 5, an overview of different approaches applied to several plant growth modelling problems is exposed. Section 6 is dedicated to the presentation of software associated with HMC and HMT analysis. Our conclusions and general perspectives are proposed in Section 7. The appendix contains a resume and list of publications.

2. Tree analysis

Our work related to tree analysis was motivated by the development of a series of models and approaches to address the following problems in tree-structured data: comparison of tree graphs, computation of distances, clustering or segmentation (either at tree scale with such distances or at vertex scale with generative models), change-point detection, classification, pattern identification (either rare or frequent patterns) and assessing the effect of covariates on tree shapes. This motivation came from the need of practitioners issued from different fields of application: signal processing, document categorisation, neuroanatomy, genomics, botany and agronomy. We believed that most methods developed in the context of genomic sequence analysis could be transposed to trees. These were all generic data analysis problems, in the sense where they could be formulated into an abstract, mathematical way that made their solutions more easily transposable from one field of application to another. We could not address all these problems, but we focused on most of them: segmentation, distance computation, clustering and generative modelling.

In what follows, it is assumed that observations are rooted trees $\mathcal{T}$, which are directed graphs $(\mathcal{V}(\mathcal{T}), \mathcal{A}(\mathcal{T}))$, where $\mathcal{V}(\mathcal{T})$ refers to the set of vertices and $\mathcal{A}(\mathcal{T})$ to the set of arcs. We simply notations as $(\mathcal{V}, \mathcal{A})$ when the reference tree $\mathcal{T}$ is clearly defined by the context. Let also $r(\mathcal{T}) = r$ denote the root vertex. Depending on the context:

- External variables X_v may be observed for each $v \in \mathcal{V}(\mathcal{T})$. In this case, $(X_v)_{v \in \mathcal{A}}$ is denoted by $\boldsymbol{X}$.
- $\mathcal{T}$ can be seen as an increasing sequence of trees with depths $n \in \{0, 1, \ldots\}$ (tree process).
- If no external variable is observed, tree shapes can be modelled by implicitly considering X_v as the number of children of v.
- The set of vertex children of every parent may be ordered, unordered or partially ordered.

2.1. Hidden Markov tree models

Generic statement of the problem

It is assumed here that $(X_v)_{v \in \mathcal{A}}$ are dependent random variables that have the same distributions within unknown random zones (which are connected components of $\mathcal{T}$), and different distributions from a zone to another contiguous zone. Equivalently, there exists some underlying state variables $\boldsymbol{S} = (S_v)_{v \in \mathcal{A}}$ such that given $\boldsymbol{S} = \boldsymbol{s}$, $[s_u = s_v \Rightarrow p(X_u | S_u = s_u) = p(X_v | S_v = s_v)]$, where p denotes a probability distribution if X_v is discrete and a probability density function (pdf) is X_v is discrete. The problem is twofold: infer the value of $\boldsymbol{S}$ (tree segmentation) and the conditional distributions $p(X_v | S_v = k)$ for every possible value of k. Tree segmentation yields some clustering at vertex scale that takes into account dependencies between the vertices of $\boldsymbol{X}$.

Let $\mathrm{pa}(v)$ denote the parent of v. The family of HMT models introduced by Crouse *et al.* (1998) makes the additional following assumptions:

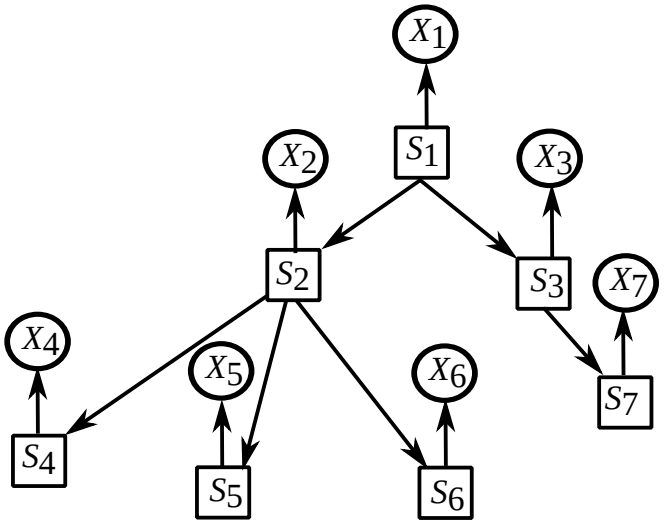

Figure 1: Independence graph of HMT models with conditionally independent children. Hidden categorical variables are in squares. Observed variables of arbitrary nature are in circles.

- $\mathcal{T}$ is deterministic.
- The set of possible values for each S_v is finite and denoted here by $\{1, \ldots, K\}$.
- Variables $(X_v)_{v \in \mathcal{A}}$ are independent given $\boldsymbol{S}$.
- Let $\mathrm{ch}(v)$ denote the set of children of v and $\boldsymbol{X}_{\mathrm{ch}(v)}$ denote $(X_u)_{u \in \mathrm{ch}(v)}$. Then $\boldsymbol{X}_{\mathrm{ch}(v)}$ are independent random variables given S_v.
- $\boldsymbol{S}$ is an homogeneous, first-order Markov tree on $\mathcal{T}$, meaning that for every vertex v, S_v is independent from the other non-descendent state vertices given $S_{\mathrm{pa}(v)}$ and that $P(S_v|S_{\mathrm{pa}(v)})$ does not depend on v.

As consequences from the assumptions above, firstly $(\boldsymbol{S}, \boldsymbol{X})$ is a directed probabilistic graphical model in the sense of Koller and Friedman (2009), whose joint distributions factorizes as

$$p(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{X} = \boldsymbol{x}) = p(S_{r(\mathcal{T})} = s_{r(\mathcal{T})}) \prod_{v \neq r(\mathcal{T})} p(S_v = s_v | S_{\mathrm{pa}(v)} = s_{\mathrm{pa}(v)}) \prod_{v \in \mathcal{V}} p(X_v = x_v | S_v = s_v).$$

Its independence graph is depicted in Figure 1.

Secondly, $p(\boldsymbol{X})$ is invariant by permutations of the children of every vertex v, so this model implicitly assumes unordered children. Note that this distribution writes

$$p(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{s}} \left[p(S_{r(\mathcal{T})} = s_{r(\mathcal{T})}) \prod_{v \neq r(\mathcal{T})} p(S_v = s_v | S_{\mathrm{pa}(v)} = s_{\mathrm{pa}(v)}) \prod_{v \in \mathcal{V}} p(X_v = x_v | S_v = s_v) \right].$$

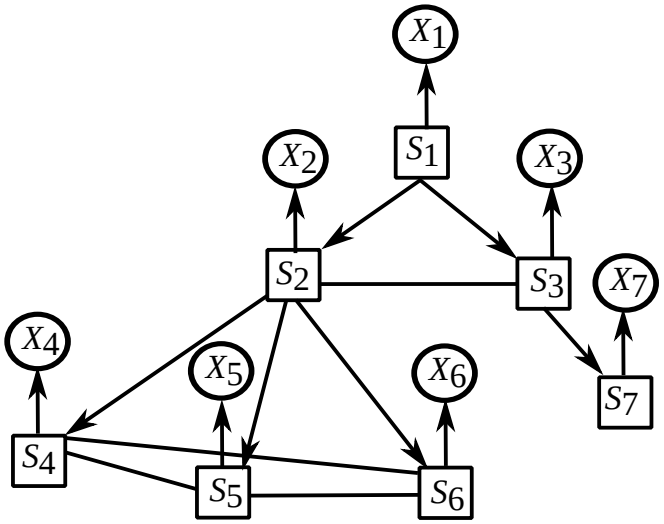

Figure 2: Independence graph of HMT models with conditionally dependent children. Hidden categorical variables are in squares. Observed variables of arbitrary nature are in circles.

Assuming that $p(X_v = x_v|S_v = k)$ does not depend on s and belongs to a parametric family $(p_\theta)_{\theta\in\Theta}$, the model has canonical parametrization $(\pi, A, \theta_1, \ldots, \theta_K)$ where $\pi_k = p(S_{r(\mathcal{T})} = k)$, $A_{jk} = p(S_v = k|S_{\mathrm{pa}(v)} = j)$ and $p_{\theta_k}(x) = p(X_v = x|S_v = k)$.

All inference methods can easily be extended to forests (i.e., sets of trees), assuming that their trees are independent replications of the same HMT model.

Contributions

Maximum likelihood parameter estimation can be addressed with the EM algorithm of Crouse *et al.* (1998). Their E step relies on two recursions: an upward recursion that computes state probabilities given increasing observed subtrees, starting from tree leaves and a downward recursion that computes state probabilities given all observations $\boldsymbol{x}$, starting from tree root. Their algorithm was subject to numeric instabilities since they relied on joint probabilities. Its complexity was in $\mathcal{O}(nK^2)$ with respect to tree size $n = \mathrm{card}(\mathcal{V})$. In Durand *et al.* (2004), we proposed a smoothing algorithm solving this issue, with an additional downward recursion with complexity still in $\mathcal{O}(nK^2)$. We also solved the restoration problem with a Viterbi algorithm, which computes

$$\arg\max_{\boldsymbol{s}} p(\boldsymbol{S} = \boldsymbol{s}|\boldsymbol{X} = \boldsymbol{x}).$$

Other algorithms dedicated to state inference and developed in this context are presented in Section 4.1.

We then relaxed the assumption of conditionally independent children states given parent state. This yielded two families of HMT models described in Durand *et al.* (2005): Markov in-trees and out-trees. Markov in-trees are probabilistic graphical models directed from leaf

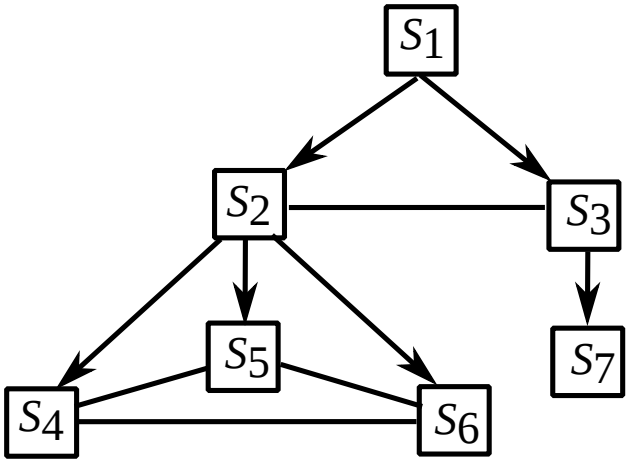

Figure 3: Independence graph of Markov out-tree models.

vertices towards root, while Markov out-trees are directed from root vertex towards leaves (not requiring conditional independence assumptions on children states). These models also have associated Viterbi and EM algorithms with explicit E steps. Their time-complexity is polynomial with respect to tree size and exponential with respect to the maximal number of children. The upward and downward recursions are given in P. Fernique's PhD thesis (2004a).

Let $\mathcal{L}(\mathcal{T})$ denote the set of leaves of $\mathcal{T}$. Markov out-trees are defined by the following factorization of $p(\boldsymbol{S})$:

$$p(\boldsymbol{S} = \boldsymbol{s}) = p(S_{r(\mathcal{T})} = s_{r(\mathcal{T})}) \prod_{v \notin \mathcal{L}(\mathcal{T})} p(\boldsymbol{S}_{\mathrm{ch}(v)} = \boldsymbol{S}_{\mathrm{ch}(v)} | S_v = s_v)$$

while Markov in-trees are defined by factorization

$$p(\boldsymbol{S} = \boldsymbol{s}) = \prod_{v \in \mathcal{L}(\mathcal{T})} p(S_v = s_v) \prod_{v \notin \mathcal{L}(\mathcal{T})} p(S_v = s_v | \boldsymbol{S}_{\mathrm{ch}(v)} = \boldsymbol{S}_{\mathrm{ch}(v)}).$$

Markov out-tree models are partially directed probabilistic graphical models, to be understood both in the sense of AMP and LWF properties (Andersson *et al.*, 2001); a more formal description is provided in Section 3. The independence graph of $\boldsymbol{S}$ is depicted in Figure 3. Random variables in $\boldsymbol{X}$ are assumed to be independent given $\boldsymbol{S}$, as in every other model considered in this section.

Markov in-tree models are directed probabilistic graphical models. The independence graph of $\boldsymbol{S}$ is depicted in Figure 4. It specificity is marginal independence properties of leaf state vertices. More generally, any set of disjoint subtrees starting from the leaves are independent.

Whereas each of these probabilities can be associated to a model parameter in the case where both K and the maximal number of children are small, in the other cases building parsimonious models requires using regression or parametric multivariate distributions as building blocks of

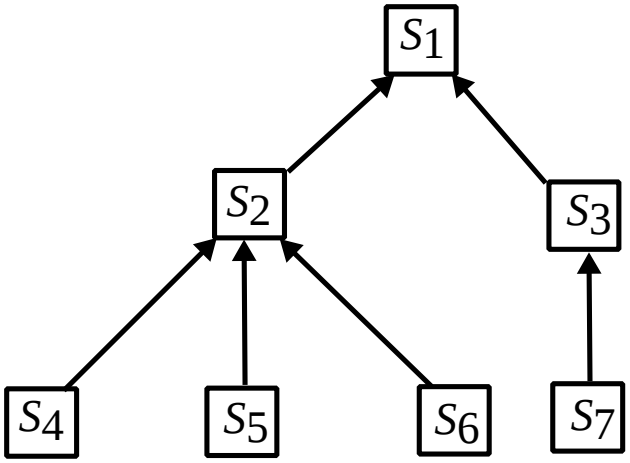

Figure 4: Independence graph of Markov in-tree models.

the model. To handle this case and to account for three possible assumptions on children ordering, the following models were proposed and implemented. Associated hidden models can be defined straightforwardly, since conditional independence assumption of $\boldsymbol{X}$ given $\boldsymbol{S}$ implies that

$$p(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{S} = \boldsymbol{s}) \prod_{v \in \mathcal{V}} p(X_v = x_v | S_v = s_v).$$

Markov out-trees: unordered case. In the case of unordered children, we showed equivalence between an assumption of invariance under any permutation of children in $p(\boldsymbol{S}_{\mathrm{ch}(v)} = \boldsymbol{S}_{\mathrm{ch}(v)} | S_v = s_v)$ and modelling that distribution using multivariate count distributions $p(\boldsymbol{N}_v = \boldsymbol{n}_v | S_v = s_v)$, denoting $\boldsymbol{N}_v = (N_v^{(1)}, \ldots, N_v^{(K)})$ and

$$N_v^{(k)} = \sum_{u \in \mathrm{ch}(v)} \mathbb{I}_{\{S_u = k\}}$$

the number of children of v in state k. If parametric families of distributions $p(\boldsymbol{N}_v | S_v = k)$ are chosen separately for each value of k, then the problem is equivalent to defining multivariate count distributions, which is addressed in more details in Section 3.

Markov out-trees: ordered case. In the case of ordered children, for any vertex v, let n_v denote its number of children and $(S_{v,1}, \ldots, S_{v,n_v})$ its children states. Parsimonious modelling of $p(\boldsymbol{S}_{\mathrm{ch}(v)} = \boldsymbol{S}_{\mathrm{ch}(v)} | S_v = k)$ relies on an assumption of compatible families of distributions, meaning that if $n_u = c$ and $n_v = c + 1$, then

$$p(S_{u,1} = s_{u,1}, \ldots, S_{u,c} = s_{u,c} | S_{\mathrm{pa}(u)} = j) = \sum_k p(S_{v,1} = s_{v,1}, \ldots, S_{v,c} = s_{v,c}, s_{v,c+1} = k | S_{\mathrm{pa}(v)} = j).$$

Without loss of generality, $p(S_{u,1} = s_{u,1}, \ldots, S_{u,c} = s_{u,c} | S_{\mathrm{pa}(u)} = j)$ can be assumed a higher-order Markov chain, taking as order the maximal number of children in $\mathcal{T}$ minus one. In practice, models of reasonable parsimony can be obtained using variable-order Markov models, whose memory trees are selected by information criteria (see Csiszár and Talata, 2006).

Markov out-trees: partially ordered case. Our attention focused to the case where a fixed number o of children of v are totally ordered: denoting by (v, w) the w-th child of v, then $(v, 1) > \ldots > (v, o)$ and the other children are mutually unordered, but they all are less than (v, o). This framework is mainly motivated by plant growth modelling, see Section 5. In this case, both previous strategies (higher-order Markov chains and multivariate count models) can be combined and yield:

$$\begin{aligned} p(\boldsymbol{S}_{\mathrm{ch}(v)} = \boldsymbol{S}_{\mathrm{ch}(v)} | S_v = k) = \\ p(N_{v,o}^{(1)} = n_{v,o}^{(1)}, \ldots, N_{v,o}^{(K)} = n_{v,o}^{(K)} | S_v = k, S_{v,1} = s_{v,1}, \ldots, S_{v,o} = s_{v,o}) \\ \times p(S_{v,1} = s_{v,1}, \ldots, S_{v,o} = s_{v,o} | S_v = k), \end{aligned}$$

where

$$N_{v,\ell}^{(j)} = \sum_{\substack{u \in \mathrm{ch}(v) \\ u < o}} \mathbb{1}_{\{S_u = j\}}.$$

Given parent state, ordered children states are then modelled by variable-order Markov models (corresponding to $p(S_{v,1}, \ldots, S_{v,o} | S_v)$) and unordered children states are modelled with multivariate count distributions given ordered children states (corresponding to $p(N_{v,o}^{(1)}, \ldots, N_{v,o}^{(K)} | S_v, S_{v,1}, \ldots, S_{v,o})$).

Markov in-trees: unordered case. In the case of unordered children, modelling $p(S_v | \boldsymbol{S}_{\mathrm{ch}(v)})$ is equivalent to modelling $p(S_v | \boldsymbol{N}_v)$. Parsimonious models can be obtained by using multiple regressions for discrete random variables based on GLMs (e.g., Poisson or their extensions to zero-inflated or overdispersed data).

Markov in-trees: ordered and partially ordered cases. In the case of ordered (resp. partially ordered) children, modelling $p(S_v | S_{v,1}, \ldots, S_{v,n_v})$ (resp. $p(S_v | S_{v,1}, \ldots, S_{v,o}, N_{v,o}^{(1)}, \ldots, N_{v,o}^{(K)})$) could be turned more difficult than in previous cases by the fact that the number of children may be variable in practice, although we did not encounter this case in applications. This would correspond to regression models with variable numbers of covariates.

Extension to random structures. If variability in the structure itself has to be accounted for, the numbers of children $(N_v)_{v \in \mathcal{V}}$ now become random variables, and trees can be seen as processes where at each time step, children are added randomly to the leaves of the tree at current step. Under Markovian-like assumptions (given the first a ancestor states of u, N_u is independent from every non-descendent (S_v, N_v)), the trees are those underlying branching

processes. This family of models is referred to as *generative Markov trees*. Their sample space are whole tree graphs with associated states, whereas multitype branching processes are sequences of multivariate counts that consider the total number of leaves within each state (because the precise vertex genealogy is unknown).

Now, if the states are observed, the joint probability writes

$$p(\boldsymbol{S}=\boldsymbol{s},\boldsymbol{N}=\boldsymbol{n}) = p(S_{r(\mathcal{T})}=s_{r(\mathcal{T})}) \\ \times p(N_{r(\mathcal{T})}=n_{r(\mathcal{T})}|S_{r(\mathcal{T})}=s_{r(\mathcal{T})}) \prod_{v\notin\mathcal{L}(\mathcal{T})} p(N_v=n_v,\boldsymbol{S}_{\mathrm{ch}(v)}=\boldsymbol{S}_{\mathrm{ch}(v)}|S_v=s_v)$$

in the case of generative Markov out-trees. Without loss of generality, $p(N_v=n_v,\boldsymbol{S}_{\mathrm{ch}(v)}=\boldsymbol{S}_{\mathrm{ch}(v)}|S_v=s_v)$ factorizes as $p(\boldsymbol{S}_{\mathrm{ch}(v)}=\boldsymbol{S}_{\mathrm{ch}(v)}|S_v=s_v,N_v=n_v)p(N_v=n_v|S_v=s_v)$ and previously-described models can be used for $p(\boldsymbol{S}_{\mathrm{ch}(v)}=\boldsymbol{S}_{\mathrm{ch}(v)}|S_v=s_v,N_v=n_v)$.

Particular attention must be given to censoring of observed leaf vertices, distinguishing cases where they were actually extinguished during the process from cases where measurements ceased, which induces censoring.

Extensions of this model to generative hidden Markov-out trees are straightforward. Note that as a particular case of these models is the one defined by the marginal $p(\boldsymbol{N}=\boldsymbol{n})$ in (1). This corresponds to generative Markov-out trees with unobserved states (only the total number of children of each vertex is observed) and can be handled with our EM algorithm, provided the model is identifiable on $\{p_{\theta_k}(N_v=n_v|S_v=k)\}_{1\leq k\leq K}$.

Generative Markov in-trees are related to coalescent processes (Lambert and Popovic, 2013) and not further considered in this manuscript.

Perspectives in model selection

In practical applications, tree-structured data generally come with ordered children, since this is more convenient for data coding and storage. However, it is sometimes known, or only conjectured, that order is not relevant regarding dependencies in statistical models. As perspectives, consistent methods for validating this assumptions should be developed in frameworks such as generative Markov tree models, including the hidden case. These could rely on either cross-validation using random permutations of children, or information criteria such as BIC (Kass and Wasserman, 1995).

More generally, model selection issues arise not only to choose or validate models among families with different ordering assumptions, but also within each of these families: choice of the number of hidden states, of some particular parametric assumption on the transition probabilities from parent to children (Markov out-trees) or from children to parent (Markov in-trees), choice of the order of the model in the case of higher- or variable-order models. Such issues have been addressed in Markovian models for sequences, but not for trees.

2.2. Edit distances

Defining distances on trees offers possibilities of applying distance-based clustering of whole trees, or clustering of subtrees of a given single tree. HMT-based approaches perform clustering at vertex scale by considering changes in the (conditional) distribution of local features.

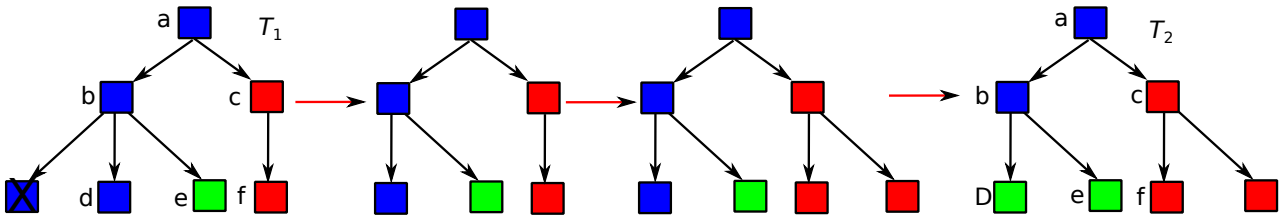

Figure 5: Sequence of optimal edit operation and mapping between two trees T_1 and T_2. It is assumed when editing the children of vertex b that changing a vertex label has a lower cost than deleting this vertex and adding a new one with the correct label, since the distance has to satisfy the triangle inequality. The mapping between T_1 and T_2 is represented by pairs of vertices with the same letter (one being capitalized if some label change occurred). Crossed vertices have no image by the mapping.

Extending these models to clustering of whole trees would require using mixtures of HMT models. These would be computationally challenging to estimate, would involve a large numbers of parameters and large quantities of data for estimation, but above all, clustering would be based on changes in the dynamics of branching processes instead on sheer tree shapes.

As an alternative, we used edit distances between trees, firstly to quantify tree asynchronism. Tree edit distance are based on a set of edit operations of tree graphs associated with costs (mainly: inserting and deleting vertices, or changing their features). The distance between trees is the minimal cost among all sequences of edit operations transforming one tree into the other (Zhang, 1995). Computing the optimal sequence also provides a mapping between both trees, distinguishing between vertices that have some homologous vertex in the other tree from the inserted or deleted vertices. This is illustrated in Figure 5. The rate of vertices mapped with exactly the same homologous state is some quantification of tree synchronism. This reveals especially useful in combination with HMT models when features are continuous multivariate, since segmentation acts as both normalization and denoising step. An illustration is presented in Subsection 5.2.1.

Then, we investigated edit distances for distance-based clustering, combining spectral and centroid-based approaches using self-nestedness properties of trees.

2.2.1. Modelling approximate tree self-nestedness

Generic statement of the problem. Modelling self-nestedness for unordered trees was addressed by Godin and Ferraro (2010) in the framework of lossy data compression. Unordered trees $\mathcal{T}$ can be represented by edge-labelled DAGs $\mathcal{R}(\mathcal{T})$ whose vertices are equivalence classes of their complete subtrees (i.e., subtrees with every descendant of their roots), and where the labels are the numbers of occurrences of each subtree. Self-nested trees $\mathcal{T}$ can be defined by either of the four properties:

- $\mathcal{T}$ is a single leaf or all the subtrees of $\mathcal{T}$ rooted at the children of the root vertex are self-nested and one of them contains the others as subtrees;
- all the complete subtrees of $\mathcal{T}$ with identical height are isomorphic;
- any two complete subtrees of $\mathcal{T}$ are either isomorphic or such that one is a subtree of the other;
- $\mathcal{R}(\mathcal{T})$ is a linear DAG (*i.e.*, it contains at least one path that goes through all its vertices).

As a consequence from the last characterisation, the DAG representation of self-nested trees is sparse.

Self-nestedness is a concept justified, in the context of plant structure modelling, by the fact that plants are build from elementary processes repeating themselves (see Section 5). However, repetitions are not perfect and are randomly perturbed by numerous physiological and environmental factors. Moreover, adding or deleting one vertex to a self-nested tree generally turns its DAG as non-linear. As a consequence, large trees obtained by natural processes are most often far from being self-nested. In such cases, lossy compression could be achieved, in principle, by approximating any tree by its closest self-nested tree, but computing this approximation is conjectured to be some NP-hard problem.

Contributions. Our approach consisted in replacing the fixed edge labels by independent random variables, which led to stochastic trees that are special cases of generative Markov out-trees with bounded depth (states of extinction). In collaboration with R. Azaïs and C. Godin, we transposed the first three characterisations of self-nested trees to the case of stochastic tree processes. These involve the definition of subtree distributions given their context (i.e., the rest of the tree), and rely on invariance of these distributions over the tree (instead of isomorphism). We showed that stochastic processes with linear DAGs satisfy those three characterisation (but found no reciprocal property). We also showed that first-order Markov out-trees are the only tree process satisfying independence of subtree distributions with respect to their context.

These results were completed by a combinatorial and topological study of the space of self-nested trees. This allowed in particular the control of distances from arbitrary to self-nested trees and thus guarantees quality of approximations. Then we provided algorithms to perform accelerated computations on DAGs and on self-nested trees, particularly computation of edit distances. Eventually, we developed several heuristics for computing the nearest self-nested tree, either using simulated annealing (post-doctoral work of R. Azaïs) or by recursively replacing subtrees by their centroid among self-nested trees, starting from the leaves (see Azaïs *et al.*, 2018). Another improvement was proposed, which consists in locally pruning some of the centroids that replace subtrees. We provided errors bounds and compared errors with the reference algorithm by Godin and Ferraro (2010) on simulated data.

Perspectives.

Optimal lossy compression and problem complexity.

The main issues remaining to be addressed in lossy compression of unordered trees are, on the

one hand, the study of optimality of our algorithms, in the case were the distance between trees and their lossy representations are constrained not to exceed given bounds. Moreover, these approach is still unrelated to information theory (see next paragraph) and the optimal representation of the DAGs themselves was not addressed yet. On the other hand, the problem on determining whether complexity of finding nearest self-nested trees is in P or NP is currently still under investigation by the Mosaic team.

Lossless optimal compression of unordered trees.
Our work mainly focused on lossy compression, but optimal lossless compression is still an open question. Some algorithms have been proposed for ordered tree-structured data (Chen and Reif, 1996; Itokawa *et al.*, 2009). However, one the one hand, the desired properties of tree compression algorithms have been poorly formalized. In the other hand, these algorithms are dedicated to ordered trees and obviously embed more information than necessary for the compression of unordered trees. Moreover, Chen and Reif (1996) claim that their algorithm belongs to the family of Lempel-Ziv compression algorithms. However the dictionary built while traversing the tree must be transmitted, which contradicts the Lempel-Ziv principle of online reconstruction of the dictionary while uncompressing the tree (Ziv and Lempel, 1978).

A lossless compression algorithm was proposed by Choi and Szpankowski (2012) to compress Erdös-Rényi graphs up to graph isomorphic mappings, but such distributional assumption is too restrictive to include tree graphs issued from real-data applications.

Thus, the issues of computing entropy bounds for Markov out-trees, designing associated optimal arithmetic and Lempel-Ziv encoders remain to be addressed. Building Lempel-Ziv encoders would offer new perspectives for quantifying the algorithmic complexity of tree families, induced semi-distances and associated clustering methods, following the framework introduced by Revolle *et al.* (2017).

2.2.2. Tree clustering

Generic statement of the problem

Subsection 2.1 addresses clustering at vertex scale for tree-structured data. Some applications rather require clustering at tree scale, i.e. considering trees as statistical units. We investigated this problem from the angle of spectral clustering.

An introduction to spectral clustering is provided by von Luxburg (2007); in summary, its principle is to build some similarity matrix between trees from the distance matrix (using edit distances, as introduced at the beginning of this section). The similarity between pairs of trees is seen as edge weights in an undirected graph whose vertices are the trees. Clusters are obtained by obtaining so-called *graph cuts* (equivalently connected components) that minimise the sum of inter-component weights. This is achieved by computing the normalized Laplacian of the similarity matrix and computing its eigen elements. This yields an Euclidean representation of the points to be clustered, which can be used to performed clustering, now in some Euclidean space (referred to as *space of spectral representation*) using standard approaches (mixture models for example).

Contributions

As in most clustering methods, the issue of selecting the number of clusters requires some particular treatment. Criteria based on Euclidean distances in the space of spectral representation would not take into account the true topology of tree spaces. We thus developed clustering methods based on cluster representatives. As in iterative K-means or centroid-based clustering algorithms (Taillard, 2003), the principle was to represent a cluster by the closest tree, on average, to the trees in that cluster. In K-means, the closest point for squared Euclidean distances has a closed-form expression while in centroid-based clustering, the representative has to be chosen within the sample. However in the case of trees, there was no known efficient algorithm to compute the closest tree representative, which has to be searched among every tree in the space, as mentioned in Subsection 2.2.1. Not only are its height and size not bounded by those of the sample but even more limiting, each distance computation between a candidate representative and the sample points has time-complexity more than the product of tree sizes times the sum of their maximal degrees. Since finding the exact solution of the minimisation problem seemed out of reach, we focused on restricting the search space while keeping it dense, in some sense, using self-nested trees as cluster representatives in the selection step.

We used prediction strength (Tibshirani and Walther, 2005) to select the number of clusters, since this method is very general and independent of the specific algorithm used for clustering. It only requires some clustering of any cloud of points, and the ability to use that clustering to predict the cluster of out-of-sample points. It is based on the stability of clusters under cross-validation procedures.

Our first results on simulated trees showed that the approach is robust to high mean intra-cluster over inter-cluster distances.

Perspectives

Our approach is quite general and could be applied to other type of data: sequences or more general types of graph. Its main rationale is to use spectral clustering (thus, Euclidean distances) to actually achieve the clustering, but to use distances in the original data space to perform model selection. The main difficulty is to define an appropriate representation of the cluster. Typically, one point of the original space (which may not be in the sample) is chosen. Finding the point minimising the mean within-cluster distance is the main issue. Then the prediction function required by prediction strength is the closest representative to test points.

Moreover in numerous applications, the question of interest is multiscale clustering of trees. In a given tree, we want to seek subtrees that share structural similarities at several nested scales, so that clusters at vertex scale could be further clustered at a some coarser, unknown scale, and so on until we reach whole tree scale. An example of application is provided in Subsection 5.2.5. A possible approach to address the problem are variational graph autoencoders (Kipf and Welling, 2016), since the principle of autoencoders is to encode structures at nested scales through dimension reduction by reducing the number of neurons per layer in multilayer neural networks. Another approach would be to duplicate the tree graph several times and connect tree layers by edges in a Bayesian model with priors on edges, and perform model selection by learning the edge posteriors with MCMC or variational approximations.

2.3. Multiple change-point detection

Generic statement of the problem

Subsection 2.1 addresses segmentation at vertex scale for tree-structured data using hidden Markov models. This approach is especially useful whenever vertex features may have the same distribution within separated connected components, if separation is due to other connected components within which vertex features have different distributions. In other words, HMMs are motivated by the possibility to return to previously-visited states and to predict future observations.

If such features are not required and if it is sufficient to find a partition of the tree, such that within each connected component, vertices have the same distribution, then multiple change-point detection approaches may be sufficient. They do not rely on dependencies assumptions between variables in $\boldsymbol{X}$. On the contrary, they assume that all variables are independent. These are identically distributed if and only if they belong to the same component.

In the case of sequential data of length n, exact algorithms in $\mathcal{O}(n^2)$ exist for likelihood maximisation. Moreover, consistency of penalized contrast functions to estimate the number of change points have been established (Lavielle, 2005). However, the problem of multiple change-point detection in tree-structured data remained unaddressed.

Contributions

Firstly, we addressed change-point detection in trees with a fixed, known number K of change points. Finding the change points is equivalent to finding a partition Π of $\mathcal{V}(\mathcal{T})$. For any $\pi \in \Pi$ and $v \in \pi$, let $p_\pi(X_v)$ denote the distribution of X_v. This distribution is assumed to belong to some parametric family $(p_\theta)_{\theta\in\Theta}$, where the parameter is denoted by θ_π. From the independence assumption, and since by definition all vertices in π have distribution p_π, the log-likelihood function writes

$$\mathcal{L}(\boldsymbol{X}, \Pi, \theta_\Pi) = \sum_{\pi\in\Pi}\sum_{v\in\Pi} \log p_{\theta_\pi}(x_v)$$

where $(\theta_\Pi) = (\theta_\pi)_{\pi\in\Pi}$. There was no known algorithm with quadratic time-complexity to determine

$$\arg\max_{\Pi,\theta_\Pi} \mathcal{L}(\boldsymbol{X}, \Pi, \theta_\Pi). \tag{1}$$

Even more, the only known possibility is to enumerate the $S(n, K+1)$ elements of Π with cardinal number $K+1$ ($S(n,k)$ denoting Stirling numbers of the second kind).

We proposed some greedy, iterative heuristic to maximise the log-likelihood function. Given a current proposal $\tilde{\Pi}$, the principle is to consider in turn for every $\pi \in \tilde{\Pi}$ all partitions of π into two sets (Fernique *et al.*, 2015 and 2016a). The number of change-points K was selected using slope heuristics adapted from the sequential framework (Baudry *et al.*, 2012). The approach was used to detect patches in mango trees and the estimated partitions were compared to a priori known partitions obtained from biological considerations (Fernique *et al.*, 2016a).

Perspectives

Firstly, the exact complexity of the combinatorial problem (1) has to be determined. Secondly, the theoretical properties of slope heuristics have to be investigated under the assumption of tree-structured data, provided we have an oracle exact algorithm to solve (1). Lastly, the greedy algorithm is likely to be improved, maybe by solving a sequence of time-series multiple change-point detection problems. It should also be compared with the approach by Thepaut and Rigaill (2019).

3. Multivariate counts and graphical models

As highlighted in Section 4, generative statistical modelling of trees using Markov out-trees requires statistical models for multivariate counts, which correspond to $p(N_v^{(1)}, \ldots, N_v^{(K)} | S_v = k)$ in this framework (recalling that this denotes the joint distribution of the number of children in each state given their parent state). Models may be chosen in distinct families for each value of k, which makes the problem equivalent to defining discrete multivariate distributions $p(\boldsymbol{N}) = p(N^{(1)}, \ldots, N^{(K)})$. Thus, the problem can be formulated in a general way, independently of using those distributions in Markovian models for trees or in other applications. Indeed, multivariate counts are often encountered in health, biological, medical and social applications – in fact any domain where a categorical (state) variable is recorded in several individuals, whenever the joint state distribution is invariant under permutations of these individuals.

Defining appropriate families of distributions is a crucial point, not only for the sake of accurate prediction, but also because dependencies between states is often of interest for interpretation of interactions between individuals. For example if some inventory is made of the species of living organisms within a given place, $N^{(k)}$ representing the number of individuals with species k, dependencies are likely to be interpreted in terms of competitions, trophic relationships or symbiosis. This section focuses on both dependencies and definition of parametric families of distributions, which are two closely related aspects.

Generic statement of the problem

The aim of this work was: I) to infer conditional independence relationships between the K count variables; II) to build parsimonious parametric models consistent with these relationships; III) to characterise and test the effects of covariates on the distribution of $\boldsymbol{N}$, particularly on the dependencies between its components.

The second objective is motivated by some impossibility or lack of efficiency to define such models by exhaustive enumeration of $p(\boldsymbol{N} = \boldsymbol{n})$, corresponding to saturated models, given that:

(i) The support of some $p(N^{(k)})$ may be unbounded;

(ii) Even if the support of every $p(N^{(k)})$ is upper-bounded by M, then defining $p(\boldsymbol{N})$ without any particular assumptions requires up to $M^K - 1$ independent parameters, which is more that the sample sizes and the values of M and K would allow in a vast majority of applications.

Context (ii) typically corresponds to cases where multivariate histograms have many cells, most of which are empty.

Contributions

To achieve these goals, we proposed an approach based on graphical probabilistic models (Koller and Friedman, 2009) to represent the conditional independence relationships in $\boldsymbol{N}$ and on parametric distributions to ensure model parsimony. Three kinds of graphs are usually considered:

either undirected (UG), directed acyclic (DAG), or partially directed acyclic (PDAG) graphs. Models and methods for graph identification were proposed in UGs and DAGs, but the case of parametric models for PDAGs has been considered less often in the literature. Using undirected models lacks of versatility, since marginal independence relationships cannot be represented if variables are conditionally dependent. If directed models are used, pairwise cyclic independence relationships cannot be represented either. To raise such limitations, we developed mixed approaches based on PDAGs.

The main issue in probabilistic graphical models (PGMs) is graph identification, essentially because it involves combinatorial searches among the set of possible graphs. This can be circumvented using Lasso techniques, but these are limited to multivariate Gaussian distributions or distributions only known up to some intractable scaling factor, which turns model selection within different families especially difficult (Friedman *et al.*, 2008; Yang *et al.*, 2012). Our contexts of application, on the contrary, are rather characterised by zero-inflated, right-skewed marginal distributions.

In the context of P. Fernique's PhD thesis (2014b), we introduced a family of parametric PDAG models, such that covariates can be introduced easily and in a flexible manner. The class of considered PDAGs is such that the joint distribution factorises as

$$P(\boldsymbol{N} = \boldsymbol{n}) = \prod_{c \in \mathcal{C}} P(\boldsymbol{N}_c = \boldsymbol{n}_c | \boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)}), \tag{2}$$

where $\mathcal{C}$ denotes the set of undirected subgraphs (so-called *chain components*) and $\mathrm{pa}(c)$ the parent chain components of c (which can be the empty set).

Each source vertex of the graph is associated with some univariate distribution chosen in parametric families and their mixtures. Each non-singleton source component of the graph is associated with some multivariate distribution chosen among diverse extensions of the multinomial family, the multivariate Poisson distribution (Karlis, 2003) and their mixtures. Each component of the graph with at least one parent is associated with the corresponding families of univariate and multivariate regression models defined hereinbefore in the case of source components. As a consequence, each factor in (2) is modelled by a parametric distribution or a regression model. An example of parametric PDAG is provided in Figure6. The parameters are estimated by maximum likelihood and the family with maximal BIC value is selected for each factor, which in the ends uniquely defines a joint distribution.

Graph search is achieved by a stepwise approach, issued from unification of previous algorithms presented in Koller and Friedman (2009) for DAGs: Hill climbing, greedy search, first ascent and simulated annealing. The search was improved by taking into account the parametric distribution assumptions, which led to caching overlapping graphs at each step. An adaptation to PDAGs of graph search algorithms for DAGs was developed, by defining new operators: edge addition and deletion on the one hand, directed edge reversal at chain component scale (instead of vertex scale) on the other hand. Two operators specific to PDAGs were added: chain component addition and deletion. On the one hand, a parent vertex can be added to its child chain component, or a child vertex can be added to one of its parent chain component, which results into deletion of one chain component in both cases. On the other hand, a vertex

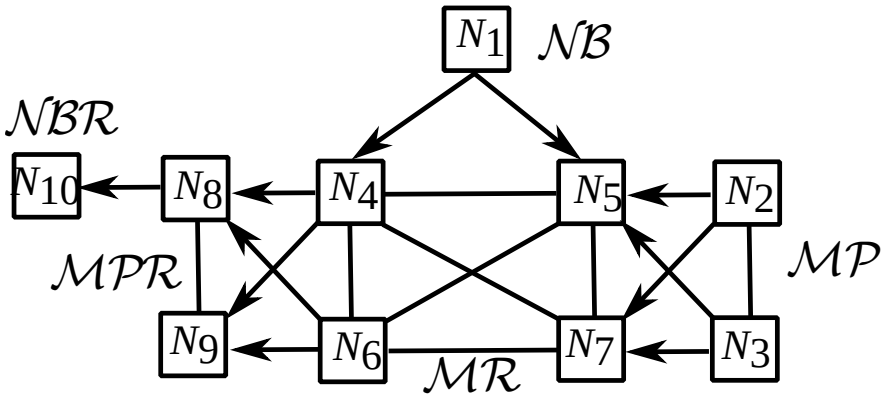

Figure 6: An example of parametric PDAG. The chain components are the graph undirected parts: $\{N_1\}, \{N_2, N_3\}, \{N_4, \ldots, N_7\}, \{N_8, N_9\}$ and $\{N_{10}\}$. Source chain components (here, $\{N_1\}$ and $\{N_2, N_3\}$) are modelled by univariate or multivariate distributions (here a negative binomial $\mathcal{NB}$ and multivariate Poisson $\mathcal{MP}$, resp.). The other chain components are modelled by univariate or multivariate regressions (here a multinomial regression $\mathcal{MR}$, multivariate Poisson regression $\mathcal{MPR}$ and negative binomial regression $\mathcal{NBR}$).

from a chain component c can be set to be a parent or a child of c, which results into addition of one chain component.

Since our model is essentially defined by chaining regression models in PDAGs, some set of covariates $\boldsymbol{Y}$ can be easily incorporated in the model. This is achieved by substituting $P(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{Y} = \boldsymbol{y})$ for $P(\boldsymbol{N} = \boldsymbol{n})$ in (2), and $P(\boldsymbol{N}_c = \boldsymbol{n}_c|\boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)}, \boldsymbol{Y} = \boldsymbol{y})$ for $P(\boldsymbol{N}_c = \boldsymbol{n}_c|\boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)})$. In the graph search step, some covariates in the set $\boldsymbol{Y}$ may be discarded in practice leading to $P(\boldsymbol{N}_c = \boldsymbol{n}_c|\boldsymbol{N}_{\mathrm{pa}(c)} = \boldsymbol{n}_{\mathrm{pa}(c)}, \boldsymbol{Y}_c = \boldsymbol{y}_c)$, in a differentiated way with respect to the different chain components c.

Comparisons between the different proposed algorithms were performed on simulated datasets to: (i) Assess gain in speed induced by caching; (ii) Compare the graphs obtained under parametric and nonparametric distributions assumptions; (iii) Compare different strategies for graph initialisation. Strategies based on several random graphs were compared to those based on a fast estimation of an UG, assumed to be the moral graph and obtained using the approach by Friedman *et al.* (2008).

In some specific applications, mixtures of parametric PDAGs models were found more adequate that PDAGs with mixture of parametric distributions as chain components, see Section 5.

Perspectives

The family of PGMs we introduced was computationally efficient from an estimation point of view and offered some improvements regarding generality, in the sense that previous state-of-the-art graphical models were particular cases of it. However, it failed to represent arbitrary

dependencies. This was due to the difficulty to define tractable parametric families of UGs for multivariate counts with arbitrary graphs. Indeed, we used parametric families as building block of PGMs, which dependency graphs were mostly either entirely connected or disconnected (Fernique *et al.*, 2016b). Thus, our work would highly benefit from relaxing such constraints, using for example axiomatic definitions of the required families, which somehow would be discrete equivalent of Gaussian distributions.

Moreover, consistency of our graph estimation procedure would need to be investigated from a theoretical point of view. Although originally, our modelling framework for PDAGs was motivated to applications in multivariate count modelling, in the end very few aspects are specific to discrete variables and the extension to multivariate Gaussian PDAGs is straightforward. We believe that combining approaches for UGs (Friedman *et al.*, 2008) and DAGs (Chickering, 2002; Verma and Pearl, 1992) are promising leads for this task.

Multiple applications for multivariate count models are also to be expected, particularly in ecology as mentioned in Section 7, but also in the analysis and modelling of human activities. In Durand and Fernique (2013c), we addressed inference of dependencies from the number of activities of various types (walk, school, work, leisure, cycling, driving, bus, etc.) performed jointly by the different members of families within a day, with particular focus on transportation.

4. Hidden Markov chains

Some methodological contributions motivated by sequence analysis with HMCs are developed here. The first one is a general set of diagnostic tools to assess uncertainty on hidden state processes and to provide alternative restorations to the Maximum A Posteriori (MAP) sequence yielded by the Viterbi algorithm. It was extended to non-sequential data (i.e., graphical HMMs). The second contribution, associated with L. Donini's PhD thesis, is related to the sequential decision problem in optimal timeout identification. The third contribution is the topic of B. Olivier's PhD thesis. It aims at modelling coupled heterogeneous sequences with regime switches subject to delays.

4.1. Quantifying uncertainty on state processes

Generic statement of the problem

Using similar notations as in Section 2, we consider some HMC model $\boldsymbol{X} = (X_t)_{t\geq 0}$ with state process $\boldsymbol{S} = (S_t)_{t\geq 0}$ and with given parameter λ (usually estimated from data). Since λ is fixed, it will generally be omitted in probabilistic notations.

The state inference problem consists in estimating values of $\boldsymbol{S}$ from λ and a finite sequence of observations $\boldsymbol{x} = (x_t)_{1\leq t\leq n}$, also denoted by x_1^n. This is particularly crucial in numerous applications where the unobserved states are expected to have some meaningful interpretation. In such cases, the state sequence has to be restored: Firstly, to validate the expected interpretation with respect to observations; secondly, to validate assumptions underlying the model itself and lastly, because restored state values may be required as inputs of further post-processing steps. For example, validation of the choice for the family of emission distributions is generally achieved by visualising and comparing histograms with conditional pdfs given the states. The use of restored states for post-processing is typically required in prediction, segmentation or denoising (Ephraim and Mehrav, 2002).

State restoration is usually achieved by optimizing the MAP criterion,

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} P(\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{X} = \boldsymbol{x}) \tag{3}$$

solved by the Viterbi algorithm.

Using the restored state sequence in the above-mentioned frameworks relies on the assumption that uncertainty on the state process given observations should be reasonably low. Quantification of local state uncertainty given an observed sequence has been addressed by either enumeration of a fixed number L of best solutions to problem (3), or by state profiles, which are state sequences summarised in a $K \times n$ array, K being the number of states (Guédon, 2007). These profiles are obtained by computing

$$\max_{(s_u)_{u\neq t}} P(S_1 = s_1, \ldots, S_{t-1} = s_{t-1}, S_t = j, S_{t+1} = s_{t+1}, \ldots, S_n = s_n | \boldsymbol{X} = \boldsymbol{x}) \tag{4}$$

and drawing curves of such probabilities indexed by t (one curve per value of j). Such methods may highlight potential relevant (i.e., significantly probable) local alternatives to the MAP-optimal $\hat{s}_t$. As a complement, smoothed probabilities $P(S_t = j | \boldsymbol{X} = \boldsymbol{x})$ are sometimes considered.

However, these approaches do not allow for global uncertainty quantification on the whole state sequence. Hernando *et al.* (2005) proposed an algorithm to compute the entropy $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$.

Contributions

The approach by Hernando *et al.* (2005) provides a well-formalized solution to global uncertainty quantification, since this relies on entropy and the information theory. However, it suffers from two shortcomings. Firstly, this is insufficient for detailed state interpretation: knowledge on how global uncertainty is distributed along the sequence is also of primary importance. Secondly, their algorithm is specific to sequences and cannot be generalized to other types of probabilistic graphical models.

We defined a general class of graphical hidden Markov models (GHMMs) and provided a result of additive decomposition of the global entropy with respect to local contributions that applies to this whole class of models (Durand and Guédon, 2016). GHMMs are defined by a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, observed variables $\boldsymbol{X}$ being indexed by $\mathcal{V}$, i.e., $\boldsymbol{X} = (X_v)_{v\in\mathcal{V}}$ and by hidden state variables $\boldsymbol{S}$ also indexed by $\mathcal{V}$, such that $\mathcal{G}$ is a perfect map for $P(\boldsymbol{S})$ and the observed variables are conditionally independent given $\boldsymbol{S}$. This family contains hidden Markov chain (HMC) and tree (HMT) models. Then the decomposition of entropy writes as:

$$H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x}) = \sum_{v\in\mathcal{V}} H(S_v|\boldsymbol{S}_{\mathrm{pa}(v)}, \boldsymbol{X} = \boldsymbol{x}),$$

where $\mathrm{pa}(v)$ denotes the parent of vertex v, for any subset U of V, X_U denotes the family of random variables $(X_u)_{u\in U}$ and by convention, $P(S_v|\boldsymbol{S}_U) = P(S_v)$ if $U = \emptyset$.

Every term of that sum is associated with one vertex in $\mathcal{V}$. Hence, these entropies can be interpreted as local contributions to global uncertainty. Since these profiles are unidimensional, they can be drawn whatever the graphical structure $\mathcal{G}$, contrarily to smoothed probabilities, which are multivariate and thus difficult to visualise on graphs. An illustration of some entropy profile on trees is provided in Figure 7.

We provided algorithms with polynomial complexity in the case of HMC models and HMOT models with independent children to compute the elements of the decomposition. It was shown using synthetic and real-case data that the obtained local entropy profiles are relevant for state uncertainty diagnosis and state interpretation. These algorithms are complementary with approaches that either enumerate the L most likely state restorations or solve problem (4). Algorithms to derive the solutions of these problems for HMOT models were also derived.

It was shown that usual smoothed probability profiles are not relevant for quantifying global state uncertainty, due to their inherent marginalization property.

Perspectives

Non-factorisable models. We obtained explicit results in computing and decomposing entropies $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ for a class of models such that $P(\boldsymbol{S} = \boldsymbol{s}|\boldsymbol{X} = \boldsymbol{x})$ has some factorization property. This is not the case any longer for more complex models, as hidden Markov random fields, in which local contributions may not be identified and $H(\boldsymbol{S}|\boldsymbol{X} = \boldsymbol{x})$ may not be

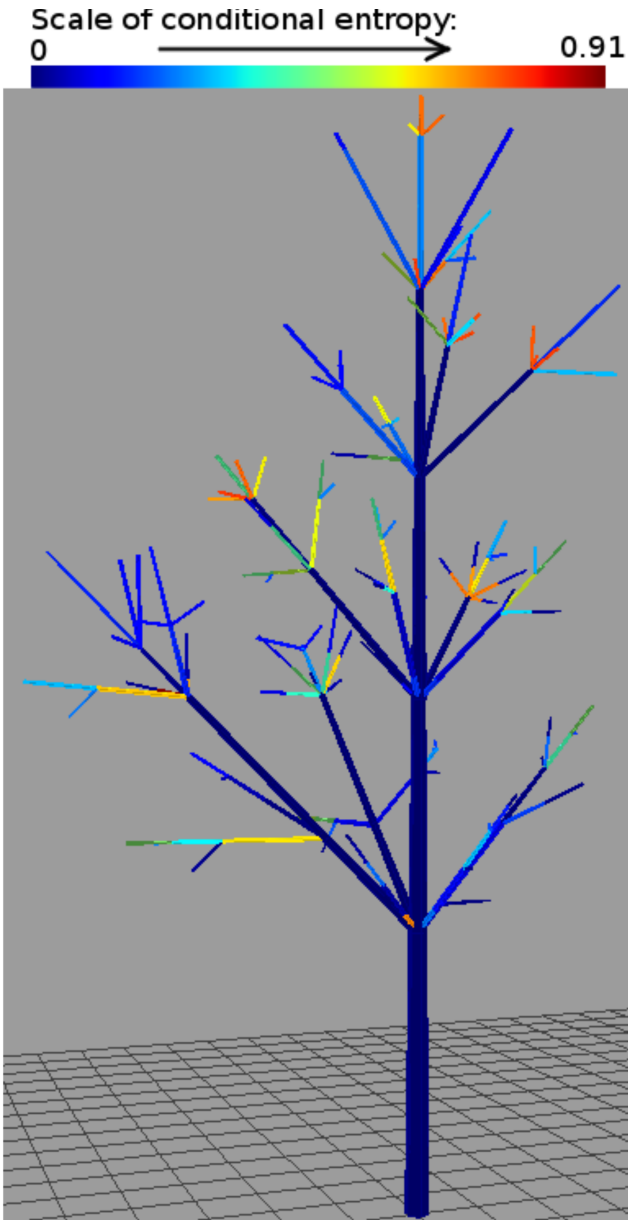

Figure 7: An example of entropy profile for HMTs. Vertices with lowest contributions to joint entropy are represented in dark blue (value: 0.0). Those with highest contributions are represented in red (value: 0.91).

computed with polynomial complexity. For such models, approximations would have to be performed using combinations of variational and mean-field methods.

Impact of parameter estimation. The above study considers the model parameters as certain. When dealing with real data sets, parameters λ are usually estimated by maximum likelihood or with Bayesian approaches. In both cases, uncertainty on the value of λ would have to be accounted for in state uncertainty; either using confidence intervals in frequentist frameworks, or by integration on λ in Bayesian frameworks, using equation

$$H(\boldsymbol{S}|\boldsymbol{X}=\boldsymbol{x}) = H(\boldsymbol{S}|\lambda,\boldsymbol{X}=\boldsymbol{x}) + H(\lambda|\boldsymbol{X}=\boldsymbol{x}) - H(\lambda|\boldsymbol{S},\boldsymbol{X}=\boldsymbol{x})$$

combined with Monte-Carlo methods sampling under the posterior $p(\lambda|\boldsymbol{X}=\boldsymbol{x})$.

4.2. Optimal timeout modelling

Generic statement of the problem

This work was in collaboration with XRCE, the Xerox Research Center in Europe. This collaboration led to the co-supervision of Laurent Donini's industrial PhD project (with S. Girard, V. Ciriza and G. Bouchard). We addressed the optimal choice of the waiting period (or *timeout*) that a device should respect before entering some sleep mode, so as to optimize a trade-off between power consumption and user impact. The optimal timeout was inferred by appropriate statistical modelling of the times between user requests. Several models were considered among the class of point processes, among which HMC-based models.

Devices have several modes associated with different power consumptions. When solicited by users, devices have to enter the so-called *operating mode*, which has highest consumption. Once the task completed, devices might be put into a mode with lower consumption after some timeout. However, switching between modes implies extra consumption. If the time X_i to the next request i were known in advance, the optimal timeout τ would be deterministic and explicitly obtained by comparing X_i to some characteristic Δ of the material, defined as the sum of mode switching costs over the difference of consumptions between modes. In the considered context, X_i was random and unknown at the end of the task, so the decision had to be taken by minimising future consumption.

Several approaches were considered in the literature to address this problem, including supervised machined learning algorithm. A detailed state of the art is provided in Durand *et al.* (2013a).

Contributions

To take precisely into account possible time-dependencies in the request process, we formulated the problem as a Markov decision process (MDP; Sutton and Barto, 2018) where the state space is the set of device modes and the decision space, the set of possible timeout values. Several criteria may be considered to represent future consumption.

In our work, we chose to minimise consumption between current and next request given times of past requests, which correspond to a reinforcement learning problem with horizon 1. We

showed that in this case, the problem has some explicit solution. Depending on whether the hazard rate z_i for the distribution of $X_i|X_1^{i-1}$ is an increasing function, a decreasing function or considering other cases, the solution is either 0, $+\infty$, or satisfies equation $z_i(\tau) = 1/\Delta$.

As a consequence of this property, the quality of the strategy is essentially determined by that of the prediction of $X_i|X_1^{i-1}$. We considered two classes of models for the request sequence: renewal processes, were the $(X_i)_{i\geq 0}$ are assumed to be independent and HMC models, which account for possible heterogeneity and abrupt changes in the conditional distribution X_i.

Under the independence assumption, three parametric families of distributions were investigated for X_i: Weibull, Gamma and Pareto. Optimal values for Δ were derived in each case. Under the HMC assumption, three strategies to predict X_i given X_1^{i-1} were considered and compared numerically on simulated and real data. As an alternative to HMC models, changes of distribution for X_i were accounted for by reestimating parameters of a renewal process using the set of past observations contained in a sliding window.

The considered data set was composed by sequences of print job submissions on different printers. In this framework, hidden states may be correspond, for example, to variable activity rates, such as business hours vs. night periods. The printer modes are the following.

- *Print mode:* The device activates its marking engine, print path and controller and completes any print requests. Power consumption is typically the highest in this mode.
- *Idle mode:* The device is ready to print immediately and therefore a certain power consumption is required to maintain the device in a state of readiness.
- *Sleep (or standby, or power-save) modes:* The device is not ready to print immediately, which induces a delay between the user request and the actual beginning of the print job. Depending on the printer, one or several such modes are available.

On the real data sets, the approaches based on renewal processes performed significantly better than state-of-the-art approaches (so-called *c-competitive*) on test sets used in cross-validation. The use of sliding windows did not significantly increase the performance. The HMC performance what comparable to that of the c-competitive approach. This may be due to the fact that heterogeneity did not impact a large proportion of requests in that specific framework.

The problem of minimising energy consumption was extended to accounting for user impact in our criterion. Indeed, in practice mode transitions of device are not instantaneous and cause some delay to availability. Such inconvenience can be incorporated into the model by adding some constant penalty δ to each occurrence of switching to the operating mode. We showed that the previous theoretical results remained valid, up to some redefinition of Δ.

Perspectives

In this study, we proposed a statistical cost-based analysis to determine optimal timeout period for devices. The theoretical formulation of power consumption in terms of a print process can be considered as a stepping stone for more complex models (e.g., incorporating covariates) that would allow the model to progressively gain completeness in the consideration of other several cost factors, as for example device ageing due to increased transitions from power saving mode

due to a more dynamic power saving policy. We also established the foundations to develop in the future a power saving strategy capable of performing accurate prediction of power saving entry as described above, but also of optimal power saving exit.

A further extension of this work is the challenging issue of optimal redirection of print jobs and power saving policy within a network of devices managed by a server. Given a user request, this consists in determining on which device the task has to be processed, and after what delay each device has to be turned into sleep mode, so as to minimize the global consumption. Modelling this problem should take into account constraints due to user impact, which are partially related to network connectivity. This could be handled in the framework of MDPs, where the spaces for states and actions now would be the space product on each device.

Finally, our approach dealt separately with model identification (parameter estimation from trajectories of user requests) and computation of optimal timeout periods (in a framework with fixed parameters). As an alternative, a unified model for handling both model identification and decision taking would be provided by the Bayesian Partially-Observed Markov Decision Processes (POMDPs) in Poupart and Vlassis (2008). Here the non-observed part of the MDP would consist in, firstly, the unknown parameters, considered as stochastic in a Bayesian framework, and secondly, potential unknown states as in the HMC models. The benefit of Bayesian POMDPs to our application would come from taking into account simultaneously the different sources of uncertainty: device and user states, value of parameter and reward (which is the opposite of expected consumption).

4.3. Coupling of hidden semi-Markov models

Generic statement of the problem

This work was motivated by the joint analysis of eye-movement signals and multi-channel electroencephalograms (EEGs). Both signals were acquired concomitantly on participants during reading tasks aiming at deciding as fast as possible whether some text is or not related to a given target topic.

In this framework, the following process features may be assumed:

- Both processes are sampled at different time scales, possibly at random times;
- Both are subject to regime switches;
- Switches of one of the processes are driven by those of the other one with random delays.

These assumptions were originally justified by our specific context, but they are applicable in other contexts, particularly in videos containing sequences of changing human activities, or monitoring persons, places or devices with several types of sensors, etc.

In our framework, the regimes are hypothesised to reflect short-term (i.e., within-trial) reading strategies. They consist in reading portions of texts more or less carefully, depending on their expected relevance to carry out the task. Information gathered through eye movements is then integrated, resorting to different zones of the brain and at different characteristic frequencies, depending on the strategy and particularly, on the degree of maturity regarding the

decision. It is thus expected that changes in strategy inferred from eye movements should be followed by specific signatures in EEGs, with some delay. Some feedback loop is also expected, since decision processes potentially perceptible at EEG level also lead to changes in strategy, i.e., in eye-movement dynamics, although this has been ignored in a first step. Sampling is performed at a fixed rate (1,000 Hz), while eye movements are sampled at fixation/saccade (and thus, random) rate. Fixations are gaze immobilisations (allowing collection of visual information), as opposed to saccades, which are brief movements of the eyes separating fixations and allowing other parts of the text to be subsequently fixed.

Coupled hidden Markov models were proposed to jointly model heterogeneous signals, particularly for EEGs. Obermaier *et al.* (2001) developed an approach based on standard HMC embedded into Simulink models; Rezek *et al.* (2002) considered discrete and continuous signals with fixed lags in a Bayesian framework. Zhong and Ghosh (2002) developed coupled HMC models with weights representing intensities of coupling. These models have been used in supervised classification contexts, e.g. to discriminate distinct tasks by building a coupled HMM for each of them. Thus, these were not designed for signal segmentation. In particular, each channel had its own segmentation, which did not allow modellers to temporally segment the whole process in a coherent manner. Moreover, non-Markovian latent processes and variable sampling rates were not taken into account.

Contributions

This work was achieved in the context of B. Olivier's PhD thesis, co-supervised with A. Guérin-Dugué. In a first step, eye-movement sequences only were taken into account and modelled with hidden semi-Markov chains (HSMCs). The approach was inspired by the work by Simola *et al.* (2008), using different protocols, types of sequences and models, though. Information criteria were used to select the number of states and to compare HSMC with HMC models. The former showed better performance in prediction and highlighted four interpretable phases in terms of information acquisition phases (or strategies): normal reading, fast reading, careful reading and decision making. Special attention was given to the choice of input variables and their coding; the impact of this choice on the robustness of segmentation (based on our work in section 4.1) was assessed and some categorical coding was chosen (fast forward and 1-step-forward reading, refixating, 1-step-backward and fast-backward reading).

The relevance and interpretation of phases were investigated using covariates; mainly, the fixation duration, saccade amplitude and reading speed in words per minute. The latter was compared with typical values provided by Carver (1992) associated with contrasted and recognized types reading activities (e.g., skimming, normal reading, learning...). Some clustering of subjects was performed, based on the frequencies of use of the different phases. The clusters were interpreted in terms of reading skills, e.g., fast, slow, careful readers. During the experiments, three types of texts were presented to subjects: highly (HR), moderately (MR) or not related (UR) to the topic. A subcategory HR+ of HR was defined as HR texts containing words from the target topic. The effect of the type of text on the phases was assessed.

The following analysis was performed to explain which semantic contents could cause phase switching. It was expected that some specific words particularly close (in HR texts) or incon-

gruent (in UR texts) to the target topic, from a semantic point of view, could lead to switch from current phase. These are referred to as *trigger words*. To quantify proximity between the target topic and words of the text, we used two existing metrics, latent semantic analysis (LSA; Dumais, 2004) and FastText (Bojanowski *et al.*, 2017). We also proposed a new metric combining the other two. The common principle is to embed words into some Euclidean space, using singular value decompositions in the first approach and artificial neural networks in the second one. The similarity between words is then defined by the cosine similarity of their representations in that vector space. This can be extended to sets of words (e.g., sentences) by summing the associated vectors. Depending on the type of text, trigger words were defined as follows:

- in HR texts, the two words that have the highest cosines with the topic (positive in principle);
- in UR texts, the two words that have the lowest cosines with the topic (negative in principle);
- in MR texts, the two words that have the highest and lowest cosines with the topic.

This is justified in MR texts by the variability of the semantic proximity to the target topic: some were rather close and some others, rather far from that topic.

As mentioned in subsection 4.1, state values and transitions are subject to uncertainty that depends on the data and model parameter. To assess the effect of trigger words on transitions, we inferred the states using MAP restoration and measured the number of fixations between times of transition and the closest trigger words. It was expected that, on the one hand, transitions occurred closer to trigger words than to other words chosen at random uniformly in the text and in the other hand, that their distance should be greater for MR than for UR and HR texts, and greater for HR than for HR+ texts.

The three metrics were compared as for their performance with respect to trigger word identification. The fastText representation proved more effective that the other two to detect trigger words.

Linear regression models were used to explain transition frequencies by the distance (in number of fixations) to the closest trigger word. Strongly negative slope coefficients were expected to highlight transitions occurring more frequently around trigger words.

The results are presented in Figure 8. In each plot, x-axes represent the distance (in number of fixations) between a transition word and the closest trigger word. Y-axes represent relative frequencies of transitions for a given distance. Three phases are considered for outgoing transitions (meaning that we are leaving the considered phase).

The effect of proximity to trigger words was particularly noticeable for transitions occurring from normal reading phases in HR+ texts, speed reading and information search in UR texts, which suggests that beginning by normal reading would be more efficient to achieve the task in HR+ texts while beginning in speed reading and information search would be more relevant in UR texts. MR slope coefficients were almost 0, pointing out that either phase transitions are not triggered by any keywords in MR texts, or that the concept of trigger words might not even

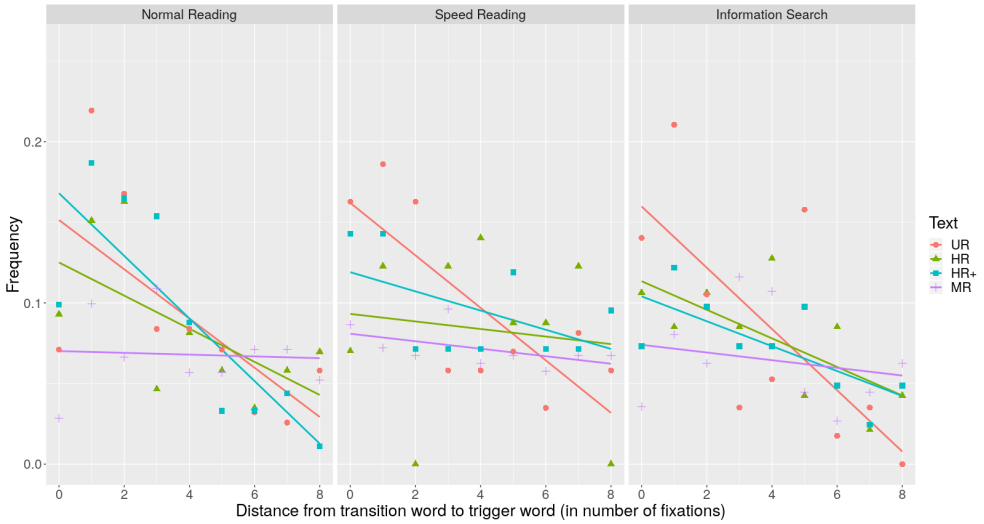

Figure 8: Frequencies of the distance between transition word to trigger word in number of fixations.

be well-defined nor relevant for some of these texts. UR texts were characterized by the most negative slopes in information search and speed reading phases and also a strongly negative slope in normal reading, showing a strong effect of incongruent words to trigger phase switches. Moreover, these texts also had the shortest times and numbers of fixations before decisions, showing that decisions are easier for such texts. HR+ texts mostly have more strongly negative slopes as HR texts do, showing that reading a word contained in the target topic has a stronger effect on phase change decision than reading a word merely semantically related to the target. This also suggests that decisions are more easily taken in HR+ than in HR texts.

In a second step, keeping the same HMSC model as before and using MAP state inference, we focused on identifying specific signatures in EEGs signals characterizing each phase. EEGs turned out to be too noisy for identifying phase-specific patterns. We thus used a time-frequency decomposition called maximal overlap discrete wavelet transform (MODWT with Daubechies-8 Least Asymmetric, Persyval 2006). MODWT is a non-orthogonal wavelet transform, in contrast to the classical discrete wavelet transform (DWT). We used MODWT because on the one hand, they yield desirable properties for wavelet correlation estimation (Whitcher *et al.*, 2000) and on the other hand, absence of decimation makes mappings between timestamps and wavelet coefficients straightforward.

EEG data were sampled at 1,000 Hz. Electrode positions are illustrated in Figure 9. Each trial had a corresponding sequence of 10 seconds and was truncated if the trial was exceeding this duration. Before each trial, 180 ms of acquisition was also available. We had a total of 2,390 trials.

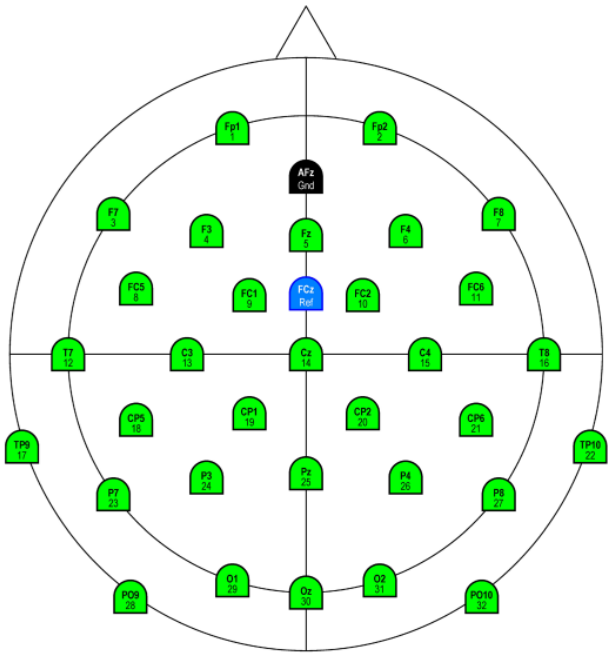

Figure 9: Electrode positions. The triangle represents subject's nose.

The correspondence between wavelet scales and brain wave frequencies is shown in Table 1. The latter have specific names (depending on their range) which are usually used in the domain literature.

Wavelet scale	Wavelet Frequency (Hz)	Brain wave	Brain wave frequency (Hz)
1	256-512		
2	128-256		
3	64-128	high-γ	32-100Hz
4	32-64	low-γ	
5	16-32	β	12.5-30
6	8-16	α	8-12
7	4-8	θ	4-7

Table 1: Wavelet scales, their equivalence in the frequency domain, and their corresponding brain waves.

Sparse anatomical representations of brain functional connectivity using graphs were obtained from inter-channel correlations using confidence intervals provided by Whitcher *et al.* (2000). Some threshold R was used for representing correlations and its choice based on Achard *et al.* (2006). The network construction methodology is summarized in Figure 10 for a given wavelet scale and a given reading strategy.

The specificity of our task lay on the decomposition of trials into phases. The aim of the study was to highlight differences of functional brain connectivity from a phase to another. To

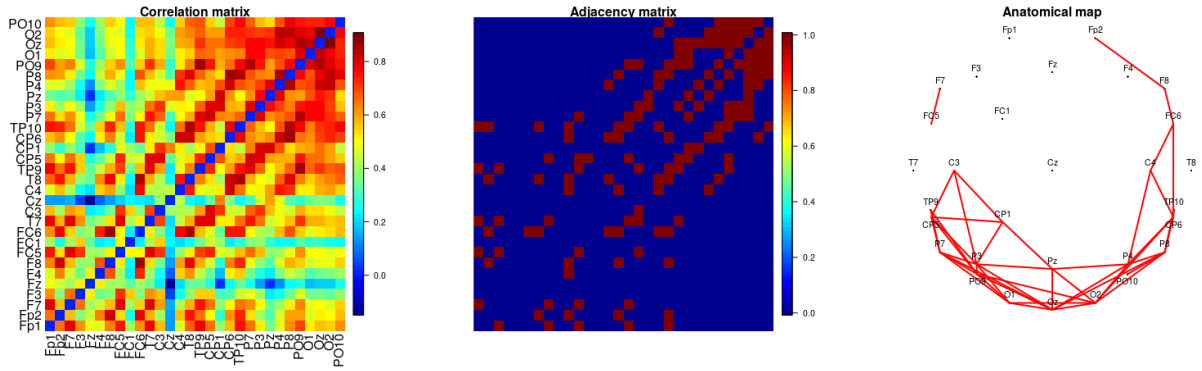

Figure 10: Network construction methodology: Correlation matrix, adjacency matrix containing significant correlations and corresponding anatomical graph for a given scale and a given reading strategy. Electrode positions in the anatomical graph are as in Figure 9.

this end, we computed the wavelet coefficients of each channel for each trial and segmented the wavelet coefficients with respect to phase changes. We then computed cross-correlations for all trials, for a given a phase, before aggregating the correlations per trial with a weighted average, weights corresponding to phase lengths.

As main outputs of this analysis, EEG activity was assessed as most salient in α and θ bands. More significant connexions were identified in the normal reading and slow confirmation phases than in information search and speed reading, the former being characterized by lower speeds and expectedly deeper sentential integration and memorization.

In a third contribution, we proposed to couple eye movements and EEGs into a single model with interpretable hidden states. Our model relies on the assumption that eye movements allow acquisition of visual information, which is then turned into semantic information and processed in different locations of the brain with an additional time delay. As rationale, since phase changes are characterized by differences in both eye-movement dynamics and channel correlations, coupled models should lead to improvements in phase-change time-localisation.

The approach applies to any multivariate process with regime switches and variable sampling rates, where switches in part of the variables drive switches for the other variables, up to some delay.

For the sake of concision, the high-rate sampling processes will be referred to as EEGs while the low-rate sampling processes will be referred to as eye movements (sampled at fixation rate). These processes are defined by the following observed quantities:

- $t \in \{1, ..., \tau\}$, the EEG temporal index in milliseconds;
- N_t, the number of fixations from 1 to t (hence N_τ standing for the total number of fixations), with $N_1 = 1$;

- T_{N_t}, the beginning of the N_t-th fixation; and similarly T_j, the beginning of the j-th fixation;
- $D_j = T_j - T_{j-1}, T_0 = 1$, the time between the j-th and the $j-1$-th fixation (i.e., the duration of the $j-1$-th fixation and associated outgoing saccade);

the notation being borrowed from the counting process terminology. Here, the potential stochastic behaviour of N_t is not accounted for.

The model is defined by some latent state process $(S_t^{(1)})_{t\geq 1}$ sampled at a fixation rate and an associated output process $(O_t^{(1)})_{t\geq 1}$, meaning that both processes have constant values between times T_j to $T_{j+1}-1$. Therefore we have, for example,

$$P(S_{T_j}^{(1)}, ..., S_{T_{j+1}-1}^{(1)}) = P(S_{T_j}^{(1)}) \mathbb{1}\{S_{T_j}^{(1)} = ... = S_{T_{j+1}-1}^{(1)}\}$$

and the distribution of $(S_t^{(1)})_{t\geq 1}$ is deduced from that of $(S_{T_j}^{(1)})_{j\geq 1}$.

EEG state and output processes are denoted by $(S_t^{(2)})_{t\geq 1}$ and $(O_t^{(2)})_{t\geq 1}$, respectively. To couple $O_{1:\tau}^{(2)}$ with $S_{1:\tau}^{(1)}$, which have different sampling rates, we define auxiliary random variables ε_j representing *lags* (or delays):

$$S_t^{(2)} = S_{T_{N_t - \varepsilon_{N_t}}}^{(1)}, \tag{5}$$

where $t \in [\![\varepsilon_1, \tau]\!]$ and ε_j represents the lag at time T_j. In practice, N_t is naturally upper-bounded by τ, the maximal sequence length, but for complexity purposes ε can be both lower- and upper-bounded, say $\varepsilon_{N_t} \in [\![0, \mathcal{L}]\!]$. Note that if $\mathcal{L} = 0$, then there is no lag and the model simply is an HSMM with multiple output processes with variable sampling rates.

Usual assumptions rule the joint process $(S_t^{(1)}, S_t^{(2)}, O_t^{(1)}, O_t^{(2)})_{t\geq 1}$, particularly: $(S_{T_j}^{(1)})_{j\geq 1}$ is an HSMC, $O_{T_j}^{(1)}$ given $S_{T_j}^{(1)} = k$ is independent from every other random variable and has distribution p_{θ_k}, $O_t^{(2)}$ given $S_t^{(2)} = k$ is independent from every other random variable and has distribution p_{λ_k}.

For the joint distribution to be fully specified, additional assumptions are required regarding $(\varepsilon_j)_{j\geq 1}$. The most realistic assumptions are the following:

- $\exists e \in \mathbf{R}, \forall j \geq 1\, \varepsilon_j = e$, where e is a deterministic parameter to be estimated (constant lag);
- $\exists (e_1, \ldots, e_K) \in \mathbf{R}^K, \forall j \geq 1\, \varepsilon_j = e_{S_{N_t}^{(1)}}$, where $(e_1, \ldots, e_K)$ are deterministic parameters to be estimated (deterministic, state-specific lags);
- $(\varepsilon_j)_{j\geq 1}$ are i.i.d. random variables with distribution p_ρ where ρ is a deterministic parameter to be estimated (random, state-independent lags);
- $(\varepsilon_j)_{j\geq 1}$ given $(S_{T_j}^{(1)})_{j\geq 1}$ are independent random variables with distributions $p_{\rho_{S_{T_j}^{(1)}}}$ where $(\rho_1, \ldots, \rho_K)$ are deterministic parameters to be estimated (random, state-specific lags);

- $(\varepsilon_j)_{j\geq 1}$ given $(S^{(1)}_{T_j})_{j\geq 1}$ are (semi-)Markov switching Markov chains with conditional transition probabilities $P(\varepsilon_j = f|\varepsilon_{j-1} = e, S^{(1)}_{T_{j-1}} = k)$, to be estimated.

As an intermediate case, a (semi-)Markov chain assumption can be considered (with marginal independence on $(S^{(1)}_t)_{t\geq 1}$). A further layer of hidden states can be introduced to model channel-specific delays. The generative procedure associated with these coupled HSMC models is depicted in Figure 11.

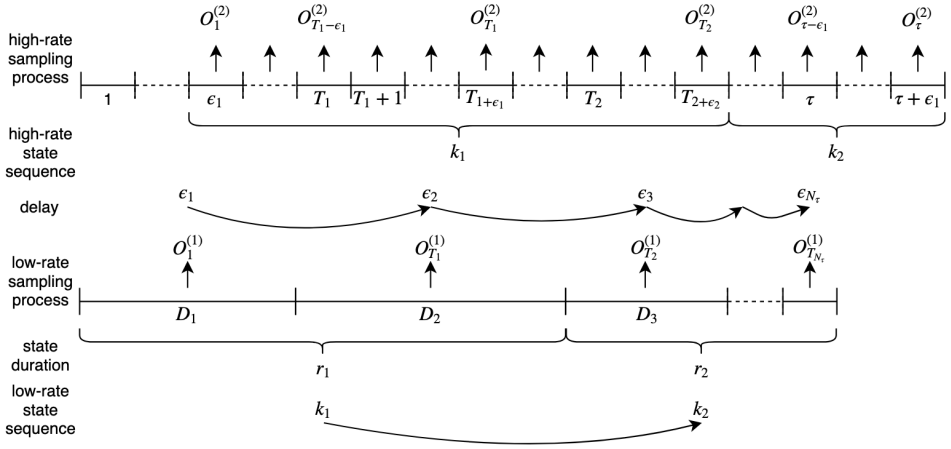

Figure 11: Coupled HSMC models: generative process. The first state $S^{(1)}_1 = k_1$ is selected using an initial probability π_{k_1}. Then, given k_1, a sojourn duration R_1 is drawn with probability $p_{k_1}(r_1)$, which means $S^{(1)}_t$ stays in state k_1 for $1 \leq t \leq T_{R_1+1} - 1$. The low-rate observations $(O^{(1)}_1, \ldots, O^{(1)}_{T_{R_1-1}})$ are sampled from emission distribution $p_{\theta_{k_1}}$. The first sampled eye-movement observation $O^{(1)}_1$ is associated with lag ε_1, intended to map the EEG to the eye-movement sampling processes. Its distribution possibly depends on state k_1. The EEG sampling process from $O^{(2)}_{\varepsilon_1}$ to $O^{(2)}_{T_1+\varepsilon_1}$, corresponding to the eye-movement observation $O^{(1)}_1$, is then sampled at each high-rate time step from $p_{\lambda_{k_1}}$, where T_1 is the beginning time of the second fixation. After that, still given $S_{T_1} = k_1$, the successive EEG outputs $O^{(2)}_{T_1+\varepsilon_1+1:T_2+\varepsilon_2}$ are drawn, as well as the associated lag ε_2, whose distribution may depend on the previous lag ε_1 and state k_1. At $t = T_{R_1+1}$ the duration in state k_1 expires and $S^{(1)}$ transits to a new state $k_2 \neq k_1$ using the transition matrix with a probability A_{k_1,k_2}. A duration R_2 is sampled for state k_2 with a probability $p_{k_2}(r_2)$ and the sampling process goes on again, until the end of the sequence.

An EM algorithm was derived for maximum likelihood parameter estimation in the general framework of semi-Markov-switching Markov delays. A specific forward-backward algorithm was developed, with complexity in $O(\tau L^2 K^2 D)$, where D is an upper bound on the support of the state duration distributions (taking $D = \tau$ if the support is infinite). A MAP restoration algorithm for the states was also developed.

It is however expected that even if in theory, maximum likelihood estimation (MLE) was consistent, in practice a huge discrepancy between the available amount of information from both sources of information (EEG and eye movements) could lead to virtually ignore the latter in EM. This is not desirable since our preliminary analyses show that the robustness of the eye-movement process (categorical process) with regard to model misspecification may be higher that than of the EEG process (assumed as conditionally multivariate Gaussian). This is why we developed an alternative EM algorithm taking benefit from the fact that $(S_t^{(1)}, O_t^{(1)})_{t\geq 1}$ marginally is an HSMC and will lead to consistent estimation of the hidden process and low-rate emission distribution, provided that sequence length tends to infinity. Then the lag parameters and high-rate emission distributions could be estimated with the other parameters considered as fixed.

Perspectives

These algorithms have not been implemented yet, so the first perspective will be to apply our coupled model on the joint EEG and eye-movement data set. This will have to be compared to a plain multivariate HSMC model. Selection will be also necessary to decide between the multiple distributional assumptions regarding delays. Our a priori assumption that coupling should reduce state uncertainty (quantified by joint state entropy, see Section 4.1) will have to be checked.

An expected limitation of the model is its incapacity to account for other sources of variability in eye movements and EEGs, mainly: individual and text effects. Our first results highlighted that individual variability in EEGs is especially high and may bias estimation of channel correlations. Such variability could be accounted for by incorporating random effects in emission distributions of HSMC models, replacing Gaussian distributions by linear mixed models (LMMs) and categorical distributions by generalized LMMs (GLMMs), as an extension of the works by Chaubert *et al.* (2010) and Peyhardi *et al.* (2017).

From a storage point of view, we have 10 seconds of acquisition of a 32 multichannel-EEG sampled at $1,000$ Hz on 7 wavelet scales for 15 subjects, 3 text types and 60 texts. Considering that this information is stored on a double of 8 bytes size, the storage requirement is approximately 45Gb. The standard EM algorithm requires all the data to be simultaneously in RAM; however such an amount may not be available on standard computers. This would lead us to turn our attention to online learning methods, which optimize the likelihood function by taking one data point or one sequence at a time. Bietti *et al.* (2015) proposed an online algorithm for HSMCs proceeding with one data point at a time, which makes it slower as necessary in our context. This issued might be solved with mini-batch versions of EM, extending to HSMCs the work by Nguyen *et al.* (2019) for mini-batch learning in mixture models with emission distributions chosen within the exponential family.

Regarding estimation of the brain functional connectivity, this has been achieved in two steps: firstly, by MLE of model parameters and MAP restoration and then, graph estimation using the restored states. This may cause some bias in graph estimation, since uncertainty on the states is not accounted for. As an alternative, estimation and selection of the structures of the covariance matrices may be performed simultaneously using penalization techniques,

with LASSO-like penalties, transferring the principles from Devijver (2017) in the context of mixtures of sparse regression models to (dependent) mixtures of Gaussian graphical models.

5. Plant structure modelling

A significant part of my research activity has been dedicated to applications of the models and methods presented in the previous sections to botanical and agronomical questions related to plant architecture (specification of spatio-temporal plant shape development).

We first present a comprehensive conceptual framework for the statistical analysis of plant architecture and then present some specific models and applications. The results presented here were mainly obtained through collaborations with the AMAP[5] and AGAP laboratories in Montpellier and particularly in AGAP, with teams AFEF[6] and Inria Virtual Plants[7].

5.1. Generic statement of the problem

For many years, plant architecture has been viewed as the result of repetitions of elementary units and patterns (Barthélémy and Caraglio, 2007) occurring through elongation and branching processes. These units may be considered at different levels of organisation: metamers (portions of stems separated by zones of insertion of leaves associated with its apical set of leaves and axillary buds), growth units (successions of metamers grown in a same cycle), annual shoots (successions of growth units grown in a same year), axes (successions of shoots) and branching systems. Here, *succession* is to be understood as opposed to *branching*, the latter resulting from development of lateral buds with respect to some reference axis. Specific patterns of growth and branching are likely to occur at several scales, which are not necessarily known in advance.

Until the beginning of the 2,000s, most approaches to describe plant architecture were qualitative and often based on *a priori* criteria, for example branching order or direction of axes (vertical, horizontal). Data were mostly acquired manually and thus, with either limited number of individuals, years of growth or spatio-temporal resolution. However on the one hand, some species turn out to be difficult to characterize that way and, on the other hand, there has been some need to progress towards quantitative methods, particularly in the perspective of plant breeding. At the same time, automatic data acquisition with laser scanners and phenotyping platforms began to emerge. The main issue regarding plant architecture was then to resume plant shape and development through a limited number of quantitative parameters.

Concomitantly, breeders and researchers in plant breeding realized that common strategies, based on quantifying heritability of global traits related to production (e.g., fruit quantity, precocity of production), had strong limitations. Optimising these traits did not seem to require any detailed analysis of growth and branching processes, which were considered as black box components of the problem. However, flowering and indirectly, fruit production are strongly connected to vegetative growth and branching. Thus, it appeared that going further into breeding of domestic plants would require more detailed modelling, quantification and

[5]https://amap.cirad.fr

[6]https://umr-agap.cirad.fr/recherche/equipes-scientifiques2/architecture-et-fonctionnement-des-especes-fruitieres/contexte-et-enjeux

[7]https://team.inria.fr/virtualplants/fr/

understanding of those interactions and eventually, of how plants gather and use resources to grow, branch and produce fruits (or wood, since similar questions may arise in forestry).

Some approaches were developed on the basis of deterministic models, such as GreenLab (de Reffye *et al.*, 2003), in which integration of various sources of structural variability was not straightforward. In contrast, we developed some battery of data-driven methods for plant architecture analysis aiming at integrating into a single decomposition model several components of structural variability; particularly:

1. Ontogeny. This component is essentially a common pattern within species. Its main architectural effect results into stationary growth phases separated by transitory or rest phases.

2. Semi-local environment. This component has a similar effect on all individuals of a same genotype at a given place.

3. Genetic. This component is common to all individuals with a same genotype, regardless of their potentially different global or semi-local environments over time.

4. Individual component. This component has an effect on all shoots within a given individual and is related to its specific (local) environment, which is *a priori* different for each individual.

From a quantitative point of view, some of these effects may be partially assessed explicitly by numerical or categorical variables. In the other cases, they have to be accounted for statistically, by specific latent variables. Interactions between several of these effects should also be considered in models.

In relation to ontogenic effects, plant components have often been reported to express gradients, e.g., more vigorous entities at the basis of annual shoots or axes for some species, or at the apex for other species. The differences between entity characteristics with respect to their positions reflect different stages of differentiation in the meristems (undifferentiated plant tissues from which new cells are formed, often at the tip a root or a stem within a bud), which are ordered in time and correspond to the notion of physiological age. Typically, the nature of the botanical entities and that of their successors tends to be equivalent while on the other hand, branching tends to induce marked qualitative changes between the bearing and borne entities.

5.2. Contributions

A generic class of models based on the notion of physiological age of meristems was proposed for tree structure and growth analysis. This is presented in subsection 5.2.1. Some extensions were developed to solve application-specific issues and are presented in the following subsections.

5.2.1. State-space models for tree structure analysis

To model ontogenic effects, we assumed that the physiological age of meristems could be assessed indirectly, i.e., deduced from measured biological characteristics of the plant entities. As a rationale, meristems with a same physiological age have a similar potential for development. We

aimed at characterising these changes by diverse quantitative or qualitative variables attached to each entity, such as the number of metamers, length, presence/absence of flowering, etc. Assuming some finite number of classes for physiological age and local dependencies owing to connected entities being produced directly or indirectly by a same meristem, we proposed HMT models (see Section 2.1) to infer physiological age from measurements. The aim was to reveal some embedded structures that were not directly apparent in the data, some regularity, patterns or levels of organisation, for instance tree-structured zones. This has been in some sense an extension to tree-structured plant data of models developed by Guédon *et al.* (2001), mainly based on Markovian models and extensions (hidden semi-Markov, hidden variable-order Markov chains).

A first model (Durand *et al.*, 2005) assumed a deterministic structure and conditional independence of children states given parent state (CIC-HMT). It was applied for illustrative purposes to model the structure of bush willows and apple trees. As expected, the model offered some quantitative synthesis of the main ontogenic aspects of tree architecture and on given individuals, a visual representation on how physiological age was spatially- and in some cases, temporally-distributed in the tree. This somehow provided some denoising of the observed variables, since their variability could be summarized through one class representing the whole distribution. State restoration could also be used to refine (i.e., reestimate) transition probabilities according to the kind of entity connexion (succession vs. branching).

This is illustrated in Figure 12 on apple trees considered at growth unit (GU) scale. The observed variables are the number of metamers and presence vs. absence of flowers (assumed to be independent given the state variable). Parametric distributions were fitted to both variables, leading us to identify four states using BIC. The states are associated to three contrasted distributions for the number of metamers. Their means thus allowed us to rename three sets of states as L(ong), M(edium) and Short. The last state was reinterpreted using the *flowering* variable, which led to S(hort vegetative) and F(lowering) states (Flowering is always short). The transition diagram associated with the transition matrix summarizes the state dynamics within trees, which roughly highlights a left-right structure: state L is mainly located at the basis of the trunk and main branches, state M at the tip of the trunk and the main branches, states S and F define a quasi-absorbing cycle corresponding to lateral twigs. This is confirmed by the restoration (depicted here on part of the tree only for the sake of readability). Distinction between transition by succession vs. branching is indicated with $<$ and $+$, respectively.

We also applied CIC-HMT models to analyse the structure of Aleppo pines (Durand and Guédon, 2016) and to quantify architectural plasticity of Beech trees and *Symphonia Globulifera* (Durand *et al.*, 2007). In the case of *Symphonia Globulifera*, individuals highlighted marked synchronism of growth and branching between branching systems along the trunk. We were able to quantify such synchronism by computing edit distances between sibling branches issued from the trunk, using the labelling provided by estimated CIC-HMT. Two variants were considered: rates of strong or weak synchronism, defined as the percentage of GUs mapped between two branches respectively without any or with state changes, as illustrated in Figure 13. In tree matching problems with multivariate vertex features, results are often sensitive to metric calibration giving unbalanced weights to insertions/deletions and the different vertex variables.

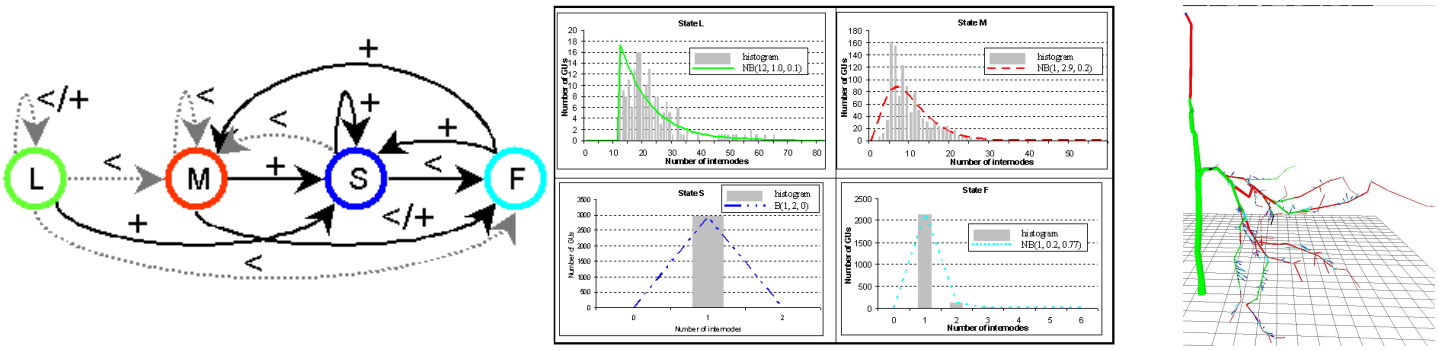

Figure 12: CIC-HMTs: an application to apple trees. Left part: transition diagram between the four states: Long (green), Medium (red), Short vegetative (dark blue) and Flowering (light blue). Edge labels $<$ and $+$ refer to transition by succession and branching, respectively. Middle part: estimated emission distributions and counts for each state. Right part: state restoration on (part of) an individual.

Our approach illustrates the added-value of HMT models in tree matching applications: state restoration acts as both a normalization and smoothing method, in which partially ordered states induce a natural metric and represent a summary of attributes, embedding them into a one-dimensional discrete space.

As useful as CIC-HMT models were to solve biological questions, they suffered from the following limitations: 1) only external variables associated to vertices could be represented (as opposed to random tree topology); 2) no potential order nor distinction between different types of children could be accounted for and 3) no direct interaction between children could be modelled. This was quite a serious limitation to address biological questions of importance such as synchronism of growth and competition (between branching systems for resources, between flowering and vegetative growth, etc.)

This is why we developed various extensions of the CIC-HMT models presented in Section 2.1. These models essentially incorporated two new aspects: random number of children and dependencies between children states given their parent state. Both were addressed within a single framework, which is multivariate count models $p(N_v^{(1)}, \ldots, N_v^{(K)} | S_v = k)$. This represents the joint distribution for the number of children $(N_v^{(k)})_{1 \leq k \leq K}$ respectively in states $1, \ldots, K$ given the parent state $S_v = k$ at vertex v, assuming K possible states and homogeneity (meaning that this distribution does not depend on s). The associated model selection issues are presented in Section 3. Given the combinatorial aspects underlying the selection of graphical dependencies, this aspect can hardly be handled simultaneously with inference of hidden states, whose distribution must satisfy constraints defined by the graph. We thus decided to proceed into three steps: 1) selection of the number of states K and state restoration using a plain CIC-HMT

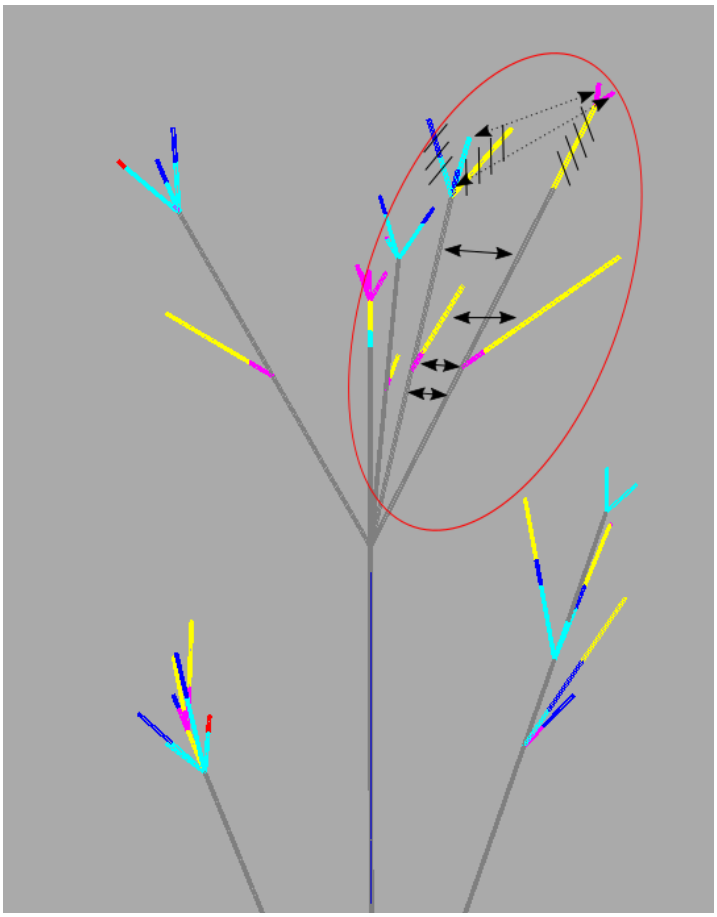

Figure 13: Mapping between GUs of two branching systems in *Symphonia Globulifera*. A mapping between two branching systems within the red ellipse is considered. Vertices mapped without state modification are represented with solid edges; those mapped with some state change are represented with dotted edges. Vertices that are not mapped are hatched.

State	Length	Flowering
0	Long	Vegetative
1	Long	Flowering
2	Medium	Vegetative
3	Medium	Flowering
4	Short	Vegetative
5	Short	Flowering

Table 2: State coding and associated features

model; 2) selection of the graphical model $\mathcal{G}$ based on the restored states and 3) reestimation of the whole model with now fixed K and $\mathcal{G}$, accounting for the conditional independence constraints induced by $\mathcal{G}$.

We illustrate this methodology through the structural analysis of two apple tree cultivars: Fuji and Braeburn (results issued from Fernique *et al.*, 2014b). Fuji is known to be alternate bearing (i.e., to tend to produce numerous flowers every two years and few flowers the other years), while Braeburn is regular. The trees are considered at annual shoot (AS) scale and modelled with unordered hidden Markov out-trees (HMOT). The states are defined as in the example in Figure 12 but since the scale of analysis is coarser, long and medium shoots may be flowering (in that case, long / medium AS are composed by a long / medium vegetative GU followed by a short flowering GU). The states are denoted as in Table 2 (even states are vegetative, odd states are flowering and states are sorted by decreasing order of vigour).

The aim of HMOT models is on the one hand, to infer dependencies between parents and children states and on the other hand, to explain how local patterns of parent / children states exclusions or competition can be translated into global bearing behaviours (bearing habits) at tree scale. Cultivars can then be compared in terms of such probabilistic patterns of exclusions. For each cultivar and each parent state value k, a parametric graphical (partially directed and acyclic, PDAG) model was estimated for $p(N_v^{(1)}, \ldots, N_v^{(K)} | S_v = k)$.

The estimated PDAGs corresponding to parent state 3 (Medium flowering) are represented in Figure 14. Each vertex k corresponds to variable $N^{(k)}$. The PDAG for Breaburn and Fuji are compared. The Breaburn PDAG highlights independence for groups of variables; in particular there is not any negative dependencies between flowering and vegetative states. Some positive correlation between $N^{(4)}$ and $N^{(5)}$ suggests that some medium flowering shoots have numerous short children, both vegetative and flowering, while some other have a low number of short children. Similar conclusion are reached for long and medium vegetative children. In contrast, the Fuji PDAG highlights negative correlations between the number of long vegetative shoots and the number of shoots in every other state but long flowering, thus showing some exclusion and alternation pattern, which is characteristic of this cultivar. The connexion between alternate bearing habits and exclusion is further discussed in Subsection 5.2.2. A more detailed case of use of graphical models in HMOTs is provided in Subsection 5.2.3.

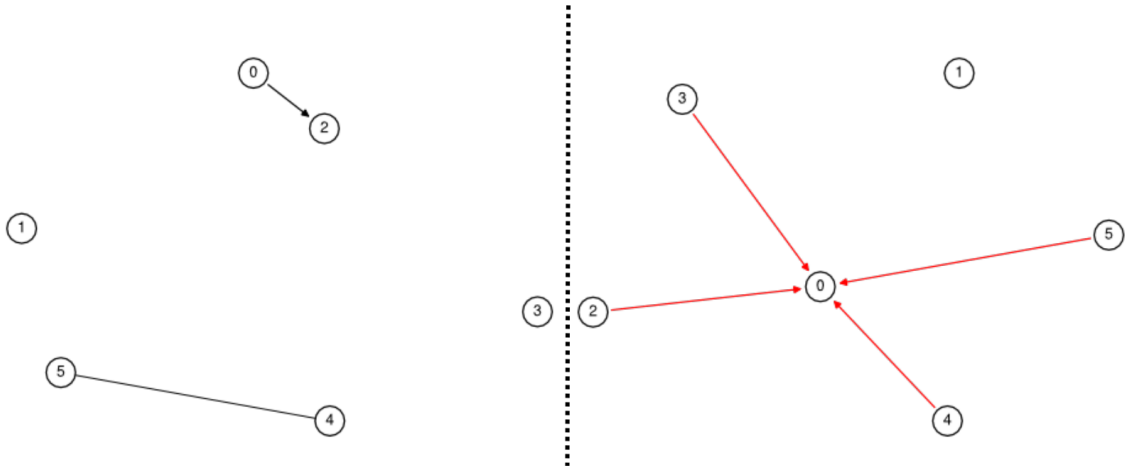

Figure 14: Selected graphical models for parent state 3 (medium flowering) for the Braeburn (left hand) and the Fuji cultivars (right hand). Edges associated with negative (resp. positive) covariances are in red (resp. black). Vertices correspond to the number $N^{(k)}$ of children in state k given the parent is in state 3.

Perspectives. In Subsection 2.1, we mentioned how partial ordering of children vertices could be accounted for in models. In the context of plant structure models, this would allow the particular role of successor child, as opposed to lateral children, to be fully taken into account. The approach is based on combinations of graphical models with some separate modelling of the successor and lateral children, obtained through conditioning on successor and parent states $p(N_{v,+}^{(1)}, \ldots, N_{v,+}^{(K)} | S_v, S_{v,<})$ where $S_{v,<}$ is the state of the successor child and $N_{v,+}^{(j)}$ the number of lateral children in state j. This would allow us, for example, to account for apical dominance or inhibition / activation of lateral flowering subsequently to apical flowering. However these models have not been implemented yet. An increased model complexity would require more data, but we expect that in future years, development of automated plant acquisition with detailed topological information (i.e., with identified AS, GU or metamers) could make such data collections accessible.

In what follows, we demonstrate how to incorporate random effects to model genetic and individual sources of structural variability in the case of sequences of phenological events. As an extension, such effects could be incorporated in PDAGs, which is discussed more thoroughly in what follows. In the same vein, some approaches developed hereafter rely on variable-order Markov chains (VOMCs), which are models for time series. The principle could be transposed to Markov tree models, where variable numbers of ancestors would be considered to model current generation distribution, depending on the values of the closest ancestors.

5.2.2. Characterisation of regularity in flowering

Irregular flowering over years is commonly observed in fruit trees. In the context of plant breeding, selection processes are slow since breeders have to wait for the trees to be sufficiently

mature to produce; in particular, this is the case to assess flowering regularity. Moreover, alternation patterns are masked during the first years of maturity by an ontogenetic trend, consisting in a general increase of the yearly numbers of flowers. We aimed at providing indices accounting for this trend (assumed as linear in what follows), achieving early prediction of bearing habits and identifying genetic determinism underlying alternation.

A first study assumed some exhaustive tree phenotyping, in which yearly total counts of flowers were available. Such data collections were unstructured, in the sense that no topological information was recorded. The performance degradation of our approach was assessed by replacing exhaustive phenotyping with sampling a reduced number of axes and replacing total flower counts by those obtained on this sample only. A second study focused on improving the previous one using sequential information brought by phenotyping axes. A third study did not consider plant breeding any more but focused on assessing the effect of water stress on flowering. The three studies were performed on apple trees and from a statistical point of view, mainly involved linear and generalized linear mixed models in higher-order or variable-order Markov chains.

Full unstructured phenotyping. The methods and results described here are issued from Durand *et al.* (2013b). A segregating population was obtained from a cross between 'Starkrimson' (STK) and 'Granny Smith' (GS) apple tree cultivars. This was used to study the bearing habits of genotypes. Two tree replicates (or just one in rare cases) were available for each genotype. The STKxGS progeny was composed of 123 genotypes.

Flowering recurrence was measured at two different scales: whole tree and AS. At the whole tree scale, the total number of inflorescences was observed during six consecutive years, from their second to their seventh year after grafting. At AS scale, the succession of vegetative vs. floral AS were observed over the same consecutive years along different axes. The data consisted in 4 to 36, 1 to 6 year-long sequences of vegetative vs. floral AS per replicate.

A two-step modelling was used to quantify biennial bearing at whole tree scale. To dissociate increase in the number of inflorescences per tree from biennial bearing, a trend model based on a linear mixed model was applied firstly. Irregular bearing was quantified afterwards using the deviations around the trend (model residuals) and combining two approaches: (i) a new index was proposed to quantify the amplitude of residuals, taking account of their order; and (ii) an auto-regressive model was estimated to characterise the dependencies between successive residuals and distinguish biennial patterns specifically from other irregular patterns, using an autocorrelation parameter. Then, clusters of genotypes that have similar patterns of annual yields were identified using both descriptors jointly.

To quantify biennial bearing using only a limited number of AS sequences, the descriptors presented hereabove were computed using the same procedure, except that the annual amount of inflorescences in the sequences was used instead of the annual amount of inflorescences at whole tree scale.

Linear mixed models were considered for the trend. These were of the form

$$Y_{g,r,t} = \beta + \beta_g + \zeta_{g,r} + (\alpha + \alpha_g + \xi_{g,r})t + \varepsilon_{g,r,t}$$

where $Y_{g,r,t}$ is the number of flowers of tree replicate r of genotype g at year t, $\beta, \beta_g, \zeta_{g,r}, \alpha, \alpha_g, \xi_{g,r}$ are unknown parameters and $\varepsilon_{g,r,t}$ is a random residual assumed Gaussian with zero mean. Different models were compared, assuming the unknown parameters as either fixed or independent random effects (except for α and β, treated as fixed). These linear models only accounted for the trend and not alternation, rather seen as structured patterns of deviation around the trend. We thus considered the following model for residuals:

$$\varepsilon_{g,r,t} = (\gamma + \gamma_g + \gamma_{g,r})\varepsilon_{g,r,t-1} + u_{g,r,t}$$

where as above, γ is a fixed parameter, γ_g, $\gamma_{g,r}$ are treated either as fixed or independent random effects and the $u_{g,r,t}$ are assumed as independent, zero-mean, unknown variance Gaussian random variables.

The fixed parameters were estimated by maximum likelihood and using these estimates, point estimates of random effects were obtained using posterior expectations. Negative values of γ_g can be interpreted as alternate bearing genotypes g, since their yearly numbers of flowers tend to alternate around the trend. Regular genotypes tend to have their γ_g values close to zero. However, γ_g does not account for the intensity of alternation relative to the total number of flowers. This is why we introduced another index, inspired by the Biennial Bearing Index (BBI) of Hoblyn *et al.* (1936) but now applied to $\varepsilon_{g,r,t}$ and normalized by the total number of flowers:

$$\text{BBI\_res\_norm}_g = B_g = \frac{\sum_r \sum_{t=2}^{T_{g,r}} |\hat{\varepsilon}_{g,r,t} - \hat{\varepsilon}_{g,r,t-1}| / \sum_r (T_{g,r} - 1)}{\sum_r \sum_{t=1}^{T_{g,r}} Y_{g,r,t} / \sum_r T_{g,r}}$$

where $T_{g,r}$ is the number of measurements for replicate r of genotype g.

Genotypes were clustered using Gaussian mixture models in the plane (B_g, γ_g) characterizing each genotype (Figure 15). Cluster 1 can be interpreted as regular bearing genotypes, cluster 2 as biennial bearing genotypes and cluster 3, as irregular bearing genotypes. The fuzzy notion of bearing habits was redefined as cluster values of genotypes.

The model was validated by out-of-sample prediction, using the number of flowers yielded during the first years to predict the last year using prediction intervals at level 0.95. The frequency of actual number of flowers being within the prediction interval was 0.74, showing underestimation of the variance. This caused some 17% error in cluster prediction.

Our B_g and γ_g indices were computed using sequences collected on axes and using yearly numbers of flowers on those axes in lieu of total numbers of flowers at whole tree scale. A binary random variable (flowering vs. vegetative) could now be associated to each event along the sample axes, so we used its sample entropy to quantify the synchronicity of flowering for every axes in a given year and averaged it out over years to obtain a third index at genotype scale. These were used to predict the bearing genotype, yielding some 40% classification error (to be compared to that of the best random classifier independent on predictors, i.e. 56%).

Ideally, breeders would not need to wait for the trees to reach maturity before assessing their values for breeding. They may use predictions obtained from their genotypes. Moreover, it is of interest for researchers to understand or at least, make assumptions on functional determinants for alternation in flowering. In that perspective, our descriptors were used for QTL detection (portion of DNA whose variation is correlated with that of descriptors). Four QTLs were

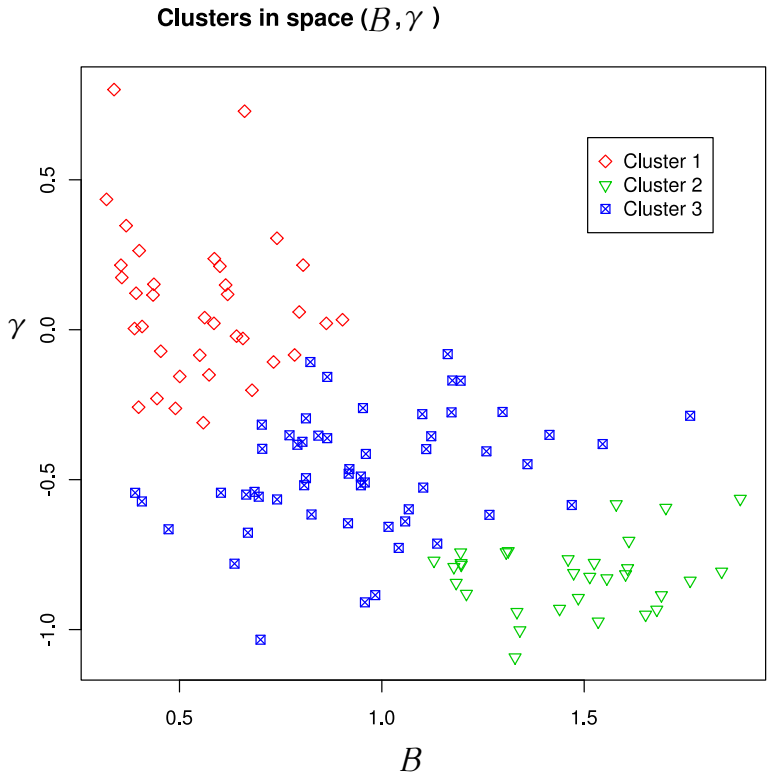

Figure 15: Clustering obtained using a three-component mixture of Gaussian distributions of B_g (x-axis) and γ_g (y-axis).

identified for BBI_res_norm and some variant thereof. No significant QTL was detected for γ_g computed at whole tree scale nor for entropy.

Both genomic regions associated with BBI_res_norm and its variant corroborated zones that were previously identified in a previous study by Guitton *et al.* (2012). The QTLs co-located with QTLs for inflorescence yield of a given year and fruit yield QTLs for the year before in Guitton *et al.* (2012). One of these seemed to be linked to the antagonist relationship of fruit production of the current year and inflorescence development for the year after that was reported by Banghert (2009).

Further QTLs were revealed on γ_g computed from subsamples on axes. These were located on zones that were not previously associated to flowering or bearing traits in this progeny, even though QTLs were detected in the same regions for branching and internode length respectively (Segura *et al.*, 2009).

Phenotyping axes. The previous study mainly aimed at providing reference indices, bearing habits and QLTs based on a limited set of progenies and on exhaustive phenotyping of yearly numbers of flowers. It also considered simplified phenotyping on a sample of axes but did not

fully account for the data structure, since we merely used the yearly numbers of flowers on those axes as if these were total numbers of flowers at whole tree scale. We assumed that the analysis of entire sequences of successive AS, combined to flowering synchronicity in each year, would provide new insights on genotype behaviours and yield more accurate predictions. We also extended the previous investigations by performing a multi-family QTL detection to enlarge the genetic basis of biennial bearing variation in apple trees. Thus, sequences of vegetative vs. floral annual shoots (AS) were observed along axes in trees now belonging to five related apple full-sib families located in two sites (Montpellier and Angers in France). The data consisted in 1 to 162, 1 to 7 year-long sequences per replicate. The methods and results presented here are issued from Durand *et al.* (2017).

Sequences were analyzed using VOMCs initially introduced in this context by Costes and Guédon (2012). In such stochastic processes, the random variable at time t depends on a variable number of past values, depending on these values. Applying a memory selection procedure by Csiszár and Talata (2006), a fixed-order Markov chain with order 2 was chosen. However, as highlighted by our previous study, alternation is partly due to genetic effects. Moreover, it can be assumed that environment has some effect, in potential interaction with memory m. Its impact is mainly characterized by synchronous fluctuations in the numbers of flowers for all trees at a given year t for a given site π. To model these interactions in sequences of binary observations, approaches based on GLMMs (Molenberghs and Verbeke, 2006) seemed relevant. Our model was thus a second-order Markov chain with transition kernels defined as logistic regression with mixed effects.

Denoting by $F_{g,r,\pi,t,\ell} = 1$ (resp. 0), the presence (resp. absence) of flower for replication r of genotype g at site π, year t, and location (or AS) ℓ in the axis, the following Markovian GLMM was considered as a transition kernel:

$$\log \frac{P(F_{g,r,\pi,t,\ell} = 1)}{P(F_{g,r,\pi,t,\ell}) = 0} = \lambda + \rho_\pi + \mu_m + \Phi_t + \theta_{g,m} + \eta_{g,t} + \zeta_{g,r}$$

where λ is a fixed intercept, ρ_π is the fixed effect of site π, μ_m is the fixed effect of memory m, Φ_t is the fixed effect of year t, treated as a qualitative variable, variables $\theta_{g,m}$ are independent random interactions between genotype g and memory m with common variance τ^2_θ, variables $\eta_{g,t}$ are independent random interactions between genotype g and year t with common variance τ^2_η. Lastly, variables $\zeta_{g,r}$ are independent replication-specific random effects with common variance τ^2_ζ. All random effects were assumed to be mutually independent and Gaussian. Parameter estimation was by restricted maximum likelihood. For a better interpretation of the model and in order to obtain new indices, the value of random effects $\theta_{g,m}$ were estimated by their conditional means (Best Linear Unbiased Predictors, BLUPs). These were used to discriminate genotypes on their low vs. high probability of AS bearing flowers at year t given they had memory m. Similarly the BLUPs of $\eta_{g,t}$ were used to discriminate genotypes on their low vs. high probabilities of bearing flowers at year t. The GLMM was also used to remove site and environment effects from the flowering probabilities and plug those into the computation of flowering entropies.

Such entropy values and the $\theta_{g,m}$ indices were added to the axis-based B_g and γ_g introduced in our previous study (Durand *et al.*, 2013b) to check the assumption that taking into account

dependencies through indices should lead to better predictions of bearing habits. The reference family, for which bearing habits were assumed as known, was STKxGS. However, some indices could not be defined for some genotypes; for example, if memory 11 did not occur in any axis of genotype g, $\theta_{g,11}$ would not be defined, resulting into missing predictors. Moreover, some predictors were highly correlated and redundant. To handle both issues of predictor absence and redundancy, a principal component (PC) analysis for partially missing data was used, using the methodology and R package by Josse and Husson (2012). Classification was performed using neural networks (NNs). Such models were also used to predict B_g and γ_g at whole tree scale. The NN parameters were estimated by least squares minimization. Since the optimal numbers of PCs to be used in classification and regression may be different, they were both chosen independently by out-of-sample validation.

Classification of bearing habits yielded a 35% (cross-validated) error rate using our new indices, to be compared with 40% obtained in the previous study. Although limited, the improvement of the error rate was significant at level 0.7% on 50 random test samples. Concerning regression, the (cross-validated) correlation between true and predicted B_g (resp. γ_g) at whole tree scale was 0.72 (resp. 0.64) when using our new indices, against 0.71 (resp. 0.60) when using just B_g, γ_g and entropy at axis scale. Thus, adding information from Markovian GLMMs did not significantly improve the prediction of the tree-scale indices.

Our new indices were used for QTL detection and QTLs were detected for all indices except for $\theta_{g,10}$ and most $\eta_{g,t}$ indices. Those with strongest evidence were associated with BBI-like but not with the new $\theta_{g,m}$ indices. Genotype estimation at QTLs was performed, bringing two types of information: Firstly, the allelic classes of parents allowed families in which QTLs segregated to be identified; Secondly, this allowed parents and founders bearing favourable, variable-specific alleles, to be highlighted.

Despite the lack of major QTL associated with our $\theta_{g,m}$ indices, their analysis brought new insight on the respective roles of synchronicity and alternation at axis scale in whole-tree scale alternation. The regular genotypes (lowest values of BBI_res_norm) exhibited AS with flowering probabilities above average at year t after flowering at year $t-1$ (as emphasized by higher $\theta_{g,01}$ and $\theta_{g,11}$ associated with memories 01 and 11). The positive correlation of entropy with γ_g and its negative correlation with B_g showed that the genotypes with highest synchronism (lowest values of entropy) were mostly biennial bearers (highest values of B_g and lowest values of γ_g). This suggested that biennial bearing at tree scale results (at least in the considered populations) from the conjunction of two phenomena: synchronism in flowering between AS in a given year and biennial alternation at AS scale between consecutive years. On the contrary, regularity at tree scale results from either asynchronous locally alternating flowering or regular flowering at AS scale. By contrast, trees with low values of B_g, medium or high γ_g values and high entropy values at the tree scale (regular bearing by total absence of structure in flowering) could not be observed in the studied populations. Irregular genotypes exhibited intermediate values for every descriptor, suggesting that such genotypes are characterized by partial biennial alternation along axes or strong biennial alternation with partial synchronism.

Our indices led to practical recommendations for breeders; particularly, selecting genotypes with regular desynchronized axes could be an appropriate strategy for avoiding poor fruit set

while reducing thinning or manipulations costs. Moreover, three descriptors should be combined to ensure regular bearing behaviour, i.e., B_g, γ_g and entropy.

Effect of water stress. The previous study on alternation in flowering consisted, from a statistical point of view, in modelling the fates of successive buds over years and in identifying the effect of past fates, year, location and genotype on flowering probabilities using Markovian models with GLMM transition kernels. The study summarized hereafter, issued from Yang *et al.* (2016), takes profit from a similar approach to model the effect of water stress on apple tree architecture and flowering.

Water deprivation generates a number of physiological and morphological responses in plants, depending on the intensity and duration of stress, plant species and development stage. In perennial plants, water stress may affect plant development through cumulative effects that modify plant functions, architecture and production over years. Fruit trees are usually irrigated and their growth and production heavily depend on water availability and irrigation in many countries. Nevertheless, the shortage of water has become a critical problem in fruit tree orchards, especially in arid and semi-arid regions. In such situations, lower irrigation supplies decrease mean apple fruit mass.

Due to the regulation of shoot growth, branching and flowering year after year in perennial plants, the effect of water deprivation should be considered in the long term, taking into account subsequent modifications of plant architecture. The latter is determined by the fates of the terminal and axillary buds that can give rise, in the particular case of apple, to reproductive or vegetative GUs of different lengths, as seen in Subsection 5.2.1. In this study the impact of long term (7 years) water deficit on the fate of terminal and axillary buds was investigated in relation to flowering occurrence and production patterns (biennial vs. regular) in the "Granny Smith" cultivar.

Our study aimed at analyzing the effect of water stress over years, at different scales of plant organization (whole tree, branch, axis and GU) on apple trees. Our hypothesis was that long-term water stress would modify the composition of shoot types (vegetative vs. reproductive, long vs. short GUs) within a branch, with potential repercussions at whole-tree scale on fruit production patterns (regularity vs. irregularity). The following questions were addressed: (1) Can a decrease in primary growth (GU length) be observed in response to water stress? (2) Are the inter-annual transitions between GUs modified by water stress? (3) Does water stress modify the floral GU frequency and production patterns at whole-tree scale?

Plant material consisted in sixteen "Granny Smith" trees organized in several rows with well-watered (WW) trees alternated with rows of trees submitted to restricted soil water supply (WS). There were originally eight trees per treatment but in each each group, two trees were damaged or died. For each tree, all the branches that arose from the first and second annual GUs of the trunk except those that were broken during the experiment were selected and analysed (from one to five branches per tree eventually). Sequences of labelled GUs were recorded, using the following labels: L(ong), M(edium), S(hort) vegetative GUs, F(loral) and D(ead) GUs, meaning they were dry and did not produce any new GU in terminal position until the end of the experiment. The data set contained 17 and 20 branches, 3525 and 3464 GUs for WW and WS trees, respectively.

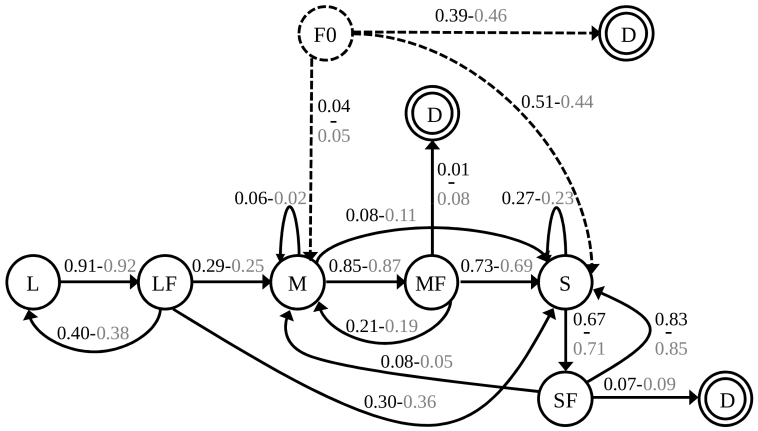

Figure 16: Transition frequencies between GU types under WW (black numbers) and WS (grey numbers) treatments. Transition frequencies were estimated from first (L, M, S for long, medium and short GU and F0 for the initial flowering) and second order memories (LF, MF, and SF for long-floral, medium-floral and short-floral). State D corresponds to the growth cessation of branches. Transition frequencies lower than 0.05 are not represented. Circles and arrows represented by dashed lines correspond to transitions and memories that can be observed only once in GU sequences after an initial inflorescence, whereas continuous lines and circles correspond to memories and transitions that can occur several times.

The GU sequences along axes were analysed using VOMCs. The maximal order, as determined by BIC in merging WW and WS trees, was two and the ML estimated transition matrix is depicted by the diagram in Figure 16.

Then the effect of water stress was incorporated into the model, using multinomial GLMs to represent transition probabilities:

$$\log \frac{P(T_{t,m,w,\ell} = c)}{P(T_{t,m,w,\ell} = S)} = \lambda_c + \mu_{m,c} + \varphi_{w,c} + \theta_{m,w,c}$$

where $T_{t,m,w,\ell}$ is the label of GU at location ℓ in tree t, with memory m and treatment w, c is either L, M, F or D, λ_c is an intercept, $\mu_{m,c}$ is the effect of memory m on transitions to label c, $\varphi_{w,c}$ is the effect of water treatment w and $\theta_{m,w,c}$ is the effect of interaction between memory m and water treatment w. The model parameters were estimated by maximum likelihood.

Five models were built: model 1 took into account the effect of memory only, model 2 took into account the water treatment effect only, model 3 included both effects and model 4 included both effects plus their interactions. The significance of the different effects (water conditions, memory and interactions) was assessed using a chi-squared likelihood ratio test. To assess whether the chosen significant level of the tests was sufficiently small, considering that the four models could all be incorrect and that the number of data was large, the models were also compared using BIC (Kass and Raftery, 1995). The 5^{th} model was obtained by adding a random "tree" effect to the best of the four previous models.

To assess the effect of water stress on the number of axillary, short and medium GUs, together with the probability of GU death and fruit production, LMMs and GLMMs with either binomial or Poisson distributions were estimated. Test were performed on their parameters, considering branches as replicates, trees as a random effects, treatments and years as fixed effects. The significance of fixed effects was then assessed using χ^2 tests. Significance of random tree effects was assessed by computing confidence intervals for the variance parameters at level 95%.

As main results of this study, branches exhibited two distinct development phases, one from the 2^{nd} to the 4^{th} year of tree growth, characterized by the occurrence of long GUs and the beginning of flower production, and one starting in the 5^{th} year, including patterns of alternation between vegetative and reproductive GUs. Both development phases were observed independently of the soil water status, which confirmed the stability of ontogenetic characteristics of plants even if they were subjected to environmental stress.

Water stress led to some significant decrease of the total number of GUs, some increase of the proportion of short GUs and of the transition probabilities towards small and dead GUs. This suggested an acceleration of ontogenetic gradients, as observed under stressful conditions on other species (e.g., walnut and almond). This could in turn be beneficial to floral induction. However, the decrease in the length of vegetative GUs and the lower number of GUs led to a reduction in tree vigour under water stress, as highlighted also by some decrease in the trunk cross-sectional area.

An increase in the proportion of flowering GUs was observed under water deficit, as shown directly by modelling the proportion of floral inflorescences along axes and indirectly, by the VOMC highlighting higher transition probabilities towards floral GUs under water stress (as depicted in Figure 16).

Regarding alternation, higher fruit numbers were observed under WS in the years following "off" years, thus reducing biennial bearing patterns. Individual fruit weight was reduced under water deficit conditions.

Conclusion and perspectives. We developed a methodology aiming at decomposing the effect of main growth components in tree architecture, summarized here through sequences of elementary tree units, by plugging GLMMs into Markovian models. The considered components were: ontogeny, environment (e.g., year, site, water stress).

This is illustrated here in applications related to apple trees: quantifying alternation in flowering or the effect of water stress on architecture. Our approach could be directly used on any species for which retrospective phenotyping of flowering is possible at AS scale. This is the case for species with terminal flowering such as pear, walnut, avocado, mango, litchi, etc. in which flowering events can be easily identified. For such species, the methodology proposed, including prediction of flowering behaviour at tree scale from *a posteriori* observations and computation of indices, would be transposable. For other species, alternation indices could be computed based on counting the total number of inflorescences measured on several successive years. Even though more time-demanding than retrospective observations, such counts may be facilitated and automatized by new technologies based on imagery.

Our results also suggest that flowering synchronicity at whole tree level could not be associated with regularity, probably because it would lead to over-cropping and major drawbacks

in an agronomic context. Knowing if flowering desynchronization has been selected during the apple domestication remains an open question.

The obtained results on bearing habits offer new perspectives to decipher the putative role of fruits and carbon economy on the inhibition of floral induction. Indeed, both hormonal and trophic hypotheses were proposed to explain this phenomenon. According to the hormonal hypothesis, floral induction is inhibited by gibberellin signalling coming from seeds, while according to the trophic hypothesis, this is limited by carbon availability. In the one hand, our models could be used to classify genotypes before investigating their physiological behaviours. On the other hand, the water stress study highlighted that higher fruit numbers were observed under WS than under WW in the years following "on" years, reducing the biennial bearing pattern. One hypothesis to explain this observation is related to the decrease in vegetative growth under WS, which could reflect non trophic mechanisms, but rather hormonal signalling or changes in cell hydraulic proprieties under moderate stress. In turn, this reduced vegetative growth could have favoured plant growth processes such as floral induction and fruit set, possibly through a higher assimilate availability.

The studies also provide information that could be relevant in simulation models. For instance, variable-order Markov chains with parameters that depend on environmental conditions could be integrated into functional-structural plant models (in particular MappleT regarding apple trees – see Costes *et al.*, 2008). Such improvements could be an interesting way to further analyse the impact of modifications in GU successions and branching at shoot scale on biennial bearing, under contrasted environmental conditions. This could open new perspectives for in *silico investigations* of agronomical scenarios.

From a methodological point of view, the main perspective is to extend analyses performed on axes with Markovian models for sequences, to whole plants using Markovian models for trees. This is partly illustrated in the next subsection and would rely on multivariate regression models for count data in the exponential family, as discussed in Section 3. The added value of handling directly tree-structured data would be to characterize more precisely the respective roles in axillary production (quantities, fates and dependencies, including fates of apical shoots) of different components: ontogeny, growth conditions and genotype. In particular, this would open new perspectives in determining the origin of synchronism / asynchronism in flowering when considering different axes of a same individual.

Moreover, Markovian dependencies were not fully taken into account in our statistical analysis. Due to the factorization property of the likelihood function, point parameter estimates, likelihood values and information criteria were obtained consistently with 2^{nd}-order mixed transition kernels assumptions, using the standard "lme4" library of the R software. However, confidence intervals and p-values are provided in this library under an independence assumption. As a consequence, further theoretical developments are required to account for Markovian dependencies in this framework, as an extension of the work by Islam *et al.* (2009).

5.2.3. Modelling phenology and patchiness

Several species in temperate and tropical zones are characterized by strong phenological asynchronisms between and within trees, entailing patchiness. The latter is defined as growth

development organized in clumps of either vegetative or reproductive shoots within the canopy. For example, some parts of the canopy may develop vegetative shoots, while other parts remain in rest or produce shoots bearing flowers (referred to as reproductive shoots). Alternation in flowering at whole-tree or axis scales, as investigated in Section 5.2.2, can be seen as other particular cases of patchiness.

Such flushes of growth, either asynchronous or heterogeneous regarding natures of produced shoots, entail various agronomic problems. These are mainly: (1) Repeated use of pesticides to protect recurrent sensitive phenological stages from pests; (2) Excessively extended periods of fruit maturity, leading to difficulties in organizing harvesting.

Our objective here was to define a statistical methodology to identify and quantify such patchiness patterns. The identification of architectural patterns in trees was in a first step addressed through modelling local dependency properties (e.g., HMT models, see Subsection 5.2.1). It is questionable here whether local dependencies are sufficient or not to model patchiness patterns. The latter may indeed require to infer *a priori* unknown scales of aggregation of similar botanical entities, which is not directly achieved in HMT models. Considering the issue of identifying structural patterns within sequences of botanical events, two families of statistical models emerged: HSMC and multiple change-point models. In the following study, we proposed firstly, to extend multiple change-point models to tree-structured data in order to identify patches at various scales in trees and secondly, to complement the approach with HMT analysis.

From an agronomic point of view, quantification of patchiness was used in cultivar comparison and as a perspective, could be integrated into varietal selection procedures or technical arrangement studies.

Regarding patch identification using multiple change-point models, the methodology presented here is mainly issued from Fernique *et al.* (2016b). The greedy algorithm presented in Subsection 2.3 was applied to the data set, after some shoot labelling. Its output was a partitioning of trees into subtrees constrained to have their adjacent subtrees significantly different from each other in terms of shoot label distributions. However on the one hand, distributions associated with non-adjacent subtrees (even those issued from different trees) may be similar. On the other hand, patches may be composed by shoots with different labels. We therefore proposed a two-stage tree-quotienting/subtree-clustering algorithm incorporated some post-processing to change-point detection based on mixture models in order to identify similar subtrees. In such mixtures, latent states were constrained to be the same within a quotient obtained by change-point detection. The number of patch types was selected using BIC.

Although the patches obtained by this approach were inferred from trees considered at the finest scale, they could actually occur at some more integrated scales. We thus chose to assign each patch to the closest available biological scale described in the data (e.g. metamer, GU, scaffold, tree). Since patches did not necessarily match any of these quotients defined *a priori*, a method to determine the minimum distance between quotients obtained from the tree-quotienting/subtree-clustering algorithm and scales of interest was necessary.

Hence, in order to determine the scale of a patch, we proceeded as follows:

- For each patch and scale of interest (defining partitions), vertices were assigned binary labels in accordance with their membership or not to the elements of the partitions.

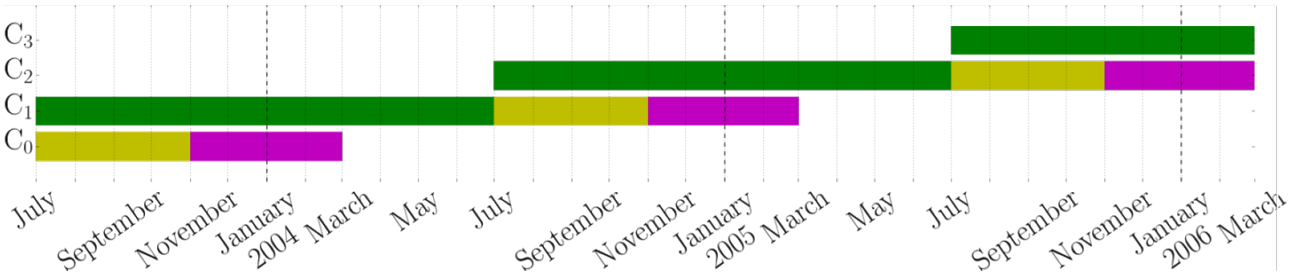

Figure 17: Mango tree growth cycles. Three phases were considered in mango tree growth cycles: vegetative (green), flowering (yellow) and fruiting phases (magenta)

- Distances between the binary partition induced by the patch and each binary partition induced by quotients at a biological scale were computed using the Rand index.
- Patches were assigned to the scale that minimizes that distance.

Finally, to take account of the temporal dimension of patchiness, directed acyclic graphs (DAGs) were built to summarize the succession of patches over time for a given observed tree.

Application to mango tree. This approach was applied to patchiness quantification in mango trees. Seven cultivars were considered: Cogshall, José, Kensington Pride, Irwin, Kent, Nam Doc Mai and Tommy Atkins. For each of them, five mango trees were described at GU scale. Since tree patchiness is a spatio-temporal phenomenon, its analysis requires the choice of a temporal resolution in order to observe the production of growth cycles over a given period. Due to this choice, the temporal component of patchiness is thus partly subjective. Nevertheless, due to the extended growth cycle in mango trees, three periods of marked interest emerge:

- The early flush period (E). This corresponds to the period when the vegetative phase of a growth cycle overlaps the flowering phase of the previous cycle.
- The intermediate flush period (I). This corresponds to the period when the vegetative phase of a growth cycle overlaps the fruiting phase of the previous cycle.
- The late flush period (L). This corresponds to the period when the vegetative phase of a growth cycle does not overlap the previous or the next cycle.

The definition of the three flushes is illustrated in Figure 17.

Patchiness was thus investigated at the flush temporal resolution, with the mango tree material being observed during seven successive flushes at most. In this way, this spatio-temporal phenomenon was decomposed into several spatial analyses at fixed times, which had to be recombined *a posteriori*. At each time step, only the tree fringe issued from GU growth, flowering or quiescence since previous time step was considered. GUs in that fringe were labelled either as V(egetative), (R)eproductive or Q(uiescent).

This is illustrated in Figure 18, where the same mango tree is observed at the end of:

- The intermediate flush of the first growth cycle, denoted as I1 (Figure 18a). The canopy is then composed of a reproductive patch containing GUs bearing fruits and a quiescent patch containing GUs that burst during a previous flush.
- The late flush of the first growth cycle, denoted L1 (Figure 18b). The canopy is then composed of a quiescent patch and a vegetative patch containing GUs that burst during current flush.
- The early flush of the second growth cycle, denoted E2 (Figure 18c). The canopy is then composed of a single quiescent patch.

Once patches have been identified, observing their succession in time directly on trees is not convenient. We proposed instead space-time representations using DAGs. In such DAGs, each vertex corresponds to some patch at a given flush (corresponding to its position on the x-axis). Each directed edge linking two vertices has a temporal meaning, accounting for splitting or merging patches in the next flush. DAGs also encode patch compositions in terms of V, R and Q labels (as determined by the mixture model) and characteristic scales of patches (see details hereafter). This is illustrated in Figure 19, where the succession of patches within sketched mango trees in Figure 18a-c is represented. Vertex colours represent patch type shapes and scales. At Flush E1, since no information was available for identifying patches, the DAG was initialized with an unlabelled single patch at tree scale. Then at Flush I1, two patches (one reproductive and one quiescent) were identified. At Flush L1, the reproductive patch split into two patches (one vegetative and one quiescent), whereas the quiescent patch at Flush I1 turned into a vegetative one. At flush E2, the three patches merged into a single quiescent patch.

Patch compositions are represented in Figure 20. Most of them were nearly pure. This highlighted the relevance of our clustering stage.

The mixture model led to identifying 5 clusters, corresponding either to dominance or rarity of some label within the patch. State proportions highlighted a slight excess of patches containing vegetative GUs, but all types of patches were clearly present. This excess of vegetative patches is biologically justified, since the observed mango trees were young and therefore not at their permanent regime of production, in which more flowering GUs would be expected.

To compare cultivars in terms of patchiness, the DAGs associated to each plant were summarized using the following properties: number of DAG vertices, proportions of flushes, scales and clusters, average in- and out-degrees, together with ratio of edge number to maximal edge number. Then some Linear Discriminant Analysis (LDA) was performed, using cultivars as classes. Since the number of variables was large regarding the number of individuals per cultivar, a sparse version by Clemmensen *et al.* (2011) was preferred, with model selection based on cross-validation. The projections of cultivars in the first plane are represented in Figure 21.

Cultivar positions in the first LDA plane highlighted contrasted patterns between some cultivars in terms of patchiness. Differences were then interpreted in terms of phenology and particularly, of alternation in flowering. For example, Irwin is known to be a regular bearer whereas José is an alternate bearer. Our results showed that this was mainly related to their relative patch sizes: Irwin had larger patches than José.

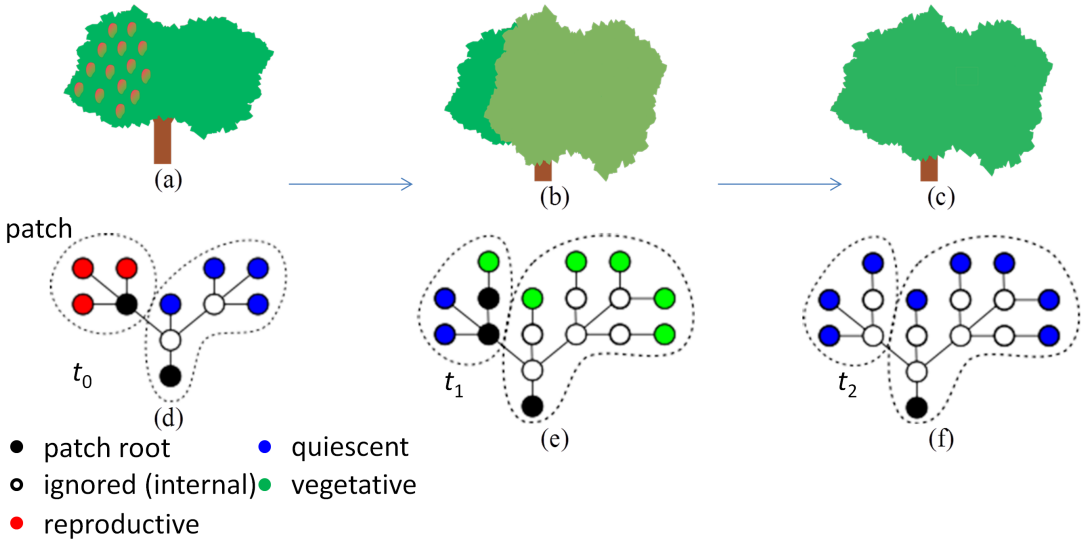

Figure 18: Schematic representation of a mango tree canopy at different flushes and its quotienting. (a) The canopy observed after the intermediate flush of the first growth cycle contains fruits in the left part and GUs that burst in a previous flush in the right part. (b) The canopy observed after the late flush of the first growth cycle contains GUs that burst during a previous cycle in the left part and new GUs in the right part. (c) The canopy observed after the early flush of the second growth cycle only contains GUs that burst during a previous flush. (d) (resp. (e) and (f)) Tree-indexed data that represents the sketched mango tree (a) (resp. (b) and (c)). Black vertices represent roots of homogeneous subtrees found using multiple change-points models. White vertices represent unlabelled vertices associated with past flushes. Red vertices represent GUs bearing fruits during the flush. Green vertices represent vegetative GUs that burst during the flush. Blue vertices represent terminal GUs that burst in a previous flush. An intermediate scale is represented by dashed dark lines grouping GUs belonging to the same quotient.

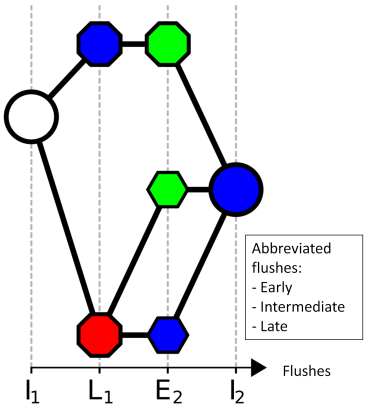

Figure 19: Directed acyclic graph (DAG) representation of patch successions over time for the sketched mango trees in Figure 18. Vertices represent patches and are located on the x-axis according to their flush. Since all temporal edges point from left to right, edge direction is not represented to simplify the drawing. The white vertex is the unlabelled root vertex. Red (resp. green and blue) vertices represent reproductive (resp. vegetative and quiescent) patches. Large vertices correspond to tree, medium to scaffold and small to growth cycle scales.

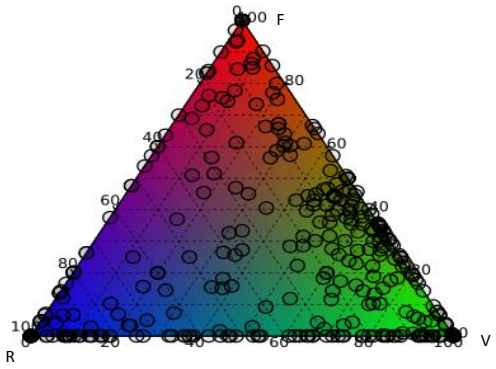

Figure 20: Results of the tree quotienting step: Ternary plot of the subtrees obtained by the multiple change-point model. Each subtree is denoted by a circle. Subtrees at the left-bottom corner are pure quiescent, those at the right-bottom corner are pure vegetative and those at the top corner are pure reproductive. Therefore, the nearer trees are to corners of the triangle, the purer they are. By contrast, patches close to edges have a very low proportion of the label represented at the opposite corner. Background colours are RGB representations associated to the proportions of each label (red for reproductive, green for vegetative and blue for quiescent). This gradient emphasizes patch compositions.

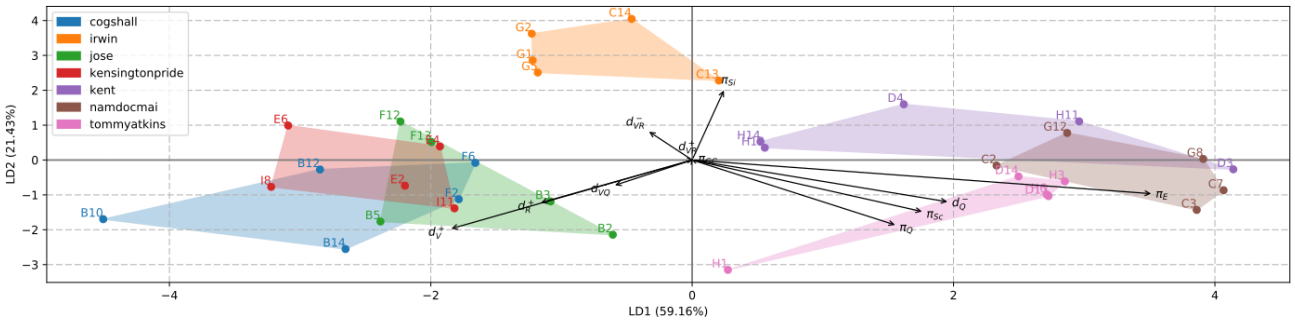

Figure 21: Representation of mango tree individuals in the first LDA plane, using cultivars as classes.

Another approach for assessing synchronism in growth. From a statistical point of view, our tree-quotienting/subtree-clustering approach presented above is based on an assumption of independence and patch-wise homogeneity between shoot labels. The independence assumption does not seem realistic *a priori*, since for example fruiting usually has an inhibitory effect on fruiting the year after along the same axis. If all terminal shoots produced both vegetative and reproductive shoots in the same proportions and synchronously, i.e. at the same burst dates, all branching systems would grow synchronously and would have the same quantitative and qualitative distributions of shoot properties (in terms of quiescence, vegetative and reproductive growth), referred to as *fate* in what follows. This is not compliant with measurements, highlighting patchiness patterns. The latter thus result from mutual exclusions, at the local scale of children shoots of a given mother shoot, between some of their burst dates and / or some of their fates. Such exclusions are observed, for example, when two different children shoot fates cannot occur from a same parent shoot with some given fate. Our model confirmed and quantified such dependencies, which not only exist between child and parent shoots but also between children shoots in the case of mango trees. These dependencies were already suggested in Dambreville *et al.* (2013a) through regression models.

As a complement to change-point detection, we thus developed a Markov-out-tree-based approach aiming at explaining patchiness trough local dependencies inducing synchronism and alternation phenomena. It is fully explained in Dambreville *et al.* (2013b). From a phenological point of view, the question is the ability of local synchronism and exclusion patterns to account or not for global patchiness patterns occurring at more integrated scales (whole tree or scaffold).

We focused especially on differences on fates and burst dates between children GUs issued from a same parent GU to define the model state-space. The latter was defined as the space product of fates, dates within growth cycle (flushes) and delays, from which non-existent combinations were removed. Let recall the possible flushes: E(arly), I(ntermediate) and L(ate). The considered fates were V(egetative), (T)erminal flowering and L(ateral flowering). The delays were either I(mmediate), meaning that current GU was grown in the same flush as its parent, or (D)elayed. Time U was a particular case of unknown date of burst for old GUs that were grown before the beginning of measurements. Thirteen states were defined for GUs as

follows: U-V, IE-V, IL-V, DE-V, DI-V, DL-V, U-F, II-T, IL-T, DI-T, DL-T, II-L and DI-L, using some Delay-Flush-Fate order to name states (e.g., IL-V means Immediate Late Vegetative). Thus using the methodology described in Section 3, thirteen PGMs were identified, each one associated with a given parent state.

To illustrate an example of results and conclusions that can be drawn from estimated PGMs, we develop the case of state II-L as a parent GU (see Figure 22).

- No transition occurred from parent state II-L to children states U-V, U-F, II-V, II-T, IL-V, IL-T, DE-V, nor II-L. This was expected for states U-V and U-F, which by definition always preceded other GUs.

- The edges originating from source vertices DL-V and DL-T and pointing toward non-source vertex DI-V with associated negative regression parameters expressed mutual exclusion between DI-V on the one hand, and {DL-V, DL-T} on the other hand. The same mutual exclusion behaviour occurred between states DL-V and DI-L. This highlights that immediate GUs from flush I with lateral inflorescences (state II-L) cannot have children GUs, the year after, successively at flushes I and then L. Hence state II-L, as a local context regarding parent GUs, is favourable to synchronism. However, this also suggests that on the one hand, children GU fates may be heterogeneous (simultaneous occurrence of V and T children). If these children, in turn, tend to propagate their fates to their own children, this could lead to patches from the viewpoint of fates. On the other hand, DI-T, DL-T and DL-V may coexist, highlighting absence of strict exclusion pattern regarding GU production at both I and L flushes.

These results showed the ability of Markov out-tree (MOT) models to identify in which contexts a given parent GU can or cannot have children GUs at different flushes or in different fates. This can be interpreted as the mechanism entailing patchiness at GU scale, combined with propagation processes favouring its emergence at coarser scales. This local point of view on asynchronism can be turned into a more integrated view by prediction, using our model, of the total number of descendant GUs at each flush and each fate at different scales (e.g., scaffold, whole tree). It is in this spirit that the model was included into a simulation scheme by Boudon *et al.* (2017).

Perspectives. In summary, we designed two complementary methods to quantify and characterize patchiness. Their outputs can be used to compare cultivars and as a perspective, could be integrated into varietal selection procedures or technical arrangement studies. The segmentation heuristic does not require particular assumptions concerning observation distributions. This approach could therefore be used for detecting patchiness resulting from the observation of numerous variables of different types. It could also be applied to other temperate or tropical plant species.

Particularly, a next step in the mango tree setting would be to analyse absolute patch sizes, which are related to amounts of carbon reserves and to distances between GUs. It is thus likely that the absolute patch size would summarize accurately cultivar fruiting patterns and their agronomic behaviours.

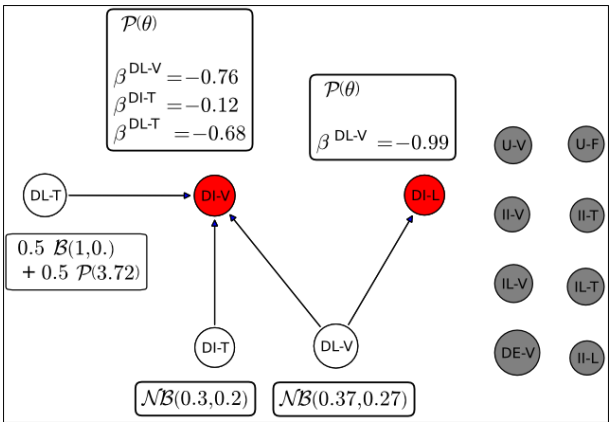

Figure 22: PGM associated with parent state II-L. Vertices of the PGM correspond to the random numbers of children GUs in each state. Grey children states correspond to absence of children in those states (deterministic). White vertices (sources) and parameters correspond to univariate distributions ($\mathcal{P}$: Poisson, $\mathcal{B}$: Binomial, $\mathcal{NB}$: Negative Binomial, except for vertex DL-T associated with a mixture of $\mathcal{B}$ and $\mathcal{P}$ with weights 0.5). Red vertices (sinks) correspond to univariate regression models. Regression parameters associated to variable i in PGM are denoted by β^i.

This approach also offers new perspectives for testing causal assumptions on patchiness, as for example the effect of fruit numbers in patches on the nature of subsequent patches, or more generally the effects of phenological or environmental factors on patch development.

One remaining question is related to sufficiency of local dependencies to account for patchiness patterns occurring at more integrated scales. To address this issue, we plan to simulate plant growth under our MOT and some null model (e.g., MOT with independent children states) and compare the empirical distribution of patchiness indices with the values obtained from the data set. An additional perspective of improvement would be to resort to new segmentation methods on trees developed by Thepaut and Rigaill (2019).

5.2.4. Reconstruction from laser scanner

In this section, the different data sets subject to statistical analysis were essentially collected in more or less handcrafted ways, involving in particular expert knowledge on how to segment plants into elementary components. For example in apple trees, identifying growth units or annual shoots requires some ability in grasping characteristic scars in axes, which can be separated from one another by less than 4 mm. In every case, acquiring tree topologies required some human intervention to record each entity into a data base.

A significant improvement towards automation in data acquisition has been offered by laser scanners. Their outputs are unstructured clouds of points, from which topology and geometry of plants have to be reconstructed. Except in controlled environments (particularly, orchards)

in which data acquisition is achieved during a defoliation period, additional steps have to be performed: segmentation of individuals, discrimination between wood and leaves.

The work by Preuksakarn *et al.* (2010) described hereafter focused on reconstruction in isolated, defoliated plants. This already is a complex problem due to occlusions and low point densities, hence difficulties in reconstructing small branches, which often are attached to inappropriate parts of the plan structure. The aim of our work was to improve robustness of state-of-the-art algorithms with respect to occlusions and coarse sampling of small branches by incorporating some statistical modelling in the analysis. At the time of publication, plant reconstruction was mainly addressed through so-called *procedural methods*, which relied on iteratively simulating plant growth from a given starting point, considering as attractors the points that were not included in the structure yet, at current iteration. The approach suffered from the following shortcoming: this involved a high number of parameters for which no efficient calibration method was available. One possible alternative was sketching, which is a manual rough outline of the shape guiding reconstruction by constraining subtrees to be delimited by given spatial zones; however, the reconstruction error remained high.

Later, some methods based on graph reconstruction were proposed. The addition and removal of edges was based on distance criteria. A skeleton was then built from the centres of segments determined from the graph, using either a K-means algorithm or bounds of distances to root vertex. However, heuristics used to handle occlusions and low point densities lacked of robustness, yielding high error rates.

Moreover, both kind of approaches lacked of principled quantitative assessment, due to insufficient methodology and benchmark data sets. The authors' contributions were mainly twofold: on the one hand, to improve robustness of the Space Colonization Algorithm (SCA; Runions *et al.*, 2007) with respect to occlusions and low point densities using automatic adaptation of parameters to local point densities, together with statistical models to determine the number of branches issued from a current axis in a given zone (my personal contribution here). On the other hand, a benchmark data set was collected and various approaches were assessed and compared, based on edit distance computations between tree skeletons.

The main steps towards robustification of SCA were to infer plant topology using a local neighbourhood graph L, obtained by connecting each point to its k closest points. Edges were weighted by distances between vertices. Since this could result into several disconnected components, a reconnection procedure was used, based on recursive search of the smallest edges needed to reconnect parts. Then to take tree topology into account, the distance between points was redefined as the length of the shortest path in L.

At each iteration of SCA, the neighbourhood the tips of current tree axes was considered. Assuming that points were sampled from the tree surface, local axis orientations around vertices were extracted using principal component analysis (PCA) on their neighbours. Normed eigenvectors associated with the first PC, corresponding to main tangential orientation at each point, were computed (Figure 23), oriented with respect to some fixed reference and represented in polar coordinates (Figure 24).

Clustering of these polar orientation values using bivariate Gaussian mixture models provided possible local directions for tree axes. The crucial point was to identify the number of axes contained in a current set of points, corresponding to the number of clusters. This was selected

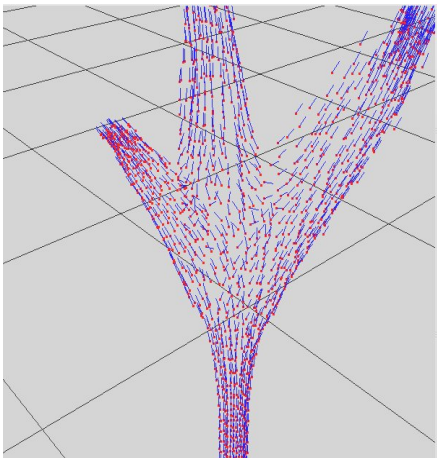

Figure 23: First eigenvectors of PCA associated to each point of a given volume, used in plant reconstruction.

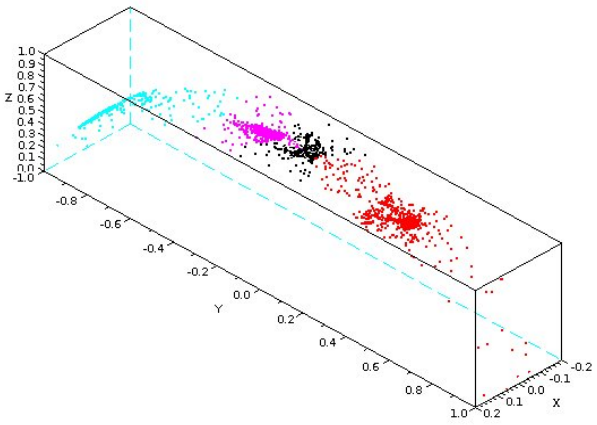

Figure 24: Polar representation of first PCA eigenvectors associated to the points in Figure 23 with the different clusters represented with colours.

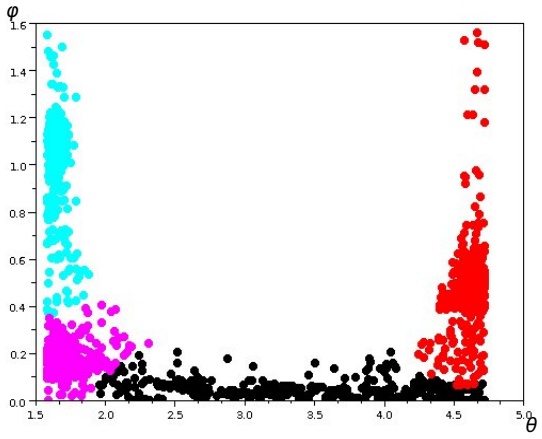

Figure 25: Polar representation of first PCA eigenvectors associated to the points in Figure 23 in the (θ, ϕ) plane (azimuth and elevation) with the different clusters represented by colours.

using the Integrated Classification Likelihood (ICL) of Biernacki *et al.* (2001). Using the selected number of clusters, the mixture parameters were estimated with the EM algorithm and the points were clustered into the (θ, ϕ) plane (azimuth and elevation, see Figure 25). Then the clustered points were mapped towards the original 3D representation, thus associating each point to a tree axis, each of them corresponding to a cluster (Figure 26). Their directions were then used to constraint SCA to use them as candidates for further directions of space colonization.

Perspectives. The presented approach is not related to any tree-scale statistical growth and structure model addressed in this section. However, we can expect that incorporating knowledge from already segmented trees by statistical modelling could improve tree reconstruction. Thus, a novel and integrated approach could rely on hierarchical Bayesian models incorporating different levels of latent variables: species or clusters of structurally similar individuals, trees (to achieve tree segmentation), axes (as vertices from a tree graph) modelled as HMOTs and points distributed as geometric 3D random variables determined by axes (e.g., uniform distributions on cylinders aligned on axes with Gaussian noise).

Such complex models would have to be inferred using approximations. Since on the one hand, inference in each of the building block (in other words, at each underlying scale) of the model is rather well understood and on the other hand, each scale includes latent states whose numbers have to be selected, VBEM-based approximations seem particularly relevant (see also Section 7).

5.2.5. Other contributions

In this subsection, diverse, somewhat minor contributions related to plant science are summarized.

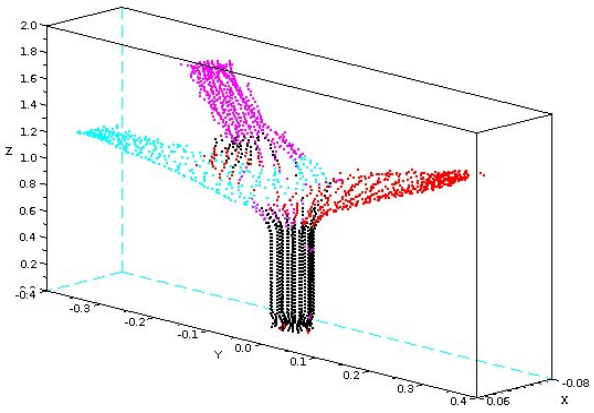

Figure 26: Mapping of the clustering obtained in the (θ, ϕ) plane (Figure 25) to the original 3D space.

Architecture of rose bush. In relation to M. Garbez' industrial PhD thesis, we studied the architecture of rose bushes. The main goal of the study was in a first step, to obtain a quantitative characterization of rose bush architecture, including its phenotypic plasticity in response to growth conditions. In a second step, we aimed at relating such characteristics to plant visual appearance, with the commercial purpose of improving product quality in crops. As bushes, rose trees usually do not have hierarchical structures and growth strategies at whole plant scale; for example the considered cultivars do not necessarily show any trunk. They rather grow branched systems from the plant basis, yielding somewhat equivalent structures, referred to as reiterated complexes. These can be characterized by their degree of similarity with some ideal branched system containing all structural and qualitative properties of shoots potentially expressed at whole tree scale. One *a priori* assumption was the existence of statistical dependencies between consumer preferences on the one hand and the quantity and degree of reiterated complexes on the other hand – which we initially wanted to test. To enhance tree phenotypical variability, plants were cultivated under a shading gradient in three distinct environments: natural conditions, under 55 and 75% shading net. Firstly, CIC-HMT were estimated on the plant material, composed by 20 Rosa hybrida "Radrazz" in each light condition. The objective was to identify clusters of axes, which defined an exhaustive catalogue (given available growth conditions) of axes used to qualify and eventually, to compare branched systems and identify degrees of reiteration. A typical segmentation of a rose bush illustrating the segmentation into axes with contrasted quantitative and structural properties is illustrated in Figure 27. The considered variables for axes were length, diameter, number of GUs, flowering status, number of children axes and stiffness coefficient. The transition diagram is represented in Figure 28. The model suggested that reiterated complexes typically have their first axes in state 1 (represented in red).

The results published in Garbez *et al.* (2018) are related to quantitative characterization of architectural development of rose bushes over time and prediction of sensory attributes, rep-

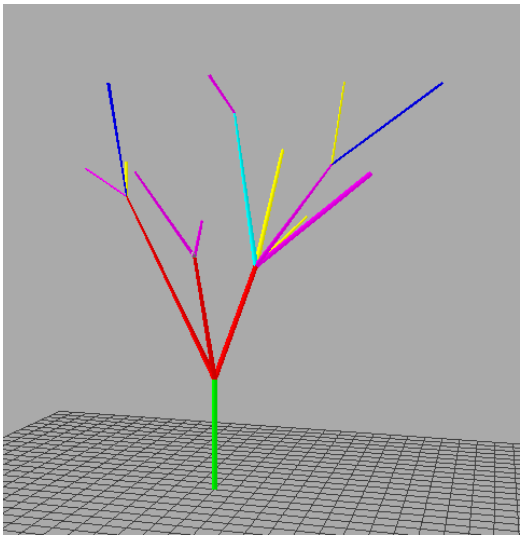

Figure 27: Segmentation of rose bush (no shading) using a CIC-HMT model. Colours represent restored states.

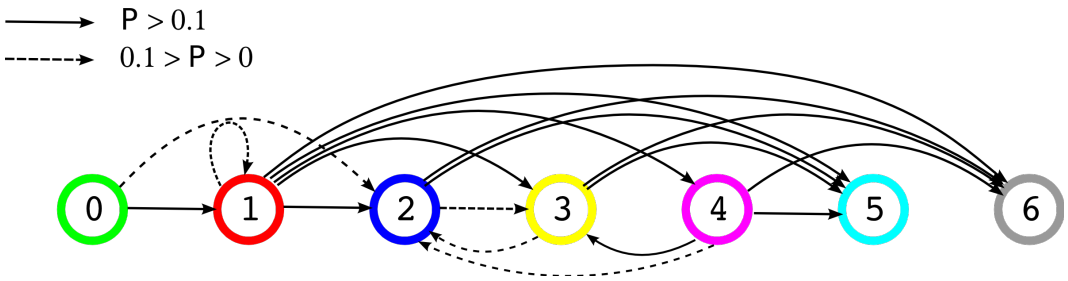

Figure 28: Transition diagram of the CIC-HMT model on rose bush (no shading). Arcs with probabilities less than 0.1 are omitted.

resenting multiple visual traits of the plants. They did not include identification of reiterated complexes with CIC-HMT models. Strong correlations were found between them and architectural variables, extracted from phytomer- to plant-scale data. Acceptable to very satisfying predictions were obtained with ordinary least squares regression and variable transformations to encompass non-linear relationships.

The proposed approach thus seemed a promising way to gain a better insight into the architecture of shrub plants, together with their visual appearance. It opened new avenues to target processes of interest in order to optimize growth conditions or select the most fitting genotypes across breeding programs, with respect to contrasted consumer preferences.

As a perspective, reiterated complexes could be identified by combining the states obtained with CIC-HMT models and comparisons with edit distances (as developed in Subsection 2.2). The CIC-HMT could incorporate covariates representing levels of shading; moreover, their effect on plant architecture could be assessed, particularly on the quantity and degree of reiterated complexes. This approach of identifying reiterated complexes with similar structural properties and local attributes is somehow specific to the problem. Thus, more generic methods for multiscale clustering of tree structures would have to be developed (see perspectives in Subsection 2.2.2).

Cell divisions. In shoot apical meristems (SAMs), organogenesis results from divisions and differentiation of cells whose identities are not predetermined, thus allowing different organs (leaves, flowers, stems, etc.) to be produced by a same meristem. While embryonic organs begin to form, their cells progressively get definite identities, which may be transferred to their descendants through divisions. Recent advances in imagery allowed researchers to acquire temporal sequences of 3D meristem scans with coarse time steps. Between acquisitions new divisions occur, so that specific methods were developed in the Inria Virtual Plants team to segment cells, estimate their geometric properties and reconstruct lineages by inferring unobserved divisions. One acquisition is represented in Figure 29. The meristem was manually segmented in this figure, delimiting several primordia with different stages of development (green: early stage, red: later stage), the central dome (orange), the central zone (cyan) composed by pluripotent cells and boundary cells (dark blue). A more detailed description of the spatial structure of SAMs is provided by Fletcher (2002).

The aim of our study, as a part of P. Fernique's PhD thesis (2014a), was to infer underlying cell identities (during early stages of flower development) through clustering. As a complement, its purpose was to highlight interactions between division rates and cell identities.

The relationships between an initial set of cells and their descendants can be represented as a forest. The data set was modelled using unordered hidden Markov out-trees, where latent states correspond to cell identities. Since the number of children was one or two and the number of possible states K remained low, saturated (so-called non-parametric) models were considered for the generation distributions $p(\boldsymbol{N}_v|S_v = k)$, S_v denoting the latent state for cell v and $\boldsymbol{N}_v$ the state vector for its descendants issued from divisions between two time steps.

The observed geometrical properties X included in the model were cell volume, epidermal surface, external surface, principal and secondary curvatures. These were assumed to be independent given their identity (hidden state S) and modelled as univariate, Gaussian-(curvatures)

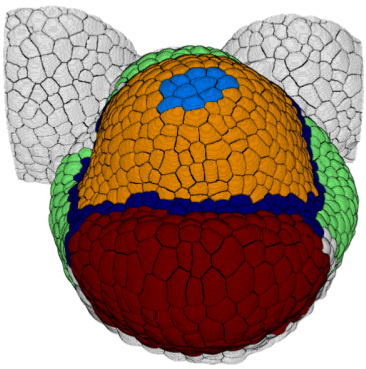

Figure 29: Spatial regions of floral meristems. The most developed sepals correspond to red cells; the latest ones to green cells. Orange cells correspond to the central dome and light blue ones to the central zone. Boundary cells are in dark blue.

or Gamma-distributed (other variables). The number of states was selected using BIC, yielding here a four-state model. HMOT models with conditionally dependent ($\mathcal{M}_1$) vs. independent ($\mathcal{M}_0$) children states were also compared using BIC. Both estimated models had rather close numbers of non-zero parameters: 22 for $\mathcal{M}_1$ against 18 for $\mathcal{M}_O$. BIC was higher for $\mathcal{M}_1$ (-18164) than for $\mathcal{M}_0$ (-18668), indicating a clear benefit of taking into account children state dependencies in cell divisions.

Cell epidermal surface, internal surface, curvatures and (especially) volume were assessed to be structuring variables in this model, since the estimated observation distributions for the different states were well separated for this five characteristics (see Figure 30).

These observation distributions allowed us to partly characterize the four states:

States 0 **and** 3 correspond to large cells and are mostly differentiated by their curvatures (both negative for state 0 and positive for state 3), state 0 corresponding to the largest cells.

State 1 corresponds to small cells with both curvatures almost of the same norm and mostly negative, this being typical characteristics of saddle forms.

State 2 is in-between considering size, but with clearly positive curvatures, corresponding to the dome area.

Using the Viterbi restoration algorithm (Figure 31), spatial regions that emerged from cell identity labelling were interpreted using meristem morphology: central dome and zone were assigned to state 2, sepal primordia were composed by states 0 and 3, while boundary zone was assigned to state 1.

State 2 is the main state of the first time point and presents a high spatio-temporal coherence from 0h to 69h. Despite an early stage of meristem differentiation at 0h, few cells are already assigned to putative sepals. At subsequent time points, the multiplication of sepal and dome cells with the apparition of boundary cells, delimiting the frontier between sepals and the dome zone, is observed. The apparition of boundary cells is unobtrusive until 44h but significant as the continuous border is clearly identified starting from 56h.

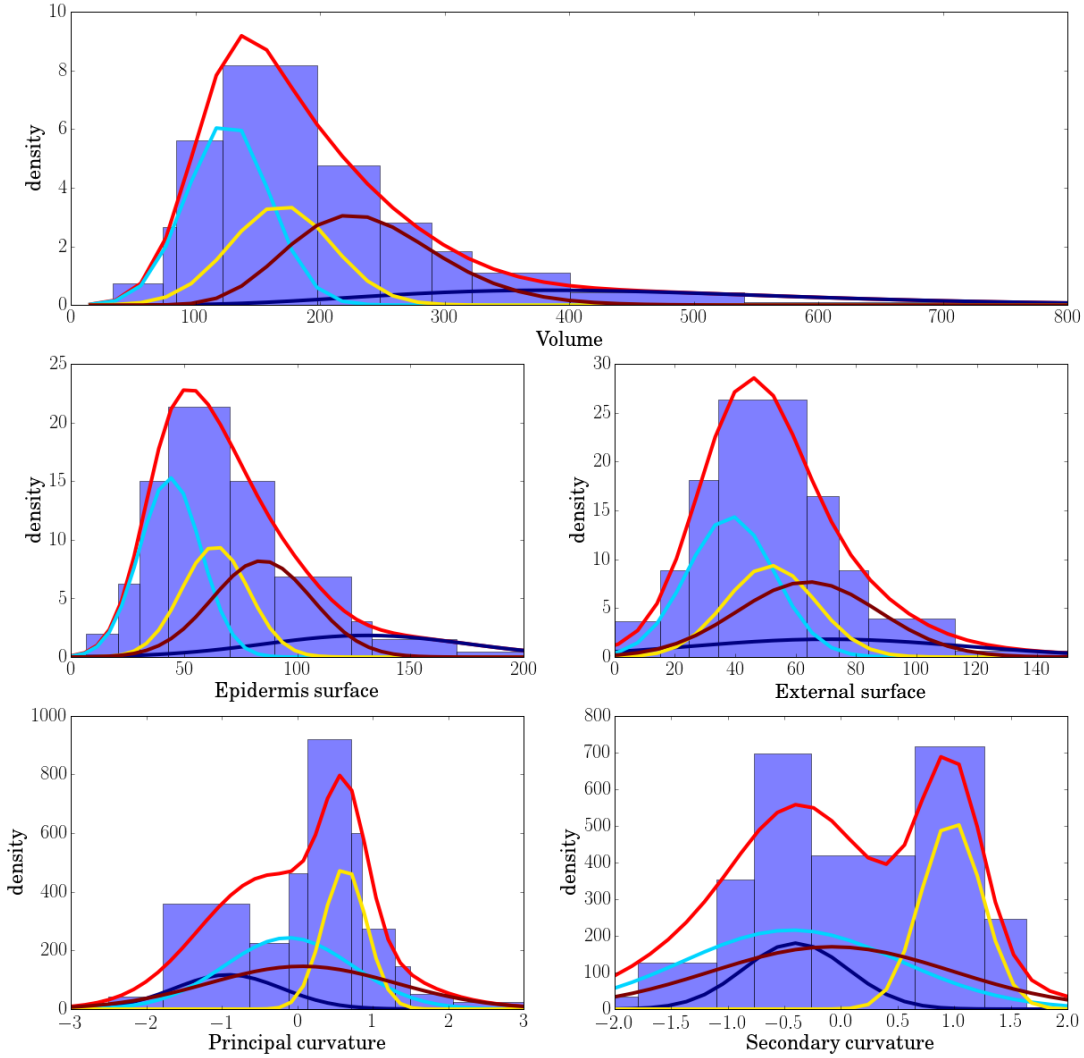

Figure 30: Histogram (given restored states), emission distributions and mixtures (bright red) for each observed variable. State 0 is in dark blue, state 1 in light blue, state 2 in yellow and state 3 in dark red. Surfaces and volumes are modeled by Gamma distributions and curvatures by Gaussian distribution. Combining state separations induced by surfaces and volume in the one hand and curvatures in the other hand indicates that these characteristics are sufficient for state discrimination.

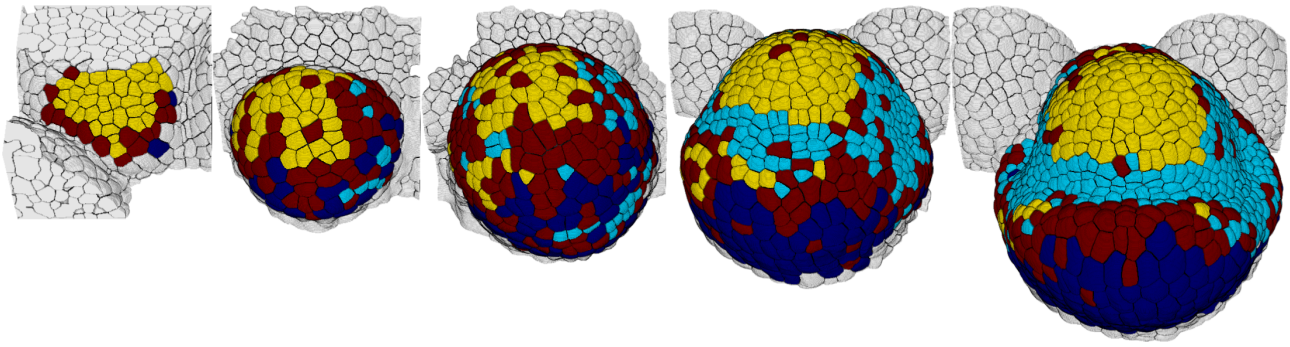

Figure 31: Viterbi hidden state restoration for Hidden Markov Out-Tree (HMOT) models on cell divisions. Images, from left to right, were taken at 0h, 26h, 44h, 56h and 69h after the beginning of the experiment. Segmentation is obtained by spatial projection of the four states obtained using for the HMOT estimated from cell epidermis surface, internal surface, volume, curvatures and inertia as variables. State 0 is in dark blue, state 1 in light blue, state 2 in yellow and state 3 in dark red. Primordia are mostly identified by considering state 0 and 3, the dome by state 2 and boundary cells by state 1.

State interpretation can be completed by the analysis of the estimated generation distributions denoting as $\Gamma_s(n_0, n_1, n_2, n_3)$ for $s \in \{0, 1, 2, 3\}$ (probability of a parent cell in state s having jointly n_k children cells in state k):

$$\Gamma_0(0,0,0,2) = 0.07, \quad \Gamma_1(0,0,0,1) = 0.13,$$
$$\Gamma_0(1,0,0,0) = 0.29, \quad \Gamma_1(0,1,0,0) = 0.45,$$
$$\Gamma_0(1,0,0,1) = 0.42, \quad \Gamma_1(0,2,0,0) = 0.35.$$
$$\Gamma_0(2,0,0,0) = 0.20.$$

$$\Gamma_2(0,0,0,1) = 0.13, \quad \Gamma_3(0,0,0,1) = 0.14,$$
$$\Gamma_2(0,0,0,2) = 0.10, \quad \Gamma_3(0,0,0,2) = 0.09,$$
$$\Gamma_2(0,0,1,0) = 0.18, \quad \Gamma_3(0,0,1,1) = 0.11,$$
$$\Gamma_2(0,0,1,1) = 0.28, \quad \Gamma_3(0,1,0,1) = 0.35,$$
$$\Gamma_2(0,0,2,0) = 0.31. \quad \Gamma_3(0,2,0,0) = 0.17,$$
$$\Gamma_3(1,0,0,0) = 0.05.$$

The reproduction and emergence of cell identities underlined by generation distributions are consistent with biological beliefs. State 3 is a hub for transitions from state 2 at 0h to other states at times greater than 44h. This is the state with highest division rate. Cells in state 3 cannot stay in that state until the next time step: they either divide or switch to state 0, which corresponds to differentiation from early cells to late ones in primordia. Divisions of state-3 cells may mainly yield one cell in the same state and a second cell in state 1 or 2. Transitions from

state 3 to state 1 correspond to emergence of boundary cells induced by primordia formation, which may be more the consequence of their differentiation than an active phenomenon.

Despite these first conclusions, biological interpretations drawn from the outputs of the HMOT model were limited by the number of available successive time points and data quality. Indeed, time intervals between successive images were large, resulting into some divisions not being observed. Such missing divisions were interpolated but this resulted into the presence of a large number of predicted cells without observed characteristics (almost 50%). In addition, the number of time points (five here) also limited detailed investigation of cell division patterns. Now new protocols would be available, allowing researchers to acquire more time points (up to 15), with reduced time intervals between successive acquisitions and raw images of better quality. This would help us to obtain more accurate segmentations and thus, measurements of cell characteristics. There is also a systematic bias in algorithms in estimating cell characteristics. This is in particular true for curvatures computed with some non-adaptive procedure, which could be replaced by the adaptive algorithm by Tong and Tang (2005).

Geometry of meristems of *Acacia mangium*. This study issued from Hatt *et al.* (2012) focused on comparing morphological and histocytological characteristics of *Acacia mangium* SAMs with respect to growth conditions (natural vs. *in vitro*) and to heteroblasty (change in form between juvenile and mature individuals). The main biologic conclusions are SAMs in natural environment being much bigger and containing more cells with larger vacuolated area for mature than juvenile type. *In vitro*, where reversions from mature to juvenile morphological traits do occur unpredictably, heteroblasty was less obvious in SAM examined characteristics. *In vitro* SAMs corresponding to the juvenile and mature types showed similarities with outdoor juvenile SAMs, but could be distinguished from these latter by a much larger vacuome, which might be induced by culture conditions.

From a methodological point of view, the considered morphological SAM characteristics were height H, diameter D and a shape index S characterizing the bulged aspect of the dome based on fitting the following superellipse equation with SAM points:

$$\left(\frac{2x}{D}\right)^S + \left(\frac{y}{H}\right)^S = 1.$$

Our contribution was on the one hand, to assess shape heterogeneity within given growth conditions by fitting mixture models to the triplets $x_i = (D_i, H_i, S_i)$ for each SAM i. A BIC-based model selection procedure gave rise to a partition of SAMs into two clusters, primarily determined by H and secondarily by S. None of these two clusters could be clearly associated with a specific SAM origin, despite cluster 1 containing mostly mature SAMs and having higher H values than SAMs from cluster 2.

On the other hand, as an alternative to multivariate mean comparisons based on multivariate Gaussian assumptions (MANOVA; Tabachnick and Fidell, 2007), original permutation tests (Good, 2005) were developed for assessing the effect of the treatments on the triplet components simultaneously. These are based on the ratio of inertia defined as follows: Let N be the number of individuals and N_k the number of individuals with a given class k, where classes are defined as the space product of 2×2 conditions. Let m_k be the mean vector for individuals with class

k and m the mean vector for all individuals (regardless of class). Our statistic was defined as $\delta = \text{ICI}/\text{TI}$, where

$$\text{ICI} = \frac{1}{N}\sum_k N_k \|m_k - m\|^2$$

is the inter-class inertia (representing the dispersion between mean vectors induced by different SAM classes) and

$$\text{TI} = \frac{1}{N}\sum_i \|x_i - m\|^2$$

is the total inertia, representing the total dispersion between all individuals.

The ratio of inertia is a generalization of the test statistic of ANOVA, used in both factorial discriminant analysis and multivariate analysis of variance (MANOVA), to quantify the class separation in multivariate settings. Values of δ far from 0 indicate a good separation between classes.

Randomized tests were implemented as follows:

1. Compute the value δ_{ref} of δ for the true dataset (individuals with non-random classes).
2. Assign random classes to individuals, such that the number of individuals of each class is preserved (assignment does not depend on the x_i's).
3. Compute the ratio of inertia δ using the classes obtained at step 2.

Random permutations of the individuals were used for assigning random classes to individuals in adequate proportions. Steps 2 and 3 were repeated n_r times (with $n_r = 5,000$). The proportion P of values $\delta_{\text{ref}} < \delta$ was computed and interpreted as usual p-values in hypothesis testing.

As a perspective, the power of this test under different sample sizes and distributional assumptions (including the comparison with MANOVA) would have to be investigated, particularly through simulation studies.

6. Software contributions

The methods presented in the previous sections were included into mainly two pieces of software dedicated to hidden Markov models and statistical analysis of plant structures.

6.1. Chainxem: a Matlab library for HMC/HMT analysis

Chainxem is a Matlab library dedicated to inference in HMC and HMT models. Originally I developed the library during my PhD thesis as an extension of the ancestor of Mixmod[8], to relax an independence assumption in mixture models and consider time or tree dependencies.

During L. Donini's PhD thesis (Subsection 4.2), the library was extended by the Xerox company to include connexions with timeout optimization and reinforcement learning models.

Then extensions to categorical and multivariate observations were motivated by new research work on eye-movement analysis, as a precursory implementation of the models developed later in B. Olivier's PhD thesis.

The library now contains about 10,000 lines of code. It is available at `http://mistis.inrialpes.fr/people/jbdurand/software.html#chainxem_en`.

6.2. Tree Statistic: statistical models on trees for plant structure analysis

Tree Statistic is a python module dedicated to the statistical analysis of plant structure data. This is a part of the Structure Analysis[9] component of the OpenAlea[10] project. OpenAlea is an open source platform aiming at connecting different models from the plant research community. It includes modules to analyse and model plant architecture, growth and functioning. The principle is to provide some software architecture to compile libraries developed with different programming languages and to make them available in python, through wrappers if they are not native python libraries.

Tree Statistic is an extension of the Sequence Analysis module, dedicated to the analysis of biological sequences. It mainly contains Markov-switching models, including hidden semi-Markov or variable-order Markov models (including Generalized Linear Mixed Models as emission distributions) and multiple change-point detection. It is more specifically oriented towards discrete observations (nominal-, ordinal-, integer-valued) but also includes some continuous observation distributions. The module was used and extended during B. Olivier's PhD thesis, in relation with hidden Semi-Markov models (see Subsection 4.3). The core of the library is composed by C++ classes with python wrappers implemented through Boost.Python.

Tree Statistic relies on the same base classes (histograms, discrete and continuous distributions) and aims at providing statistical models for the analysis of tree-structured data. It provides data structures and (hidden) Markov models on trees, as presented in Sections 2 and 5. As explained in these sections, general dependency assumptions within trees rely on modelling

[8] `http://www.mixmod.org/?lang=en`
[9] `https://github.com/openalea/StructureAnalysis`
[10] `https://github.com/openalea`

multivariate count data and identifying probabilistic graphical models, also included as components of the library. Since my post-doctoral fellowship at CIRAD/AMAP, I have been the maintainer and main developer of the library. During his PhD studies, P. Fernique contributed significantly by adding graphical models and parametric multivariate count distributions. The library now contains about 50,000 lines of code, among which 75% is C++, 15% is native python and the remainder is Boost.Python. The library is interfaced with other Structure Analysis components: sequence analysis, multiscale tree graphs (Godin and Caraglio, 1998) and PlantGL. The latter is a graphical toolkit for the creation, simulation and analysis of 3D virtual plants and was used to produce the 3D figures of this manuscript (e.g., Figures 23 and 27 in Section 5). We plan to unify data structures with the Tree Matching library dedicated to computation of edit distances and mappings between tree graphs (see Subsection 2.2) and make them a coherent "Tree Analysis" component of Structure Analysis.

7. Conclusion and perspectives

As a conclusion, our main contribution is the specification, implementation and development of numerical methods for hidden Markov models on sequences and tree structures, particularly for the analysis of plant growth and architecture. This led us to develop new tools and selection methods in graphical probabilistic and multivariate count modelling. Motivated by specific contexts of applications, new families of hidden Markov chains were also proposed: coupling HMCs with decisions processes and optimization in the context of optimal timeout modelling, coupling HSMCs in the case of heterogeneous, asynchronous signals with application to joint EEG and eye-movement analysis. These models try to address generic questions and to go beyond the original motivation justifying their development. Plant structure analysis was, in a first step, mainly driven by cognitive motivations: can we characterize plant structures through quantitative models? How well do they predict tree shapes, their variability and the effect of environmental or other factors? Our hidden Markov tree models provided partial answers to the question: They often had to be complemented by other approaches, such as edit distance computation, tree matching, supervised and unsupervised machine learning at different scales, whenever thorough model integration seemed hardly possible. In such contexts, workflows of treatments were applied, often not accounting for the uncertainty or variability underlying the models at previous steps of the analysis.

We originally analysed data sets that were rather sparse and collected manually. However some years later, the automatic collection of data allowed plant researchers to access bigger collections of structured measurements, the structure often being accessible only indirectly. Thus, a need for reconstruction methods emerged and we participated (modestly) to this new challenge. As a corollary, measurements of structured data on whole progenies were accessible. This raised the question of accounting for new sources of variability (now genetic) with their potential interactions with ontogeny and environment. Then, how to assess their impact on complex structures (tree graph, multiscale tree or more general graphs)? We addressed the problem by incorporating mixed effects in models acting at local scales, with the assumption that global properties would essentially emerge from local interactions included in the models.

Although perspectives are provided separately in each subsection of the manuscript, we would like to present ongoing work and more global future research avenues. Some research partners focused on ecological questions, now considering larger spatial scales. Instead of isolated individuals or forest / orchad plots, tropical forests are the object of interest, with the aim of monitoring their evolution, particularly in relation with global warming, from the points of view of composition, leaf area density and production (covering both storage and exchange with atmosphere). Data acquisition is achieved by satellite imaging, terrestrial and aerial LiDAR. In such environments, data are characterized by much more sources of heterogeneity, primarily related to devices, climate conditions at various time scales – particularly both the day of measurement (rain, wind) and at coarser time scales, thus impacting tree growth globally – soil and forest composition or other semi-local environmental features, stage of plant development, repeated measurements on a same vs. different individuals and interception of laser beams by leaves vs. wood. From a statistical point of view, these sources may be viewed as many latent

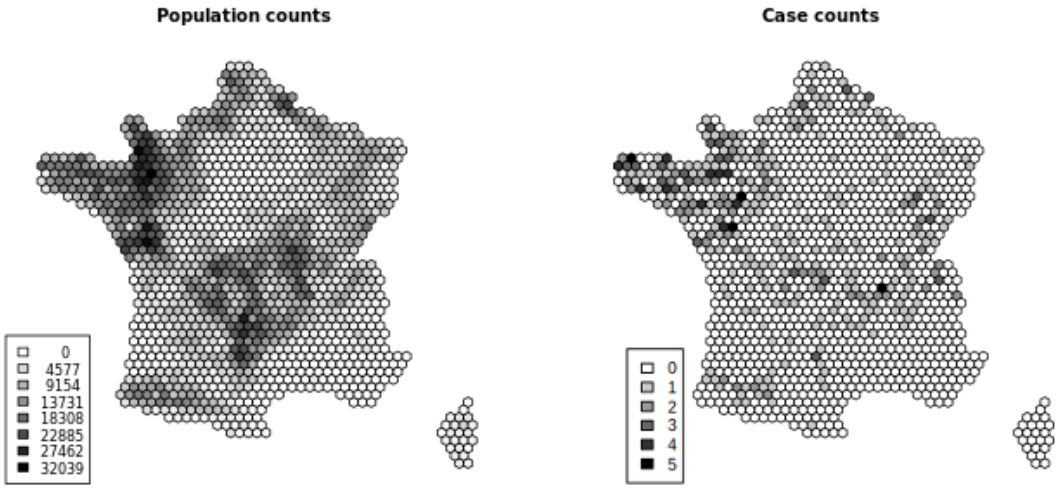

Figure 32: Total cow population (left-hand part) and BSE (right-hand part) counts in the different districts. Levels of grey correspond to intervals whose bounds are indicated in legends, e.g. white corresponds to 0 to 4576 cows.

variables subject to some hierarchy through spatial and time scales, from plant elementary components to forests.

From a statistical point of view, new topics were recently investigated in my research team, particularly regarding hidden Markov random fields (HMRFs) and Bayesian non-parametric (BNP) analysis. Contributions from both fields joined when considering image segmentation. HMRFs are a popular approach (Kato and Zerubia, 2012) but suffer from the high complexity of model selection, which determines the unknown number of segments. HMRFs are state-space models where categorical latent state variables take their values in $\{1, \ldots, K\}$, K being an unknown number of states. In classical settings, different models have to be estimated separately for each considered value of K. Then these models have to be compared using statistical criteria. One possibility to avoid successive estimation and assessment of each model is offered by reversible jump MCMC methods (Kato, 2008), in which the very high dimension of parameter and latent variable space can make convergence critically slow. Chatzis and Tsechpenakis (2010) and later, Lü *et al.* (2019) proposed a Bayesian non-parametric prior on the MRF, thus allowing *a priori* unbounded values of K. Inference was addressed by a variational Bayesian EM (VBEM) algorithm, which is less CPU-demanding than MCMC. The emission distributions were Gaussian. During F. Dama's masters internship, we developed an extension to discrete observation processes (modelled in a first step as Poisson) and are now considering further extensions to zero-inflated or over- and underdispersed data. Firstly, we applied this model to disease mapping in cases of bovine spongiform encephalopathy (BSE) in France. The data set is depicted in Figure 32. The associated segmentation and estimated emission parameters are represented in Figure 33.

We now believe that forest monitoring and reconstruction from LiDAR data could be addressed in a principled way using an integrated Bayesian framework. The global model would

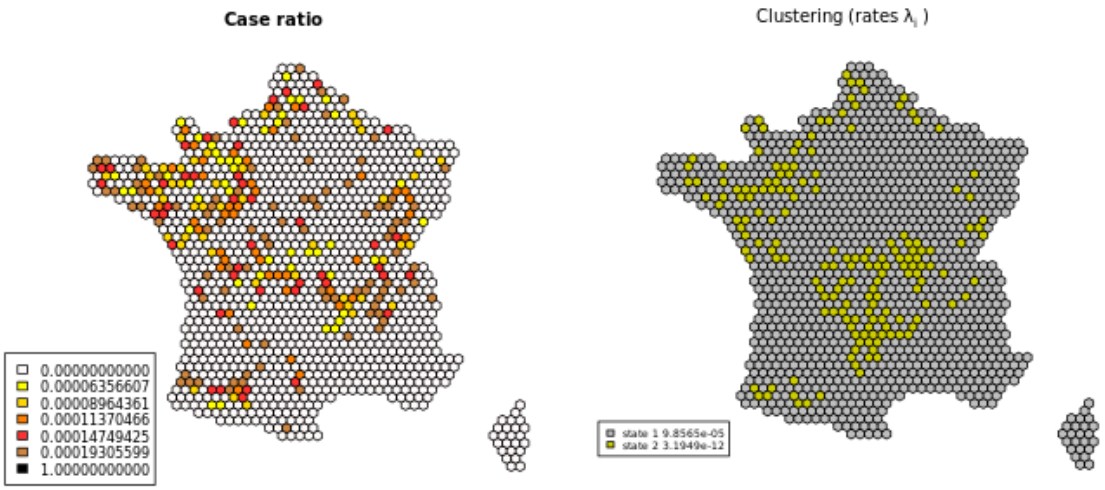

Figure 33: BSE case ratio (left-hand part) and BNP-MRF segmentation with MAP (right-hand part). For BSE case ratio, colours correspond to intervals whose bounds are indicated in legend, e.g. white corresponds to 0 to 6.4 cases per 10^5 cows.

include a hierarchy of latent variables and model blocks associated with the different scales of representation (plant elementary components, axes, individuals, species or ecological clusters, homogeneous forest plots, whole forest), including available covariates. For example, the model block associated with homogeneous forest plots could be a BNP-MRF with multivariate count distributions (e.g., multinomial distributions if homogeneous plots are defined by similar species proportions, or graphical discrete models as in Section 3 if changes in dependencies and total numbers of trees are considered as discriminant). To make the approach feasible from the viewpoint of computing time and memory, the data set could be processed in partially redundant slices, accounting for conditional independence relationships: given the cluster value representing homogeneous growth conditions, individuals are assumed to be independent, etc. However, since observations are essentially one cloud of points, slices would have to be overlapping to ensure that each individual is fully contained within at least a slice. Once individuals are segmented, points are not relevant any more at coarser scales and can be freed in memory.

In a different but related context, we aim at estimating leaf area densities (LADs) from similar measurements, except that they consists in the number of hits of laser beams with leaves, together with segment lengths between hits (or cases of non-interception), referred to as free path lengths, instead of 3D points.

Let (θ, Φ) be respectively the elevation and azimuth components of the beam incidence direction. On the one hand and under classical assumptions, LAD is related to an attenuation coefficient $k(\theta, \Phi)$, a clumping factor Ω and a ratio $G(\theta, \Phi)$ of foliage area projected in direction (θ, Φ) to actual area by the following equation:

$$k(\theta, \Phi) = G(\theta, \Phi).\Omega.\text{LAD}.$$

On the other hand, $k(\theta, \Phi)$ is related to the observed free path lengths $(l_i)_{1 \leq i \leq n}$ by the likelihood function

$$L(\theta, \Phi) = \prod_{i=1}^{n} k(\theta, \Phi)^{S_i h_i} e^{-k(\theta,\Phi) S_i l_i}$$

where S_i is the incoming beam section and h_i is a binary variable indicating absence of hits.

Using both equations, LAD estimation is achieved by estimating Ω, specifying a model for $G(\theta, \Phi)$ with a normalization ensuring identification of LAD, and estimating its parameters by likelihood maximization. Moreover, G is determined by the leaf angle distribution $p(\theta_{L,j}, \Phi_{L,j})$ at leaf j, which in the end is the main focus of statistical modelling. State-of-the-art models assume independence of θ_L and Φ_L, uniformly distributed Φ_L and i.i.d. $\theta_{L,j}$ (see Pisek *et al.*, 2013 and Pimont *et al.*, 2018). However, it is expected that within tree crowns, $(\theta_{L,j})_j$ are spatially clustered into aggregated patches. This is why we aim at taking into account spatial aggregation and dependencies by incorporating the distribution of θ_L within a BNP-MRF model operating at voxel scale, thus representing spatial dependencies of leaf orientation between contiguous voxels.

References

[1] Achard, S., Salvador, R., Whitcher, B., Suckling, J. and Bullmore, E.D. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience* **26**(1), 63–72 (2006)

[2] Andersson, S., Madigan, D. and Perlman, M. Alternative Markov Properties for Chain Graphs. *Scandinavian Journal of Statistics* **28**(1), 33–85 (2001)

[3] Åström, K.J. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications* **10**, 174–205 (1965)

[4] Azaïs, R., Durand, J.-B. and Godin, C. Approximation of trees by self-nested trees. In *Proceedings of the 21st Meeting on Algorithm Engineering and Experiments (ALENEX 2019)*, 7-8 January 2019, pp. 39-53. San Diego (USA).

[5] Bangerth, F. Floral induction in mature, perennial angiosperm fruit trees: Similarities and discrepancies with annual/biennial plants and the involvement of plant hormones. *Scientia Horticulturae* **122**, 153–163 (2009)

[6] Barthélémy, D. and Caraglio, Y. Plant Architecture: A Dynamic, Multilevel and Comprehensive Approach to Plant Form, Structure and Ontogeny. *Annals of Botany* **99**(3), 375–407 (2007)

[7] Baudry, J.-P., Maugis, C. and Michel, B. Slope heuristics: overview and implementation. *Statistics and Computing* **22**(2), 455–470 (2012)

[8] Biernacki, C., Celeux, G. and Govaert, G. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7), 719–725 (2001)

[9] Bietti, A., Bach, F. and Cont, A. An online EM algorithm in hidden (semi-) Markov models for audio segmentation and clustering. In *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 19-24 April 2015, p. 1881–1885. IEEE Publishers, South Brisbane, Queensland, Australia.

[10] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146 (2017)

[11] Boudon, F., Jestin, A., Briand, A.-S., Fernique, P., Lauri, P.-É., Dambreville, A., Guédon, Y., Grechi, I. and Normand, F. The role of structural and temporal factors in the architectural development of the mango tree: evidences from simulation. *Acta Horticulturae* **1160**, 83–90 (2017)

[12] Carver, R.P. Reading rate: Theory, research, and practical implications. *Journal of Reading* 36(2), 84–95 (1992)

[13] Chatzis, S.P. and Tsechpenakis, G. The Infinite Hidden Markov Random Field Model *IEEE Transactions on Neural Networks* **21**(6), 1004–1014 (2010)

[14] Chaubert-Pereira, F., Guédon, Y., Lavergne and Trottier, C. Markov and semi-Markov switching linear mixed models used to identify forest tree growth components. *Biometrics* **66**(3), 753–762 (2010)

[15] Chen, S. and Reif, J.H. Efficient lossless compression of trees and graphs. In *Proceedings of the IEEE Data Compression Conference (DCC'96)*, 1 March–3 April 1996. Snowbird, UT (USA)

[16] Chickering, D.M. Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3**, 507–554 (2002)

[17] Choi, Y. and Szpankowski, J. Compression of graphical structures: fundamental limits, algorithms, and experiments. *IEEE Transactions on Information Theory* **58**(2), 620–638 (2012)

[18] Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. Sparse discriminant analysis. *Technometrics* **53**(4), 406–413 (2011)

[19] Costes, E., Smith, C., Renton, M., Guédon, Y., Prusinkiewicz, P. and Godin, C. MappleT: simulation of apple tree development using mixed stochastic and biomechanical models. *Functional Plant Biology* **35**, 936–950 (2008)

[20] Costes, E. and Guédon, Y. Deciphering the ontogeny of a sympodial tree. *Trees* **26**, 865–879 (2011)

[21] Crouse, M.S., Nowak, R.D. and Baraniuk, R.G. Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing* **46**(4), 886–902 (1998)

[22] Csiszár, I. and Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory* **52**(3), 1007–1016 (2006)

[23] Dambreville, A., Lauri, P.-É., Trottier, C., Guédon, Y. and Normand, F. Deciphering structural and temporal interplays during the architectural development of mango trees. *Journal of Experimental Botany* **64**(8), 2467–2480 (2013a)

[24] Dambreville, A., Fernique, P., Pradal, C., Lauri, P.-É., Normand, F., Guédon, Y. and Durand, J.-B. Deciphering mango tree asynchronisms using Markov tree and probabilistic graphical models. In *Proceedings of the Seventh International Workshop on Functional-Structural Plant Models (FSPM2013)*, 9-14 June 2013 (b). Saariselkä (Finland)

[25] de Reffye, P., Goursat, M., Quadrat, J. and Hu, B.-G. *The dynamic equations of the tree morphogenesis GreenLab model.* B.-G. Hu, M. Jaeger (Eds.), Plant Growth Modeling and Applications, Springer/Tsinghua University Press, Beijing, China (2003)

[26] Devijver, E. Joint rank and variable selection for parsimonious estimation in high-dimensional finite mixture regression model. *Journal of Multivariate Analysis* **157**, 1–13 (2017)

[27] Diligenti, M., Frasconi, P. and Gori, M. Image Document Categorization using Hidden Tree Markov Models and Structured Representations. In *Second International Conference on Advances in Pattern Recognition. Lecture Notes in Computer Science,* Singh, S., Murshed, N. and Kropatsch, W. (2001)

[28] Dumais, S. T. Latent semantic analysis. *Annual Review of Information Science and Technology* **38**, 188–230 (2004)

[29] Durand, J.-B., Gonçalvès, P. and Guédon, Y. Computational Methods for Hidden Markov Trees – An Application to Wavelet Trees. *IEEE Transactions on Signal Processing* **52**(9), 2551–2560 (2004)

[30] Durand, J.-B., Guédon, Y., Caraglio, Y. and Costes, E. Analysis of the Plant Architecture via Tree-structured Statistical Models: the Hidden Markov Tree Models. *New Phytologist* **166**(3), 813–825 (2005)

[31] Durand, J.-B., Caraglio, Y., Heuret, P. and Nicolini, E. Segmentation-based approaches for characterising plant architecture and assessing its plasticity at different scales. In *Proceedings of the Fifth International Workshop on Functional-Structural Plant Models (FSPM07)*, 4–9 November 2007. Napier (New-Zealand)

[32] Durand, J.-B., Girard, S., Ciriza, V. and Donini, L. Optimization of power consumption and device availability based on point process modelling of the request sequence. *Journal of the Royal Statistical Society Series C* **62**(2), 151–162 (2013a)

[33] Durand, J.-B., Guitton, B., Peyhardi, J., Holtz, Y., Guédon, Y., Trottier, C. and Costes, E. New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. *Journal of Experimental Botany* **64**, 5099–5113 (2013b)

[34] Durand, J.-B. and Fernique, P. Approche graphique pour la modélisation statistique de la dépendance entre activités journalières. In *Workshop "Statistique, Transport et Activités"*, Laboratoire Jean Kuntzmann, novembre 2013(c). Grenoble (France)

[35] Durand, J.-B. and Guédon, Y. Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models. *Statistics and Computing* **26**(1), 549–567 (2016)

[36] Durand, J.-B., Allard, A., Guitton, B., Van de Weg, E., Bink, M. and Costes, E. Predicting Flowering Behavior and Exploring Its Genetic Determinism in an Apple Multi-family Population Based on Statistical Indices and Simplified Phenotyping. *Frontiers in Plant Science* **8**, 858–872 (2017)

[37] Ephraim, Y. and Merhav, N. Hidden Markov processes. *IEEE Transactions on Information Theory* **48**, 1518–1569 (2002)

[38] Fernique, P. *A statistical modeling framework for analyzing tree-indexed data. Application to plant development on microscopic and macroscopic scales.* PhD dissertation, Université Montpellier 2 (2014a)

[39] Fernique, P., Durand, J.-B. and Guédon, Y. Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes. In *Compstat2014*, 19-22 August 2014 (b). Geneva (Switzerland)

[40] Fernique, P., Durand, J.-B. and Guédon, Y. Détection de motifs disruptifs au sein de plantes : une approche de quotientement/classification d'arborescences. In *47èmes Journées de Statistique (JDS2015)*, 1-5 June 2015. Lille (France)

[41] Fernique, P., Dambreville, A., Durand, J.-B., Pradal, C., Lauri, P.-É., Normand F. and Guédon, Y. Characterization of mango tree patchiness using a tree-segmentation/clustering approach. In *Proceedings of the International Conference on Functional-Structural Plant Growth Modeling, Simulation, Visualization and Applications (FSPMA2016)*, 7-11 November 2016. Qingdao (China)

[42] Fernique, P., Peyhardi, J., Durand, J.-B. Multinomial distributions for the parametric modeling of multivariate count data. Preprint hal-01286171, `https://hal.archives-ouvertes.fr/hal-01286171v1` (2016)

[43] Fletcher, J.C. Shoot and floral meristem maintenance in Arabidopsis. *Annual review of plant biology* **53**(1), 45–66 (2002)

[44] Friedman, J., Hastie, T. and Tibshirani, R. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9**(3), 432–441 (2008)

[45] Garbez, M., Symoneaux, R., Belin, É., Caraglio, Y., Chéné, Y., Donès, N., Durand, J.-B., Hunault, G., Relion, D., Sigogne, M., Rousseau, D. and Galopin, G. Ornamental plants architectural characteristics in relation to visual sensory attributes: a new approach on the rose bush for objective evaluation of the visual quality. *European Journal of Horticultural Science*, **83**(3), 187–201 (2018)

[46] Godin, C. and Caraglio, Y. A multiscale model of plant topological structures. *Journal of Theoretical Biology* **191**, 1–46 (1998)

[47] Godin, C. and Ferraro, P. Quantifying the degree of self-nestedness of trees. Application to the structural analysis of plants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7**(4), 688–703 (2010)

[48] Good, P. Chapter 3: testing hypotheses. In: *Permutation, parametric and bootstrap tests of hypotheses, 3rd Edition*, Good, P. (ed), p. 33-63. New York, USA: Springer (2005)

[49] Guédon, Y., Barthélémy, D., Caraglio, Y. and Costes, E. Pattern Analysis in Branching and Axillary Flowering Sequences. *Journal of theoretical biology*, **212**(4), 481–520 (2001)

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

[51] Guitton, B., Kelner, J.-J., Velasco, R., Gardiner, S.E., Chagné, D. and Costes, E. Genetic control of biennial bearing in apple. *Journal of Experimental Botany* **63**, 131–149 (2012)

[52] Josse, J. and Husson, F. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis* **56**, 1869–1879 (2012)

[53] Haccou, P., Jagers, P. and Vatutin, V.A. *Branching Processes: Variation, Growth, and Extinction of Populations.* Cambridge, UK: Cambridge University Press (2005)

[54] Hatt, C., Mankessi, F., Durand, J.-B., Boudon, F., Montes, F., Lartaud, M., Verdeil, J.-L. and Monteeuis, O. Characteristics of *Acacia mangium* shoot apical meristems in natural and *in vitro* conditions in relation to heteroblasty. *Trees - Structure and Function* **26**(3), 1031–1044 (2012)

[55] Hernando, D., Crespi, V. and Cybenko, G. Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Transactions on Information Theory* **51**(7), 2681–2685 (2005)

[56] Hoblyn, T.N., Grubb, N.H., Painter, A.C. and Wates, B.L. Studies in biennial bearing. *International Journal of Pomology and Horticultural Science* **14**, 39–76 (1936)

[57] Islam, M.A., Chowdhury, R.I. and Huda, S. *Markov models with covariate dependence for repeated measures.* New-York, USA: Nova Science Publishers (2009)

[58] Itokawa, Y., Katoh, K., Uchida, T. and Shoudai, T. Dictionary-based compression algorithms for tree structured data. In *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2009)*, Vol I, 18-20 March 2009. Hong-Kong

[59] Karlis, D. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* **30**(1), 63–77 (2003)

[60] Kass, R.E. and Wasserman, L. A Reference Bayesian Test for Nested Hypothesis and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association* **90**(431), 928–934 (1995)

[61] Kato, Z. Segmentation of color images via reversible jump MCMC sampling. *Image and Vision Computing* **26**(3), 361–371 (2008)

[62] Kato, Z. and Zerubia, J. Markov random fields in image segmentation. *Foundations and Trends® in Signal Processing* **5**(1-2), 1–155 (2012)

[63] Kipf, T. N. and Welling, M. Variational graph auto-encoders. *ArXiv preprint arXiv:1611.07308* (2016)

[64] Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques.* Cambridge, Massachussetts, USA: MIT press (2009)

[65] Lambert, A. and Popovic, L. The coalescent point process of branching trees. *The Annals of Applied Probability* **23**(1), 99–144 (2013)

[66] Lavielle, M. Using penalized contrasts for the change-point problem. *Signal Processing* **85**(8), 1501–1510 (2005)

[67] Lü, H., Arbel, J. and Forbes, F. Bayesian nonparametric priors for hidden Markov random fields. Preprint hal-02163046v2, `https://hal.archives-ouvertes.fr/hal-02163046v2` (2019)

[68] Molenberghs, G. and Verbeke, G. *Models for Discrete Longitudinal Data.* Springer Series in Statistics, New York, USA: Springer-Verlag. (2005)

[69] Nguyen, H.D., Forbes, F. and McLachlan, G.J. Mini-batch learning of exponential family finite mixture models. *ArXiv preprint arXiv:1902.03335* (2019)

[70] Obermaier, B., Guger, C., Neuper, C. and Pfurtscheller, G. Hidden Markov models for online classification of single trial EEG data. *Pattern recognition letters* **22**(12), 1299–1309 (2001)

[71] Olivier, B., Durand, J.-B., Guérin-Dugué, A. and Clausel, M. Eye-tracking data analysis using hidden semi-Markovian models to identify and characterize reading strategies. In *Proceedings of the 19th European Conference on Eye Movements (ECEM2017)*, 20-24 August 2017. Wuppertal (Germany)

[72] Percival, D.B. and Walden, A.T. *Wavelet methods for time series analysis.* Cambridge, United Kingdom: Cambridge University Press (2006)

[73] Peyhardi, J. , Caraglio, Y. , Costes, E. , Lauri, P. , Trottier, C. and Guédon, Y. Integrative models for joint analysis of shoot growth and branching patterns. *New Phytologist* **216**, 1291–1304 (2017)

[74] Pimont, F., Allard, D., Soma, M. and Dupuy, J.-L. Estimators and confidence intervals for plant area density at voxel scale with T-LiDAR. *Remote Sensing of Environment* **215**, 343–370 (2018)

[75] Pisek, J., Sonnentag, O., Richardson, A.D. and Möttus, M. Is the spherical leaf inclination angle distribution a valid assumption for temperate and boreal broadleaf tree species? *Agricultural and Forest Meteorology* **169**, 186–194 (2013)

[76] Polavaram, S., Gillette, T.A., Parekh, R. and Ascoli, G. Statistical analysis and data mining of digital reconstructions of dendritic morphologies. *Frontiers in Neuroanatomy* **8**(138) (2014)

[77] Poupart, P. and Vlassis, N. Model-based Bayesian Reinforcement Learning in Partially Observable Domains. In *International Symposium on Artificial International Symposium on Artificial Intelligence and Mathematics (ISAIM 2008)*, 2-4 January 2008. Fort Lauderdale, Florida (USA)

[78] Preuksakarn, C. , Boudon, F., Ferraro, P., Durand, J.-B., Nikinmaa, E. and Godin, C. Reconstructing Plant Architecture from 3D Laser scanner data. In *Proceedings of the Sixth International Workshop on Functional-Structural Plant Models (FSPM2010)*, 12-17 September 2010. University of California, Davis (USA)

[79] Revolle, M., Cayre, F. and Le Bihan, N. Clustering and causality inference using algorithmic complexity. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO 2017)*, 28 August-2 September 2017, p. 843–847. IEEE Press. Kos (Greece)

[80] Rezek, I., Gibbs, M. and Roberts, S.J. Maximum a posteriori estimation of coupled hidden Markov models. *Journal of VLSI signal processing systems for signal, image and video technology* **32**(1-2), 55–66 (2002)

[81] Runions, A., Lane, B. and Prusinkiewicz, P. Modeling Trees with a Space Colonization Algorithm. In *Proceedings of Eurographics Workshop on Natural Phenomena*, Ebert, D.S. and Mérillou, S. Eds, 3-7 September 2007, p. 63–70. Eurographics Association 2007, Prague (Czech Republic)

[82] Segura, V., Durel, C.-E. and Costes, E. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects: QTL mapping. *Tree Genetics & Genomes* **5**, 165–179 (2009)

[83] Simola, J., Salojärvi, J. and Kojo, I. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive systems research* **9**(4), 237–251 (2008)

[84] Sutton, R.S. and Barto, A.G. *Reinforcement learning: An introduction.* Second Ed. Cambridge, Massachusetts, USA: MIT press (2018)

[85] Tabachnick, B.G. and Fidell, L.S. Chapter 7, Multivariate Analysis of Variance and Covariance. In *Using multivariate statistics, 5th Edition*, Tabachnick, B.G and Fidell, L.S. Eds, p. 243–310. Pearson Education Inc./Allyn and Bacon, Boston, USA (2007)

[86] Taillard, É.D. Heuristic Methods for Large Centroid Clustering Problems. *Journal of Heuristics* **9**(1), 51–73 (2003)

[87] Thepaut, S. and Rigaill, G. Une méthode statistique pour détecter des ruptures multiples dans un arbre. In *51e Journées de Statistique*, 3–7 June 2019. Nancy (France)

[88] Tibshirani, R. and Walther, G. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* **14**(3), 511–528 (2005)

[89] Tong, W.-S. and Tang, C.-K. Robust estimation of adaptive tensors of curvature by tensor voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3), 434–449 (2005)

[90] Verma, T. and Pearl, J. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth international conference on uncertainty in artificial intelligence (UAI '92)*, Dubois, D., Wellman, M.P., D'Ambrosio, B. and Smets, P. Eds, 17-19 July 1992, p. 323–330. Morgan Kaufmann Publishers Inc. Stanford, California (USA)

[91] von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17**(4), 395–416 (2007)

[92] Whitcher, B., Guttorp, P. and Percival, D.B. Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research: Atmospheres* **105**(D11), 14941–14962 (2000)

[93] Yang, E., Ravikumar, P., Allen, G. and Liu, Z. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Pereira, F., Burges, C.J.C., Bottou, L. and Weinberger, K.Q. Eds, p. 1358–1366. Curran Associates, Inc., Red Hook, USA (2012)

[94] Yang, W., Pallas, B., Durand, J.-B., Martinez, S., Han, M. and Costes, E. The impact of long-term water stress on tree architecture and production is related to changes in transitions between vegetative and reproductive growth in the "Granny Smith" apple cultivar. *Tree Physiology*, **36**(11), 1369–1381 (2016)

[95] Zhang, K. Algorithms for the constrained editing distance between ordered labeled trees and related problems. *Pattern Recognition* **28**(3), 463–474 (1995)

[96] Zhong, S. and Ghosh, J. HMMs and coupled HMMs for multi-channel EEG classification. In *Proceedings of the 2002 International Joint Conference on Neural Networks (IJCNN '02)*, Vol 2, 12-17 May 2002. Honolulu, USA

[97] Ziv, J. and Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* **24**(5), 530–536 (1978)

A. Curriculum Vitæ

Details

Jean-Baptiste Durand	
Date and place of birth	January 16th, 1976 in Grenoble, France
Nationality	French
Personal Address	7 rue André Chénier, 38 400 Saint Martin d'Hères
Professional Address	Laboratoire Jean Kuntzmann CS 40 700 – 38058 Grenoble cedex 9 France
Phone	+33 (0)4 57 42 17 33
Email	`Jean-Baptiste.Durand@univ-grenoble-alpes.fr`
Webpage	`http://mistis.inrialpes.fr/people/jbdurand/`

Education and Qualifications

- 1999-2002: **PhD** in Applied Mathematics, Université Grenoble I (Joseph Fourier), France. Title: "Hidden Markov models: inference, model selection and applications" under the supervision of G. Celeux, INRIA Rhône-Alpes.
- 1998-1999: **DEA (MSc)** in Applied Mathematics at Université de Grenoble I. Subfield: Statistics and Probability.
- 1996-1999: **Engineering degree at ENSIMAG**, école Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble. Option: Decision Mathematics and Economy / Statistics and Probability, with honours.

Work Experience

- 2019-2020: detached to Inria, Rhône-Alpes, Mistis research team.
- 2004-today: Assistant Professor at Ensimag, Grenoble INP. Member of Laboratoire Jean Kuntzmann, Inria Mistis research team.
- 2009-2011: detached to Inria Sophia-Antipolis, Virtual Plants team in Montpellier.
- 2003-2004: Post-doctoral fellow at AMAP laboratory (botanique et bioinformatique de l'architecture des plantes) in Montpellier.
- Working part-time (80 %) in 2008-2009, 2012-2013, 2014-2015.

Co-supervision

Post-doctoral fellows

- 2013-2014: R. Azaïs (Inria). Multitype branching processes and applications to lossy compression and clustering of trees. R. Azaïs is now a researcher at Inria (Mosaic team).

PhD Students

- 2015-2019: B. Olivier (Université Grenoble Alpes). Topic: "Joint analysis of eye movements and EEGs using hidden Markov models". Cosupervision with M. Clausel (10%) and Pr. A. Guérin-Dugué (40%), defended on June 26$^{\text{th}}$ 2019. I obtained a derogation from the doctoral school to be the official supervisor. B. Olivier is employed by MeteoSwift in Grenoble.
- 2011-2014: P. Fernique (Université de Montpellier 2). Topic: "A framework for statistical modelling of tree-structured data – Applications to plant growth modelling at microscopic and macroscopic scales". Cosupervision with Y. Guédon, defended in December 2014. P. Fernique is now the head of a project in biostatistics at Limagrain, Clermont-Ferrand.
- 2007-2010: L. Donini (Université de Grenoble 1). Topic: "Statistical learning for structured multivariate data. Application to the optimization of printer networks". Industrial PhD with the Xerox Company, cosupervised with S. Girard (Inria, 25%), V. Ciriza and G. Bouchard (XRCE, 25% each). L. Donini did not defend the PhD and quit after the end of the 3$^{\text{rd}}$ year in favour of a position into another company.

PhD oral examinations

- November 17th, 2016: B. Henry (Université de Lorraine). Supervised by N. Champagnat and D. Ritchie. Topic: "Non-Markovian branching processes in population dynamics and genetics".

PhD committees

- 2014-2015: M. Garbez (Agrocampus Ouest, Centre d'Angers, industrial PhD). Topic: "Modelling the architecture and visual components of rose trees" (defended on November 23rd, 2016).

- 2011-2012: J. Peyhardi (Université de Montpellier 2). Topic: "A new generalized linear model (GLM) framework for analysing categorical data; application to plant structure and development" (defended on December 9th, 2013)

Supervision of Master students

- 2020: J. Li (Master 2 MSIAM, Université Grenoble Alpes and Grenoble INP). Topic: "Bayesian nonparametric models for hidden Markov random fields with medical applications". Cosupervision with S. Achard and F. Forbes (Inria Mistis).

- 2019: F. Dama (Master 2 MSIAM, Université Grenoble Alpes and Grenoble INP). Topic: "Bayesian non-parametric models for hidden Markov random fields on count variables: application to disease mapping". Cosupervision with F. Forbes and J. Arbel (Inria Mistis).

- 2015: L. Liu (Master 2 MSIAM, Université Grenoble Alpes and Grenoble INP). Topic: "Coupled hidden Markov models for joint analysis of eye movements and EEGs". Cosupervision with S. Achard and A. Guérin-Dugué (GIPSA-lab).

- 2011: P. Fernique (Master 2 in Bioinformatics, Université de Montpellier 2). Topic: "Modelling branching properties of plants with multitype branching processes".

- 2010: J. Peyhardi (Master 2 Ingénierie Mathématiques Statistiques et économie, subfield Statistique et Fiabilité, Université de Bordeaux 2). Topic: "Markov models with generalized linear mixed transition kernels".

Synthesis of research, teaching and other activities

Research

My career has been guided by two main directions of research: hidden Markov models and applications to plant structure analysis. The precise contributions obtained in these fields are provided in the main body of this document; however in what follows, an overview of my involvement in some formal research projects is provided.

As I left the AMAP laboratory in 2004 and joined the University of Grenoble, my collaborators and I had the opportunity to answer a call named ACI from the Ministry of Research. We gathered researchers from the plant science community in Montpellier, P. Ferraro who was issued from this community and then appointed as a teaching assistant in LaBRI (Bordeaux) and newcomers in applied probability and statistics from the University of Lyon. The topic of the project was the development of new methods in statistics and combinatorics dedicated to the analysis of tree-structured data, with a main focus on multitype branching processes (on the stochastic side) and edit distance computations (on the deterministic side), with their applications on plant architecture modelling. The project lasted from 2004 to 2007 and regarding the first axis, we applied hidden Markov tree (HMT) models to new data sets: Beech trees and *Symphonia Globulifera*. From a methodological point of view, the focus was on the one hand on combining HMTs and edit distances to obtain biologically relevant comparisons of tree structures and on the other hand, on validating our models by defining new statistical indices on trees, determining their distributions under HMT assumptions and assessing the fit between data and these distributions.

Then I had a professional mission at Inria in 2009. The national system of calls for project changed and was renamed "ANR calls"; P. Ferraro wrote a project with the same core of partners as the ACI project, now extending targeted applications of tree analysis (*e.g.*, to RNA data secondary structure). The project was not funded and I was in charge of writing a new project, focused this time on the analysis of tree-structured data issued from plant phenotyping. The scope of the project was to develop a similar set of models and algorithms to analyse these data as what was developed in the case of sequence analysis issued from omics: segmentation, alignment and comparison of DNA and RNA sequences. From a cognitive point of view, the aim was to separate different components of plant growth (which are detailed in Subsection 5.1). I thus coordinated the writing of a 15-pages document involving three laboratories in applied mathematics or plant biology. The project was not funded and was extended in 2010 to involve an additional partner: the "Institut de Recherche en Horticulture et Semences" laboratory in Angers, whose interest in our project was mainly focused on modelling rose bushes. I visited the lab to present our project and collect the needs for structured data analysis of their researchers. Once again the project was not funded but we could recycle it in a call from University of Montpellier 2. Formally the project (entitled as *Modelling architectural plasticity of plants: effect of genetic factors in response to environmental stress*) was headed by C. Trottier, who was with the University. In 2009, we co-supervised J. Peyhardi's master thesis in Montpellier; he then went into a PhD thesis co-supervised by C. Trottier and Y. Guédon. In 2010 I supervised P. Fernique's master thesis and our project yielded 50% in

his fundings for future PhD studies. His work, although focused on applications in plant growth modelling, was actually quite broad in terms of statistical models since it addressed the issue of identifying dependencies in multivariate counts. During the same year, I participated in recruiting a post-doctoral fellow at Inria, R. Azaïs and supervised him with C. Godin. During this 2-years period, I also took part to the presentation of a stand for two days at "Palais de la Découverte" in Paris. The topic was "From genes to flowers" and the exhibition, headed by C. Godin, was advertised on `http://www.palais-decouverte.fr/fr/ressources/docs-1chercheur1manip/modelisation-de-la-croissance-des-plantes/`.

In 2013, I initiated some discussions with an occasional teacher in statistics at Ensimag, I. Joly, on our research activities. As a researcher in transportation economics, he was interested in new models accounting for the choices of modes (or durations) of transport for the members of families, together with the associated activities requiring transport. More specifically, the aim was to predict and explain the dependencies within family members as well as regarding (possibly shared) resources (car, bicycles, season tickets). We answered a call for research projects from University of Grenoble 2 entitled as *Statistical modelling of multivariate counts. Applications to the analysis of transportation habits.* We involved other members of my research team, among whom M.-J. Martinez who formally headed the project since she was with University of Grenoble 2, while I was in charge of writing and coordinating the project. J. Peyhardi joined the project, given his interest for count data and economics. The funding was mainly dedicated to a few travels, material and for organizing a workshop on statistics and transports.

During the same period, I started preliminary analyses with A. Guérin-Dugué (Gipsa-Lab, Grenoble) on eye-movement experiments. We focused on scanpath modelling using HMC models (using our Matlab library *Chainxem*) and since HMCs revealed themselves unable to provide meaningful segmentations, we used HSMC models (using the Sequence Analysis component of the OpenAlea[11] project). However, these were not appropriate for joint modelling of eye movements and EEGs and we thus identified needs for new, generic models for analysing coupled, heterogeneous processes with asynchronous regime switches. It became manifest that researchers from four laboratories in Grenoble shared common interests and needs for stochastic models to analyse eye movements, though in different contexts (free exploration of images or texts, colour vision models, comparison of image exploration between children and adults) or scales (micro- vs. macro-saccades). We thus submitted a multidisciplinary project called *Oculo-Nimbus*[12] to a university call. Originally J.-F. Cœurjolly was the corresponding member for LJK but he went on a work leave program to Canada and I took the responsibility of representing LJK to present our project (often with A. Guérin-Dugué) to institutional meetings and to make periodic reports on project advances. *Oculo-Nimbus* provided funding for B. Olivier's PhD thesis and other post-doctoral or PhD fellowships involving the other axes only. This also funded a workshop on eye-movement analysis we organized in Grenoble in 2018.

While I was supervising J. Peyhardi's master thesis in Montpellier, E. Costes (AGAP laboratory, Montpellier) proposed that together with C. Trottier and Y. Guédon we took part to a project on joint modelling of growth and alternation in flowering in apple trees. In 2010, E.

[11] `https://github.com/openalea`

[12] `https://persyval-lab.org/fr/sites/oculo-nimbus`

Costes imagined an extension of this work, focusing on a reduction of the phenotyping workload required to assess yearly numbers of flowers. We co-supervised Y. Holtz's master thesis on that topic and the combination of these new methods and results led to a publication in the *Journal of Experimental Botany* (Durand *et al.*, 2013a). Now we could extend measurements to new families of apple trees and take benefit from both increased sample sizes, genetic diversity and links between families to enhance statistical detection of QTLs; this was at the core of A. Allard's PhD thesis and led to a publication in *Frontiers in Plant Science* (Durand *et al.*, 2017). These studies demonstrated the existence of genetic determinisms in alternation, supporting the assumption of hormonal control on floral induction involving developing fruits. But the role of nutritional competition between reproductive and vegetative growth should also be considered. From 2015 to 2019, I took part to a French-German ANR research project called AlternApp, *Genetic mechanisms underlying alternate cropping in apple (Malus x domestica)*[13], which focused on examining these two assumptions by genetic and genomics approaches. I was in charge of applying and extending our statistical methodology to new data sets and phenotyping protocols.

Teaching

When I became appointed as a teaching assistant at Ensimag in 2004, I was in charge of creating a new 36H course in statistical learning. I was totally free to decide the contents and chose to teach neural networks, probabilistic graphical models, mixture and hidden Markov models, computational statistics for Bayesian models and model selection. These topics had the advantages of being useful for engineers and to belong to research fields. The course contained about one half practical work. This course received the best ratings by the students among all the courses I ever gave; unfortunately the contents of the 3-years training program of engineers changed the year after and since this course could not fit in the schedules any more, it was discarded. I also succeeded some colleagues in introductory courses in statistics and multivariate analysis. I contributed to other courses in probability, proposed and supervised some student projects.

From 2006 to 2008 I coordinated "Information and Communication Technologies in Education" at Ensimag, together with F. Hetroy in computer science. This globally consists in pedagogic and technology watch, especially but not exclusively in new technologies for education. We particularly focused on problem- and project-based learning and went in Louvain-La-Neuve for a week to follow a training program. Then I applied these principles in my course in multivariate statistics and promoted the method at Ensimag. I also trained colleagues in statistics to multivariate analysis using the R software and headed two projects. The first one is an online French-English dictionary of mathematical concepts with their pronunciations (including audio recordings). The second one is a platform to organize student projects satisfying hierarchical constraints in nested groups, including forming teams, scheduling defences and gathering deliverables. The latter project was funded by a competitive university call; I thus wrote the project including its specification and hired a technician to develop the platform.

[13] `https://umr-agap.cirad.fr/recherche/projets-de-recherche/alternap`

In 2004, O. Gaudoin and I decided to propose training sessions in the continuing education department of the university (35H per year). We thus built industry-oriented training programs in data analysis, or introductory courses in statistics and probability. This lasted about three years, until the global load of the teachers became too heavy regarding conventional training and we had no time to ensure the sessions any longer.

From 2006 to 2009 I was in charge of coordinating 5-weeks full-time projects in probability, statistics and finance for teams of second-year students (call to projects, manage student teams, enforce rules and schedules, gather reports, organise and attend defences).

From 2014 to 2019, I was the head of the teaching staff in probability, statistics and finance at Ensimag. This mainly consists in checking the consistency of courses and training programs, the application of reforms decided by our university, committee of accreditation or ministry, compute the total teaching load of the staff, check the repartition between teachers and ensure that our topics remain visible and promoted in the various training pathways of the school. I had also to watch evolutions in educational programs at bachelor level, so that our pre-requisites and refresher courses remained adequate.

In 2014, Grenoble INP Ensimag and Grenoble Ecole de Management offered a new training program dedicated to big data [14]and I was in charge of defining its contents in machine learning.

In 2014 and 2015, I was a trainer in training sessions in Python for higher school preparatory teachers (introduction to scientific libraries).

Around 2013-2014, the number of research master's students in applied mathematics had fallen low, partly due to lack of possibilities for our Ensimag students to join the programs. I contributed in conceiving new programs in data science (now accessible to Ensimag students too). These were shared by two universities and two departments, Applied Mathematics (MSIAM[15]) and Computer Science. From 2015 to 2019, I was the head of the Statistics and Data Science track (with the help of O. Gaudoin in 2015). This occupation was particularly time-demanding. The co-accreditation by two universities and departments was a source for numerous administrative complications. The number of students increased regularly, from 12 in 2015 to 50 in 2019. This position involved assessment of application files, information and selection of students regarding fellowships, collecting and validating students' choices for optional courses, maintaining multiple university webpages, organizing committees to improve the training program, creating partnerships with other universities (particularly MIPT Moscow). Some shared courses had more than 100 students. Part of the courses were shared with other master programs (signal processing, operation research). The complexity was such that I spent more than 300 hours per year just in organization, with sometime deficient support of administrative staff (actually no support at all during several months). The heads of the shared tracks and master program agreed on the relevance of sharing more courses: modelling activities, data competitions, but there was some lack of volunteering teachers so I had to get personally involved and in the end, was in charge of four different courses in the master, which added to the Ensimag courses. In 2019, we were the first French program to obtain an ECMI (European Consortium for Mathematics in Industry[16]) accreditation. We published the principles

[14]https://ensimag.grenoble-inp.fr/fr/formation/big-data-analyse-management-et-valorisation-responsab
[15]http://msiam.imag.fr/
[16]https://ecmiindmath.org/

of a training program in data science in the journal *Statistique et Enseignement* (Amini *et al.*, 2016) and published the scientific contents in a book: *Data Science. Cours and exercices.* Eyrolles (éd.), 2018.

Since 2016, I have been in charge of a training project on data competitions. The aim is to organise and promote competitions in data analysis as training activities at the scale of the joint universities in Grenoble. The project includes a platform dedicated to hosting and running data competitions and a multimodal classroom to favour team working. The classroom includes clusters of tables, plugs and screens with mobile chairs and shelves, with shareable large screens. The project was presented in conferences (Durand 2017, 2019), one of which was in Grenoble; I was part of the organising committee.

The teaching load regularly increased from 2004 (192H) to 2018 (242H), with a pike in 2016 (263H), not mentioning unrewarded hours.

Other responsibilities

From 2011 to 2017, I was in charge of organising the seminars in Applied Probability and Statistics at LJK. This is the weekly seminar of our department.

From 2008 to 2013, I was the Health and Safety Manager at LJK. During the first year, this mainly consisted in transmitting information regarding occasional intrusions and damages caused to the building. However after one year the task was much more demanding since the university required the list of every room in the building with an inventory of each possible risk (in a very broad sense: having one's feet tangled in wire, the use of chained multi sockets or keeping stacks of paper, etc.).

Organizing committees at workshops and conferences

- Grenoble Workshop on Models and Analysis of Eye Movements, Université de Grenoble, June 6–8 2018.
- CFIES'2017 (Colloque Francophone International sur l'Enseignement de la Statistique), Université de Grenoble, September 6–8 2017.
- Workshop of AIGM network (Algorithmic Issues for inference in Graphical Models): 2011 session in Montpellier and 2015 session in Grenoble.

Selection committee for assistant professor positions

- 2018: Grenoble INP / Ensimag (position n. 0664, sections 26-27 Applied Mathematics and Computer Science)

Reviewing activity for the following journals:

- Statistics and Computing
- Behavior Research Methods
- PLOS Computational Biology (invited editor)
- Journal of Mathematical Biology
- Signal Processing Letters
- International Journal of Wavelets, Multiresolution and Information Processing
- IEEE Transactions on Signal Processing
- IEEE Transactions on Image Processing
- Transactions on Reliability
- International Journal of Computational Materials Science and Engineering

B. List of publications[1]

Foreword related to authors' order in publications.
Publications in Applied Mathematics usually sort authors by decreasing importance of contributions, or by alphabetical order in case of equal contributions.
Publications in Plant Science frequently place as last authors PhD supervisors or team leaders, even if their contribution is quite significant.

Preprint

- P. Fernique, J. Peyhardi and J.-B. Durand. Multinomial distributions for the parametric modeling of multivariate count data. `https://hal.inria.fr/hal-01286171`

Book chapter

- M. Clausel and J.-B. Durand. Modèles Génératifs. In *Data Science. Cours and exercices.* Eyrolles (éd.), p. 125-156, 2018.

Peer-reviewed international journals

- M. Garbez, R. Symoneaux, É. Belin, Y. Caraglio, Y. Chéné, N. Dones, J.-B. Durand, G. Hunault, D. Relion, M. Sigogne, D. Rousseau and G. Galopin. Ornamental plants architectural characteristics in relation to visual sensory attributes: a new approach on the rose bush for objective evaluation of the visual quality. *European Journal of Horticultural Science* 83(3):187–201, 2018.

- J.-B. Durand, A. Allard, B. Guitton, E. Van de Weg, M. Bink and E. Costes. Predicting Flowering Behavior and Exploring Its Genetic Determinism in an Apple Multi-family Population Based on Statistical Indices and Simplified Phenotyping. *Frontiers in Plant Science* 8:858-872, 2017.

- W. Yang, B. Pallas, J.-B. Durand, S. Martinez, M. Han and E. Costes. The impact of long-term water stress on tree architecture and production is related to changes in transitions between vegetative and reproductive growth in the "Granny Smith" apple cultivar. *Tree Physiology*, 36(11):1369-1381, 2016.

[1]Electronic version of documents available at `http://mistis.inrialpes.fr/people/jbdurand/`

- J.-B. Durand and Y. Guédon. Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models. *Statistics and Computing*, 26(1):549-567, 2016.

- J.-B. Durand, B. Guitton, J. Peyhardi, Y. Holtz, Y. Guédon, C. Trottier and E. Costes. New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. *Journal of Experimental Botany*, 64:5099-5113, 2013(a).

- J.-B. Durand, S. Girard, V. Ciriza and L. Donini. Optimization of power consumption and device availability based on point process modelling of the request sequence. *Journal of the Royal Statistical Society Series C*, 62(2):151–162, 2013(b).

- C. Hatt, F. Mankessi, J.-B. Durand, F. Boudon, F. Montes, M. Lartaud, J.-L. Verdeil and O. Monteeuis. Characteristics of Acacia mangium shoot apical meristems in natural and in vitro conditions in relation to heteroblasty. *Trees - Structure and Function*, 26(3):1031–1044, 2012.

- J.-B. Durand, Y. Guédon, Y. Caraglio and E. Costes. Analysis of the Plant Architecture via Tree-structured Statistical Models: the Hidden Markov Tree Models. *New Phytologist*, 166(3): 813–825, 2005.

- J.-B. Durand and O. Gaudoin. Software reliability modelling and prediction with hidden Markov chains. *Statistical Modelling - An International Journal*, 5(1):75-93, 2005.

- G. Celeux and J.-B. Durand. Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics*, 23(4):541–564, 2008.

- J.-B. Durand, P. Gonçalvès and Y. Guédon. Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, 2004a.

Peer-reviewed journals with national audience

- M.-R. Amini, J.-B. Durand, O. Gaudoin, E. Gaussier and A. Iouditski. Data Science : une formation internationale de niveau Master en science des données. *Statistique and Enseignement*, Société Française de Statistique, 7(1):95-102, 2016.

- J.-B. Durand, L. Bozzi, G. Celeux and C. Derquenne. Analyse de courbes de consommation électrique par chaînes de Markov cachées. *Revue de Statistique Appliquée*, LII(4):71–91, 2004b.

Peered-reviewed international conferences

- R. Azaïs, J.-B. Durand and C. Godin. Approximation of trees by self-nested trees. In *Proceedings of the 21st Meeting on Algorithm Engineering and Experiments (ALENEX 2019)*, 7-8 January 2019, pp.39-53. San Diego (USA).

- B. Olivier, J.-B. Durand, A. Guérin-Dugué and M. Clausel. Eye-tracking data analysis using hidden semi-Markovian models to identify and characterize reading strategies. In *Proceedings of the 19th European Conference on Eye Movements (ECEM2017)*, 20-24 August 2017. Wuppertal (Germany).

- J.-B. Durand, A. Allard, B. Guitton, E. Van de Weg, M. Bink and E. Costes. Genetic determinism of flowering regularity over years in an apple multi-family population. In *International Symposium on Flowering, Fruit Set and Alternate Bearing*, 19-23 June 2017. Palerme (Italy).

- B. Pallas, J. Ngao, J.-B. Durand, S. Martinez, S. Bluy, J.-J. Kelner and E. Costes. The Analysis of the Impact of Carbon Source-sink Relationships on Flowering Patterns Reveals That Apple Tree Growth and Functioning are Determined by Mechanisms Occurring at the Tree and Shoot Scales. In *Acta Horticulturae, Proceedings of the XI International Symposium on Integrating Canopy, Rootstock and Environmental Physiology in Orchard Systems*, 28 August-2 September 2016. Bologna (Italy).

- P. Fernique, A. Dambreville, J.-B. Durand, C. Pradal, P.-É. Lauri, F. Normand and Y. Guédon. Characterization of mango tree patchiness using a tree-segmentation/clustering approach. In *Proceedings of the International Conference on Functional-Structural Plant Growth Modeling, Simulation, Visualization and Applications (FSPMA2016)*, 7-11 November 2016. Qingdao (China).

- R. Azaïs, J.-B. Durand and C. Godin. Lossy compression of unordered rooted trees. In *Data Compression Conference DCC2016*, 29 March-1st April 2016. Cliff Lodge, Snowbird, Utah (USA).

- J.-B. Durand and Y. Guédon. Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles. In *Compstat2014*, 19-22 August 2014. Geneva (Switzerland).

- P. Fernique, J.-B. Durand and Y. Guédon. Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes. In *Compstat2014*, 19-22 August 2014. Geneva (Switzerland).

- A. Dambreville, P. Fernique, C. Pradal, P.-É. Lauri, F. Normand, Y. Guédon and J.-B. Durand. Deciphering mango tree asynchronisms using Markov tree and probabilistic graphical models. In *Proceedings of the Seventh International Workshop on Functional-Structural Plant Models (FSPM2013)*, 9-14 June 2013(c). Saariselkä (Finland).

- J.-B. Durand, B. Guitton, J. Peyhardi, Y. Holtz, Y. Guédon, C. Trottier and E. Costes. Estimating the genetic value of F1 apple progenies for irregular bearing during first years of production. In *Proceedings of the Seventh International Workshop on Functional-Structural Plant Models (FSPM2013)*, 9-14 June 2013(d). Saariselkä (Finland).

- C. Preuksakarn, F. Boudon, P. Ferraro, J.-B. Durand, E. Nikinmaa and C. Godin. Reconstructing Plant Architecture from 3D Laser scanner data. In *Proceedings of the Sixth International Workshop on Functional-Structural Plant Models (FSPM10)*, 12-17 September 2010. University of California, Davis (USA).

- J.-B. Durand, Y. Caraglio, P. Heuret and E. Nicolini. Segmentation-based approaches for characterising plant architecture and assessing its plasticity at different scales. In *Proceedings of the Fifth International Workshop on Functional-Structural Plant Models (FSPM07)*, 4-9 November 2007. Napier (New-Zealand).

- P. Heuret, J.-B. Durand, E. Nicolini, S. Coste and Y. Caraglio. Exploring morphogenetical gradients plasticity using hidden Markov tree models in young individuals of the tropical specie *S*ymphonia globulifera (Clusiaceae). In *Proceedings of the Fifth International Workshop on Functional-Structural Plant Models (FSPM07)*, 4-9 November 2007. Napier (New-Zealand).

- S. Dufour-Kowalski, C. Pradal, N. Dones, P. Barbier de Reuille, F. Boudon, J. Chopard, D. DaSilva, J.B Durand, F. Theveny, P. Ferraro, C. Fournier, Y. Guedon, C. Smith, S. Stoma, C. Godin and H. Sinoquet. OpenAlea: An open-software plateform for the integration of heterogenous FSPM components In *Proceedings of the Fifth International Workshop on Functional-Structural Plant Models (FSPM07)*, 4-9 November 2007. Napier (New-Zealand).

- C. Pradal, F. Boudon, N. Dones, J.-B. Durand, P. Barbier De Reuille, C. Fournier, H. Sinoquet and C. Godin. OpenAlea - A platform for plant modelling, analysis and simulation. In *Europython conference*, 3-6 juillet 2006. Geneva (Switzerland).

- J.-B. Durand, Y. Guédon, Y. Caraglio and E. Costes. Analysis of the Plant Architecture via Tree-structured Statistical Models: the Hidden Markov Trees In *Proceedings of the Fourth International Workshop on Functional-Structural Plant Models (FSPM04)*, édité by Godin *et al.*, UMR AMAP publisher, 7-11 June 2004, p. 61-64. Montpellier (France).

- J.-B. Durand and O. Gaudoin. Software reliability modelling and assessment with hidden Markov chains In *4th International Conference on Mathematical Methods in Reliability*, World Scientific Publishing publisher, Series on Quality, Reliability and Engineering Statistics, June 2004. Santa-Fe (USA).

- G. Celeux and J.-B. Durand. Choosing the order of a hidden Markov chain through cross-validated likelihood. In *Compstat2002*, 24-28 August 2002. Berlin (Germany).

- J. Martin and J.-B. Durand. Automatic Handwriting Gestures Recognition using Hidden Markov Models. In *Proceedings of the Fourth IEEE International Conference on Face and Gesture Recognition (FG2000)*, IEEE Press, New Jersey, Piscataway, 28-30 March 2000, pp. 403–409. Grenoble (France).

- K. Schwerdt, J.L. Crowley and J.-B. Durand. Robustification of detection and tracking of faces. In *Joint TMR Workshop on Computer Vision and Mobile Robotics*, September 1998, pp. 155–161. Santorini (Greece).

9 Communications at "Journées de statistique"[17] (not detailed)

[17]French-speaking conference

Other communications

- J.-B. Durand. Compétitions d'analyse des données à l'Université Grenoble Alpes: motivations, organisation et retours d'expérience. In *Colloque Francophone International sur l'Enseignement de la Statistique (CFIES2019)*, 25-27 September 2019. Strasbourg (France).

- J.-B. Durand. Challenges d'analyse de données : une formation par la pratique transversale and multidisciplinaire en science des données. In *Colloque Francophone International sur l'Enseignement de la Statistique (CFIES2017)*, 6-8 September 2017. Grenoble (France).

- J.-B. Durand, A. Guérin-Dugué and B. Lemaire. Analysis of eye movements and EEGs in reading tasks. In *Workshop on eye-movement analysis*, GIPSA-LAB, October 2015. Grenoble (France).

- J.-B. Durand and P. Fernique. Approche graphique pour la modélisation statistique de la dépendance entre activités journalières. In *Workshop "Statistique, Transport and Activités"*, Laboratoire Jean Kuntzmann, November 2013(e). Grenoble (France).

- P. Fernique, J.-B. Durand and Y. Guédon. Parametric modelling of multivariate count data using probabilistic graphical models. In *3rd Workshop on Algorithmic issues for Inference in Graphical Models – AIGM13*, September 2013(f). Paris (France).

- J.-B. Durand. Statistical models of sequences and trees in OpenAlea. In *First OpenAlea workshop on Functional-Structural Plant Modelling*, 14-15 June 2011. Montpellier (France).

- L. Donini, V. Ciriza, J.-B. Durand and S. Girard. A Statistical Model for Optimizing Power Consumption of Printers. In *XIG R&T Conference*, Xerox Corporation, May 2008. Webster (USA).