

THÈSE

Pour obtenir le grade de
Docteur

Délivré par **UNIVERSITE MONTPELLIER 2**

Préparée au sein de l'école doctorale SIBAGHE
Et de l'unité de recherche mixte AMAP

Spécialité : **BIO INFORMATIQUE**

Présentée par **GUILHEM BRUNEL**

**CARACTERISATION AUTOMATIQUE
D'ORGANISATIONS CELLULAIRES DANS
DES MOSAÏQUES D'IMAGES
MICROSCOPIQUES DE BOIS.**

Soutenue le 1^{er} octobre 2014 devant le jury composé de

M. Francis COLIN, chargé de recherche, INRA, Nancy	Rapporteur
M. Marc JAEGER, directeur de recherche, CIRAD, Montpellier	Examineur
M. Jean-Michel LEBAN, directeur de recherche, INRA, Nancy	Examineur
M. Grégoire MALANDAIN, directeur de recherche, INRIA, Sophia-Antipolis	Rapporteur
M. Gérard SUBSOL, chargé de recherche, CNRS, Montpellier	Examineur
M. Jean-Frédéric TERRAL, professeur, UM2, Montpellier (SIBAGHE)	Examineur
M. Philippe BORIANNE, ingénieur de recherche, CIRAD, Montpellier	Invité
M. Yves CARAGLIO, ingénieur de recherche, UM2, Montpellier	Invité

UNIVERSITE MONTPELLIER 2
ECOLE DOCTORALE SIBAGHE

Doctorat
Bio-informatique

GUILHEM BRUNEL

**CARACTERISATION AUTOMATIQUE D'ORGANISATIONS CELLULAIRES
DANS DES MOSAÏQUES D'IMAGES MICROSCOPIQUES DE BOIS.**

Application à l'identification des files cellulaires.

Thèse codirigée par Marc JAEGER et Gérard SUBSOL

Soutenue publiquement le 1^{er} octobre 2014

Jury :

Francis COLIN	Chargé de recherche, INRA, Nancy	Rapporteur
Marc JAEGER	Directeur de recherche, CIRAD, Montpellier	Examineur
Jean-Michel LEBAN	Directeur de recherche, INRA, Nancy	Président
Grégoire MALANDAIN	Directeur de recherche, INRIA, Sophia-Antipolis	Rapporteur
Gérard SUBSOL	Chargé de recherche, CNRS, Montpellier	Examineur
Jean-Frédéric TERRAL	Professeur, UM2, Montpellier (SIBAGHE)	Examineur
<i>Philippe BORIANNE</i>	<i>Ingénieur de recherche, CIRAD, Montpellier</i>	<i>Invité</i>
<i>Yves CARAGLIO</i>	<i>Ingénieur de recherche, UM2, Montpellier</i>	<i>Invité</i>

RESUME

Résumé

Ce travail se concentre sur l'analyse d'images numériques biologiques. Il vise à définir et mettre en œuvre des processus de mesures automatiques de données biologiques à partir d'images numériques dans un cadre de traitement de masse, et aborde notamment : l'incidence des choix méthodologiques sur la stabilité des résultats, l'étude de la validation des mesures produites et les limites de la généralité des méthodes et modèles appliquées à la biologie végétale.

La réflexion est menée dans le cadre de l'étude de certaines organisations cellulaires, et plus particulièrement de l'identification et l'analyse automatique de files cellulaires dans des mosaïques d'images microscopiques de bois. En effet, l'étude des tendances biologiques le long de ces structures est nécessaire pour comprendre la mise en place des différentes organisations et maturations de cellule. Elle ne peut être conduite qu'à partir d'une grande zone d'observation du plan ligneux. Pour cela,

- nous avons mis en place un nouveau protocole de préparation (rondelles de bois poncées) et de numérisation des échantillons permettant d'acquérir entièrement la zone d'observation sans biais
- nous avons développé une chaîne de traitement permettant l'extraction automatique des files cellulaires dans des mosaïques images numériques.
- nous avons proposé des indices de fiabilité pour chaque mesure effectuée afin de mieux cibler les études statistiques à venir.

Les méthodes développées dans la thèse permettent l'acquisition et le traitement rapide d'un volume important de données. Ces données devraient servir de base à de nombreuses investigations : des analyses architecturales des arbres avec le suivi de file cellulaire et/ou la détection de perturbations biologiques, des analyses de variabilité intra et inter arbres permettant de mieux comprendre la croissance endogène des arbres.

Mots-clés : Traitement d'image, file cellulaire, image microscopique de bois, fusion de données, mosaïque, reconnaissance d'organisations.

Abstract

This study focuses on biological numeric image processes. It aims to define and implement new automated measurements at large scale analysis. Moreover, this thesis addresses : the incidence of the proposed methodology on the results reliability measurements accuracy definition and analysis proposed approaches reproducibility limits when applied to plant biology.

This work is part of cells organization study, and aims to automatically identify and analyze the cell lines in microscopic mosaic wood slice pictures. Indeed, the study of biological tendencies among the cells lines is necessary to understand the cell migration and organization. Such a study can only be realized from a huge zone of observation of wood plane. To this end, this work proposes :

- a new protocol of preparation (slices of sanded wood) and of digitizing of samples, in order to acquire the entire zone of observation without bias,
- a novel processing chain that permit the automated cell lines extraction in numeric mosaic images,
- the definition of reliability indexes for each measurement allowing further efficient statistical analysis.

The methods developed during this thesis enable to acquire and treat rapidly an important volume of information. Those data define the basis of numerous investigations, such as tree architectural analysis cell lines following and/or detection of biological perturbations. And it finally helps the analysis of the variability intra- or inter- trees, in order to better understand the tree endogenous growth.

Keywords: *image processing, cell files, wood microscopic images, reliability indexes, pattern recognition, data fusion.*

REMERCIEMENTS

Je tiens à remercier mes directeurs de thèse, Marc Jaeger et Gérard Subsol, pour la confiance qu'ils m'ont accordée en acceptant d'encadrer ce travail doctoral, pour la rigueur qu'ils m'ont imposée et les conseils avisés qui m'ont permis de progresser dans cette phase délicate de « *l'apprenti-chercheur* ».

Je tiens tout particulièrement à remercier Philippe Borianne pour son encadrement, ses conseils avisés, sa disponibilité et son coaching efficace tout au long de ces trois années. Et Yves Caraglio pour son aide et ses conseils précieux en biologie.

Une thèse ne pouvant se faire seul, je remercie les techniciens toujours disponibles et efficaces (Merlin Ramel, François Pailler et Michael Guérault), ainsi que Nora Bakker et Marie-Laure De Guillen pour avoir géré les tracasseries administratives, les autres scientifiques de l'équipe (Christine Heinz, Frédéric Borne et Sébastien Griffon) et enfin les autres thésards, stagiaires ou post-doctorants que j'ai pu côtoyer pendant cette période (Pol, Jimmy, Dimitri, Annes, Diane, Vincent, entre autres...). Plus généralement, je remercie les membres de l'UMR AMAP au sein de laquelle j'ai réalisé ce travail pour leur agréable présence au quotidien.

Un grand merci à tous les extérieurs qui m'ont aidé dans la réalisation de ce travail : Jean Frédéric Terral (UM2), Philippe Montésinos (LG2IP), Bruno Clair (LMGC), Christophe Godin (INRIA), Cyrille Rathgeber (LERFoB) et Vincent Bonhomme (University of Sheffield, UK).

Je tiens à remercier également le Labex NUMEV et l'Ecole doctorale SIBAGHE qui ont cru en mon sujet de thèse et qui l'ont financé.

Je n'oublie pas mes parents, ma sœur et ma famille pour leur contribution, leur soutien et leur patience. Je tiens à exprimer ma reconnaissance envers Emmanuelle Coque qui a eu la gentillesse de me supporter pendant ces 3 ans de thèse.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis (Anthony, Cyril, Guilhem, Quentin, Nicolas, Paul, et bien d'autres), qui m'ont toujours soutenu et encouragé au cours de la thèse et de la réalisation de ce mémoire.

PREFACE

Le manuscrit de thèse écrit sous la forme de compilation d'articles s'adresse à un large public d'anatomistes du bois et de bio-informaticiens. Pour une lecture aisée et agréable pour les deux communautés, chaque chapitre comporte des informations importantes pour la compréhension des thèmes abordés et ce, sans rentrer dans les détails techniques tels que le paramétrage des protocoles, filtres, etc. Pour connaître les détails, une notification des publications avec le chapitre concerné est donnée afin de pouvoir s'y référer.

Dans un souci de clarté, la structure de chaque chapitre, bâti sur le même plan, contient trois parties. D'abord, nous présentons un rappel du contexte avec ou non un complément bibliographique afin de préciser l'objet, le cadre ou la portée de la partie. Ensuite, nous donnons la description des contributions et des méthodes développées ou utilisées. Enfin, nous discutons des résultats marquants déjà publiés, et, lorsque nécessaire, nous ajoutons des compléments d'informations commentés.

LISTE DES PRODUCTIONS SCIENTIFIQUES REALISEES DANS LE CADRE DE LA THESE

Ce travail prend la forme d'une thèse sur articles et est basé sur quatre publications respectivement numérotées de I à IV (cf. liste ci-dessous), acceptés dans des revues à comité de lecture ou des conférences internationales à comité de lecture avec acte.

Publications acceptées dans des revues ou conférences internationales à comité de lecture

I. PMA'12. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio Y.** 2012. *Automatic characterization of the cell organization in light microscopic images of wood: Application to the identification of the cell file.* In Guo, Y., Kang, M. Z., Dumont, Y. Eds, Plant growth modeling, simulation, visualization and applications. The Fourth International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications, IEEE, Shanghai, China, 31 October-3 November, pp. 58–65.

II. AOB'14. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio Y.** 2014. *Automatic Identification and Characterization of Radial File in Light Microscopic Images of Wood.* In: Annals of Botany. doi: 10.1093/aob/mcu119.

III. FSPM'13. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio, Y.** 2013. *Defining reliability coefficients in an automated method of identification and characterization of radial files in microscopic images of gymnosperms.* In Nygren, P., Lintunen, A., Godin, C., Nikinmaa, E., Sievänen, R. Eds, Proceedings of the 7th International Conference on Functional Structural Plant Models, Saariselka, Finland, 14 June 2013, pp. 82–84.

IV. SCIA'13. **Brunel G, Borianne P, Subsol G, Jaeger M.** 2013. *Simple-Graphs Fusion in Image Mosaic: Application to Automated Cell Files Identification in Wood Slices.* In Kämäräinen, J. K., Koskela, M. Eds, Image analysis: 18th Scandinavian Conference, SCIA 2013, Espoo, Finland, June 17-20, 2013. Proceedings, 2013, pp. 34–43.

Autres productions

Publication

Borianne P, Brunel G. 2012. *Automated valuation of leaves area for large-scale analysis needing data coupling or petioles deletion.* In Guo, Y., Kang, M. Z., Dumont, Y. Eds, Plant growth modeling, simulation, visualization and applications. The Fourth International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications, IEEE, Shanghai, China, 31 October-3 November, pp. 50–57.

Posters

1^{ère} Journée Scientifique du Labex NUMEV 2012 – Montpellier, France.

→ http://gbrunel.fr/posters/GBrunel_NUMEV_2012.pdf

7th International Conference on Functional-Structural Plant Models – Saariselka, Finland.

→ http://gbrunel.fr/posters/GBrunel_FSPM_2013.pdf

18th International Conference of Scandinavian Conferences on Image Analysis – Espoo, Finland.

→ http://gbrunel.fr/posters/GBrunel_SCIA_2013.pdf

2^{nde} Journée Scientifique du Labex LABEX NUMEV 2013 – Montpellier, France.

→ http://gbrunel.fr/posters/GBrunel_NUMEV_2013.pdf

3^{ème} Journée Scientifique du Labex NUMEV 2014 – Montpellier, France.

→ http://gbrunel.fr/posters/GBrunel_NUMEV_2014.pdf

TABLE DES MATIERES

0	ETAT DE L'ART	14
1	MATERIEL	21
1.1	LE MATERIEL VEGETAL	21
1.1.1	LES CLADES ET ESPECES : ESSENCES DE TRAVAIL	21
1.1.2	LA CROISSANCE ET L'ANATOMIE DES VEGETAUX	22
1.2	PROTOCOLES DE PREPARATION ET D'ACQUISITION DES ECHANTILLONS	24
1.2.1	LES PREPARATIONS DU PLAN LIGNEUX	25
1.2.2	PROTOCOLE 1. LES COUPES HISTOLOGIQUES COLOREES	25
1.2.3	PROTOCOLE 2. PONÇAGE DE RONDELLES DE BOIS	26
1.2.4	LA NUMERISATION DES ECHANTILLONS	26
1.3	RESULTATS DES ACQUISITIONS & DISCUSSIONS	27
2	METHODES	31
2.1	SEGMENTATION DES IMAGES	31
2.1.1	GENERALITES SUR LES METHODES DE SEGMENTATION	31
2.1.2	CHAINE DE TRAITEMENT : DE L'IMAGE A L'ELEMENT STRUCTURANT	33
2.1.3	APPLICATION DE LA SEGMENTATION AUX CELLULES	36
2.1.4	BILAN DE MISE EN ŒUVRE DE LA SEGMENTATION	37
2.2	MODELISATION DES ARRANGEMENTS	39
2.2.1	DEFINITION DE LA MODELISATION DANS LE CADRE DES FILES CELLULAIRES	39
2.2.2	PROCESSUS D'IDENTIFICATION DES ARRANGEMENTS	40
2.2.3	MISE EN ŒUVRE DE L'IDENTIFICATION DES FILES CELLULAIRES	44
2.2.4	CONCLUSION SUR LA MISE EN ŒUVRE DE LA MODELISATION	46
2.3	FUSION DES DONNEES	47
2.3.1	MOSAÏQUES DES RESULTATS	47
2.3.2	PRESENTATION D'UNE METHODE GENERIQUE DE FUSION DE GRAPHE	48
2.3.3	AGREGATION DES RESULTATS DE FILES CELLULAIRES	48
2.3.4	CONCLUSION	51
2.4	MESURES ET FIABILITES	54
2.4.1	RESULTATS NUMERIQUES	54
2.4.1.1	Mesure des paramètres	54
2.4.1.1.1	La circularité	54
2.4.1.1.2	La surface	55
2.4.1.1.3	Le périmètre	55
2.4.1.2	Séparations des composants	58
2.4.2	FIABILITE	58
2.4.2.1	Niveau global	59
2.4.2.2	Niveau des composants	60
2.4.2.3	Niveau de l'élément	61
2.4.2.3.1	Le périmètre	62

2.4.2.3.2	La surface	63
2.4.2.3.3	La circularité	64
2.4.3	CONCLUSION SUR LA MISE EN ŒUVRE DE LA FIABILITE DES MESURES	64
2.5	TYPAGE CELLULAIRE	65
2.5.1	DESCRIPTION DE LA METHODE	65
2.5.2	ETUDES PROSPECTIVES	66
2.5.3	CONCLUSION SUR LA MISE EN ŒUVRE DU TYPAGE CELLULAIRE	68
3	APPLICATION ET RESULTATS	70
3.1	IMPLEMENTATION	70
3.2	APPLICATION AUX FILES CELLULAIRES	71
3.2.1	JEU D'ESSAI	71
3.2.2	MISE EN ŒUVRE DE LA DETECTION AUTOMATIQUE DES FILES CELLULAIRES	73
3.2.3	LIMITES	76
3.2.4	PERFORMANCES	77
3.3	AUTRES APPLICATIONS	78
3.3.1	LA ROBUSTESSE DE LA METHODE DE SEGMENTATION	78
3.3.2	EVALUATION QUANTITATIVE DE LA SEGMENTATION	82
3.3.3	INTERET DES APPROCHES MULTI-ECHELLES POUR MAITRISER LA SUR-SEGMENTATION DU WATERSHED	84
3.3.4	GENERICITE DE LA METHODE POUR LA DETECTION DE CELLULES	86
4	CONTRIBUTIONS ET PERSPECTIVES	93
5	REFERENCES BIBLIOGRAPHIQUES	99
6	ANNEXES	106
6.1	ANNEXE 1 : IMAGES ET RESULTATS	107
6.2	ANNEXE 2 : PUBLICATIONS	116
7	INDEX	117

0 ETAT DE L'ART

Les plantes sont-elles des ordinateurs ? Les objets illustrés à la figure 1 ne sont pas la création d'un génie de l'informatique doué d'une intuition artistique. Il s'agit d'une tête de chou romanesco (*Brassica oleracea*), un proche parent comestible du brocoli et d'une inflorescence de lobélie géante (*Lobelia telekii*). Chacun des fleurons du chou romanesco est une reproduction en miniature du chou romanesco entier, ce qui lui confère sa beauté fascinante. Les mathématiciens appellent ces formes répétitives des fractales (Mandelbrot 1983). Par son schéma de croissance suivant une séquence d'instructions répétitives, le chou romanesco semble avoir été généré par un ordinateur.

Comme dans la plupart des végétaux, l'apex (extrémité) des tiges en croissance élabore de façon répétée un motif de feuille, de bourgeon, de tige... Ces schémas de développement répétitif sont à la fois génétiquement déterminés et soumis à la sélection naturelle. Cette prédétermination peut donner toutes sortes de formes en fonction du type de croissance, de l'échelle considérée et de la maturation des objets considérés. Ces structures peuvent être représentatives de la croissance des plantes.

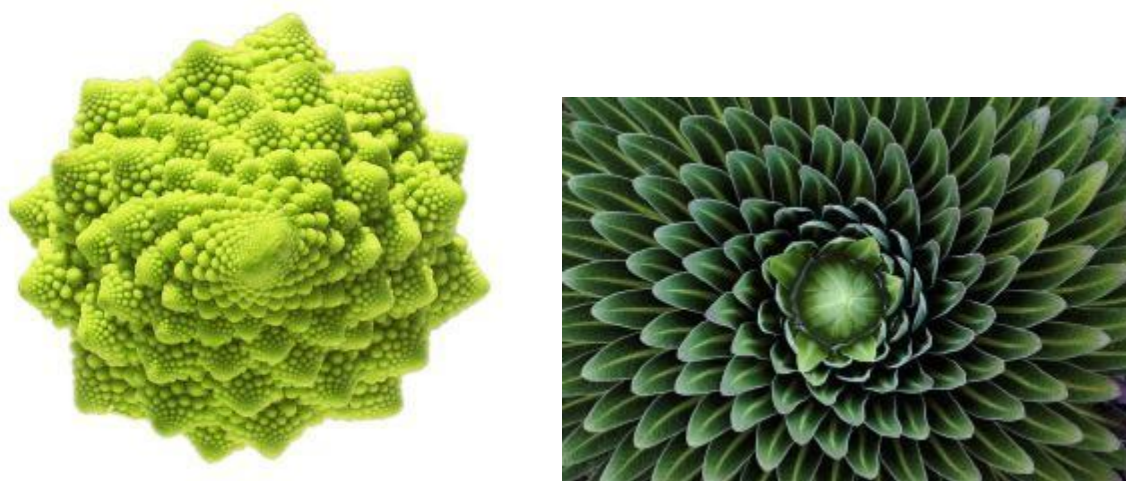


Figure 1. Plante ou art informatique ? A droite, un chou romanesco (*Brassica oleracea*) ; à gauche, une lobélie géante (*Lobelia telekii*). Les végétaux ont des schémas de développement répétitif pouvant former des structures telles que des spirales, fractales, lignes, etc.

Comment poussent les arbres ? Contrairement à la plupart des animaux, les végétaux ont une croissance qui ne se limite pas aux périodes embryonnaires et juvéniles, mais, au contraire, qui perdure toute la vie : c'est la croissance indéfinie (ou indéterminée). Ainsi, à tout moment de leur vie, les plantes possèdent des organes embryonnaires, des organes en croissance et des organes matures. Si les végétaux croissent de façon indéfinie, c'est qu'ils produisent constamment des tissus indifférenciés appelés méristèmes, qui se divisent quand les conditions le permettent pour produire de nouvelles cellules. Le développement de l'arbre résulte de la croissance primaire (en

rose sur la schéma ci-dessous) concernant l'allongement et la ramification des axes, et de la croissance secondaire (en bleu sur la schéma ci-dessous) portant sur leur épaisseur au cours du temps (Wilson 1984).

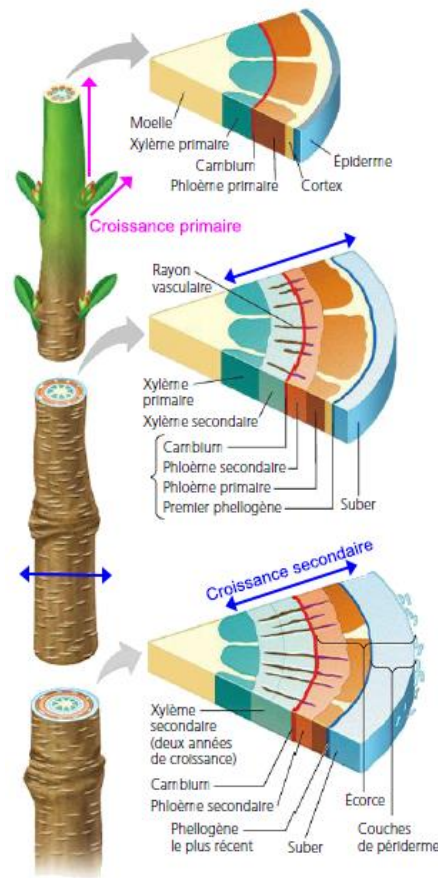


Illustration issue de Campbell Biologie 4ème édition

Figure 2. Les croissances primaire et secondaire d'une tige ligneuse. On peut suivre la progression de la croissance secondaire en examinant les sections dans lesquelles apparaissent en succession ses parties les plus vieilles, i.e. les cernes visibles sur une coupe de tronc d'arbre.

La réduction de notre empreinte environnementale et l'utilisation efficace des ressources naturelles sont des enjeux sociétaux et scientifiques du 21^{ème} siècle. De plus en plus d'approches ont tendance à considérer la plante sous des aspects structuraux et fonctionnels pour les problématiques de séquestration du carbone et de bois-énergie (Guo et al. 2011). Mais ces approches, qu'il s'agisse de la croissance secondaire et de ses variations inter ou intra spécifiques ou des relations entre croissances secondaire et primaire, s'appuient très souvent sur des études fragmentaires du fait des coûts d'obtention des données (Rossi et al. 2009).

La disposition des tissus, la forme et la taille des cellules, est appelée plan ligneux. Celui-ci est caractéristique de chaque essence. La description de la croissance primaire peut être plus ou

moins étendue grâce à des approches rétrospectives (Barthélémy and Caraglio 2007 ; Krause et al. 2010) alors que les études dendrochronologiques sont généralement limités à sur des parties du plan ligneux (carottage). Mais, plus le niveau d'observation est fin, plus les données existantes concernent une partie réduite du plan ligneux. La croissance secondaire résulte de l'activité cambiale qui est entre autre à l'origine du bois (xylème) et donc des différents éléments cellulaires qui le constituent (plan ligneux).

Pour mettre en rapport les observations internes avec les interprétations réalisées dans le cadre de l'observation de la morphologie externe, il est nécessaire d'étudier simultanément différents éléments de l'architecture végétale (Barthélémy and Caraglio 2007) en particulier de distinguer les caractéristiques anatomiques des éléments constitutifs de la plante (i.e. axes principaux, secondaires et réitérations) au fil du temps et de constituer un échantillon suffisamment significatif pour utiliser les outils statistiques.

La maturation et le typage biologique (Lachaud et al. 1999) des cellules peuvent être approchés par l'étude de leur forme, de leur taille, de leur contenu, et de leur distribution spatiale indépendamment de leur organisation spatiale. Ces notions permettent d'aborder les changements anatomiques qui caractérisent les phases de production et de différenciation du xylème (Thibeault-Martel et al. 2008), sans pour autant réellement expliquer les mécanismes de production ou de croissance du bois.

En effet, la compréhension des mécanismes de la croissance de l'assise cambiale passe par l'étude de la rythmicité des motifs cellulaires, de leur perturbation ou de leur modification dans l'espace et dans le temps. Les fluctuations et les influences de l'environnement sur la différenciation des éléments du bois (issus des divisions des cellules cambiales) demandent de suivre la production au sein d'organisations cellulaires spécifiques. L'introduction d'une trajectoire spatio-évolutionnelle est nécessaire pour aborder ces problèmes. Pour simplifier, deux organisations sont considérées : d'une part le cerne de croissance qui représente la production cellulaire à un temps donné (Heinrich 2007), et d'autre part la file cellulaire qui représente l'activité d'une initiale au cours du temps (Rozenberg et al. 2004).

Par exemple, (Nicolini et al. 2003) soulignent l'importance de l'organisation de la croissance secondaire en étudiant le long de cernes de croissance successifs les motifs cellulaires, leur rythmicité spatiale, leur variabilité pour caractériser les stratégies d'occupation de l'espace mises en place par la plante. Pour autant cette approche ne permet pas d'aborder les mécanismes de différenciation des cellules et de mise en place des motifs et structures cellulaires, c'est-à-dire les mécanismes dépendant du temps. De son côté, le suivi dynamique très précis de la production cambiale au niveau des initiales (Rossi et al. 2006 ; Rathgeber et al. 2011) est techniquement limité à une petite portion du plan ligneux. Ainsi, l'étude des files cellulaires, résultat *a posteriori* de la production cambiale et des fluctuations locales et globales, se présente comme

une piste prometteuse pour permettre d'appréhender la mise en place, la différenciation et la rythmicité temporelle des cellules (Gindl 2001).

Les questions biologiques portant sur les interactions entre le développement et la croissance des arbres sous contrainte de l'environnement (Moëll and Donaldson 2001) demandent de pouvoir :

- faire la part de la contribution aux fonctions de conduction et de stockage des différents types cellulaires dans le plan ligneux ;
- appréhender la gamme de variabilité des caractéristiques des éléments cellulaires (éléments de vascularisation, fibres, cellules de rayons, parenchymes verticaux) ;
- décomposer la variabilité en identifiant et différenciant celle due à l'ontogénèse de celle due à la réponse environnementale.

Mais ces questions biologiques et les caractérisations associées se trouvent extrêmement contraintes du fait de la lourdeur des protocoles et de la difficulté de l'acquisition des informations de masse. En effet, la compréhension fine des mécanismes de croissance et leur modélisation passent par la recherche des invariants au sein d'échantillons d'observation importants restituant la variabilité intra et inter espèces.

L'évolution technologique permet de limiter significativement les temps de préparation des sections transversales complètes d'axes et d'élargir ainsi les champs d'observation. Par exemple, l'utilisation d'une polisseuse (voir §1.2.3) automatique programmée peut nous éviter de recourir à des coupes histologiques trop fastidieuses à préparer. Combinée à une acquisition d'image pilotée par une platine à déplacements programmés, ce type de support nous donne rapidement accès à une information de qualité sur des zones élargies du plan ligneux. Mais cette technologie se heurte à la capacité de cumul des données et de son traitement. L'acquisition d'une rondelle de 10 cm de diamètre à un grossissement de x200, nécessite une mosaïque d'environ 1000 images. Il y a en moyenne 500 cellules par image. Evaluer manuellement les surfaces des cellules nécessite environ 40 mn par image à l'aide d'un éditeur d'image (ImageJ, Photoshop). L'identification d'organisations cellulaires impose d'étudier plusieurs cernes successifs dans plusieurs unités de croissance, ce qui représente une quantité d'images trop importante pour un traitement manuel. Habituellement, le comptage ne s'effectue donc que sur 3 files par coupe (Rossi et al. 2006), du fait des coûts prohibitifs en temps des traitements.

L'automatisation de l'étude des files cellulaires permettrait d'aborder des liens plus étroits avec les aspects fonctionnels et l'écologie des espèces en pouvant considérer une meilleure appréciation de la gamme de variabilité des phénomènes : le bois représente un enregistrement en continu des modifications du développement de l'arbre (partie ontogénique, la moins connue) et de son environnement (le plus travaillé notamment en dendrochronologie et dendroclimatologie). Cependant, il est fondamental de s'assurer de la fiabilité des résultats

produits et des solutions proposées. Dans l'idéal, il serait souhaitable de pouvoir introduire des éléments quantitatifs caractérisant la qualité dans les traitements automatisés et leur production.

L'automatisation de l'observation des structures cellulaires est l'un des nouveaux enjeux de la botanique structurale (Quelhas et al. 2011) qui fait appel à des compétences pluridisciplinaires. En (bio)-imagerie, la segmentation des cellules est un problème largement traité dans la littérature (Baggett et al. 2005 ; Fourcaud et al. 2008). Par exemple, (Park and Keller 2001) combinent quatre approches classiques d'analyse d'image : l'algorithme du watershed (Vincent and Soille 1991), les *snakes* (Mcinerney and Terzopoulos 1999), l'analyse multi-résolution (Jeacocke and Lovell 1994), et la programmation dynamique pour segmenter les cellules dans les images. Dans le contexte spécifique de l'imagerie du bois, la segmentation des cellules est généralement combinée au typage cellulaire, c'est-à-dire à la caractérisation biologique de la cellule. Par exemple, (Marcuzzo et al. 2009) et (Wang et al. 2008) typent les cellules à l'aide de classifieurs *Support Vector Machine* (Scholkopf and Smola 2001), méthode supervisée développée à partir des considérations théoriques de Vladimir Vapnik (Vapnik 1963) sur les statistiques de l'apprentissage. La difficulté majeure porte ici sur la constitution de jeux d'apprentissages suffisamment représentatifs pour rendre compte de la variabilité intrinsèque des individus.

Pour l'étude des organisations cellulaires, certains auteurs introduisent des modèles géométriques à base topologique permettant de décrire le voisinage des cellules. Par exemple, (Jones and Bischof 1996) ou (Kennel et al. 2010) s'appuient sur un graphe d'adjacence orienté de cellules décrivant leur relation topologique locale pour extraire les files cellulaires d'images de Gymnospermes. Plus précisément les cellules sont extraites par application de l'algorithme de ligne de partage des eaux sur l'image convertie en niveau de gris ; le graphe d'adjacence des cellules est alors construit à partir du diagramme des bassins versants. Un typage des cellules obtenu par un arbre de décision construit à la volée à partir d'un jeu de données annotées (CART) (Breiman et al. 1984) permet alors d'extraire de ce graphe les alignements de trachéides. Cette méthode est sensible au jeu d'apprentissage utilisé lors de la construction de l'arbre de décision et son paramétrage.

Du point de vue technique, nous n'avons pas trouvé de solution logicielle permettant une détection automatique des files cellulaires. Des outils commerciaux spécialisés, comme WinCell (Hitz et al. 2008), existent pour l'analyse des cellules du bois sans toutefois permettre de reconnaître et caractériser sans aide de l'opérateur les organisations cellulaires. Les possibilités d'enrichissement de fonction sont restreintes voire impossibles. Des plateformes commerciales telles que Visilog (Travis et al. 1996), proposent un environnement riche mais pas suffisamment spécialisé : il est possible de réaliser des macros fonctions (assemblage et paramétrage de fonctionnalités de base) mais sans possibilité d'ajouter des fonctionnalités de bases qui se révèlent incontournables dans notre cas pour modéliser les files cellulaires.

Les plateformes open source de type ImageJ (Clair et al. 2007), Fiji (Schindelin et al. 2012), Orfeo (Inglada and Christophe 2009) restent donc une option intéressante car elles permettent d'enrichir les codes sources de fonctionnalités particulières et de définir des macros spécialisées tout en facilitant le partage et la diffusion des méthodes. Il est donc possible de développer de nouvelles fonctionnalités qui permettent l'automatisation des traitements. Nous privilégierons donc cette solution de développement technique.

Notre objectif est donc de définir et construire une chaîne de traitement comprenant la préparation ainsi que l'acquisition des échantillons et la définition d'une méthode générique d'identification automatique de composants et de leurs arrangements à partir d'images numériques microscopiques. Pour suivre les arrangements sur de larges zones d'observation, nous avons mis au point une méthode d'agrégation des résultats en mosaïque pavant tout l'espace d'étude. Cette méthode est appliquée et testée au cas particulier des files cellulaires pour l'étude des tendances biologiques le long des files, nécessaire à la compréhension de la mise place des différentes organisations et maturations cellulaires, c'est-à-dire à la compréhension de la croissance endogène des plantes.

Pour réaliser ce travail, nous avons décidé d'adopter une approche hybride consistant à plaquer sur une approche relativement classique un regard ou un questionnement plus large. En particulier nous nous sommes intéressés à :

- l'incidence de choix méthodologique sur la stabilité des résultats, i.e. comment les étapes du traitement vont interagir et compenser les faiblesses des opérations en amont et en aval ;
- la validation des mesures produites, i.e. comment « garantir » l'exactitude des mesures réalisées : quels approches ou outils peuvent être développés pour les qualifier ;
- la généricité des méthodes et modèles appliquées à la biologie végétale, i.e. à partir de quand une méthode doit être spécialisée pour produire des résultats utilisables.

Le chapitre 1 est consacré au matériel biologique, aux méthodes de préparation et d'acquisition de ce matériel. Le lecteur intéressé pourra se reporter à une première partie (§ 1.1) traitant des spécificités des végétaux et des essences échantillonnées. La seconde partie (§ 1.2) porte sur un état de l'art des méthodes de préparation des coupes histologiques, ainsi que sur la mise en place d'un nouveau protocole consistant à préparer des rondelles de bois poncées pour la numérisation. Enfin la dernière partie est consacrée au protocole d'acquisition et au jeu de données “images” à traiter.

La méthode d'analyse d'image est décrite dans le chapitre 2. Nous argumentons cette présentation sous forme de “blocs”, puis décrivons le détail de chaque bloc. Le paragraphe 2.1 explicite les procédés conduisant à la segmentation. Ensuite le (§ 2.2) détaille la méthode d'identification des files cellulaires grâce à un parcours sous double contrainte d'un graphe. Une

fois les files créées par image nous montrons comment créer une mosaïque de files (§ 2.3) tout en évitant les problèmes de stitching qui apparaissent quand on fusionne des images pour en faire une mosaïque. Les deux dernières parties portent respectivement sur le calcul des paramètres et sur leur fiabilité.

Enfin, les résultats obtenus au cours de la thèse sont présentés et discutés dans le chapitre 3. Ils sont principalement produits par l'application de la méthode dans le cadre de la recherche automatique de files cellulaires. D'autres applications seront présentées dans le paragraphe (§ 3.3).

Les 4 grandes parties de l'étude ont fait l'objet chacune d'une publication dans une revue internationale ou dans les actes d'une conférence internationale à comité de lecture. Les articles sont regroupés en annexe et respectivement numérotés I à IV. Pour faciliter la lecture de ce manuscrit, chaque partie reprendra les grandes lignes de l'article lui correspondant et précisera les aménagements réalisés ou les résultats produits depuis la parution de l'article.

En complément, la figure 3 illustre les connexions entre les différents articles ainsi que les analyses additionnelles réalisées. Les flèches verticales épaisses représentent la trame principale, basée sur la mise au point des algorithmes de segmentation et d'identification. Les flèches horizontales concernent les applications de la méthode aux files cellulaires.

Les résultats qui découlent de ce travail peuvent servir de base aussi bien à des investigations scientifiques d'analyse de la distribution de caractéristiques internes (cellules) en fonction de facteurs divers tels que les conditions de croissance ; une meilleure compréhension des mécanismes de croissance des arbres et une amélioration des connaissances physiologiques (zone de conductivité du tronc, pattern de croissance), qu'à des applications industrielles vérifiant la qualité du bois à partir de méthodes non destructives de coupe de rameaux secondaire ou de micro-carottage.

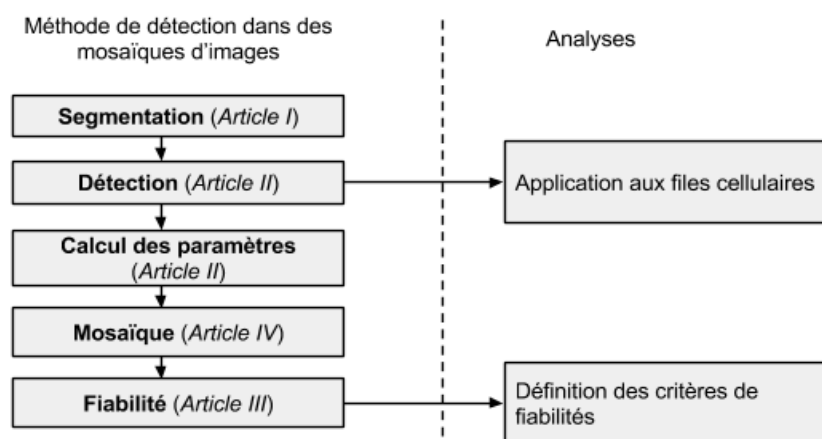


Figure 3. Liens entre les différentes parties (articles et analyses complémentaires) constituant le corps de la thèse.

1 MATERIEL

1.1 Le matériel végétal

1.1.1 Les clades et espèces : essences de travail

Depuis que les végétaux ont commencé à coloniser la terre ferme, il y a environ 500 millions d'années (cf. figure 4), ils se sont fortement diversifiés, de sorte qu'on compte aujourd'hui plus de 260 000 espèces, dont certaines occupent les milieux les plus hostiles, tels les pics montagneux, les régions désertiques et polaires.

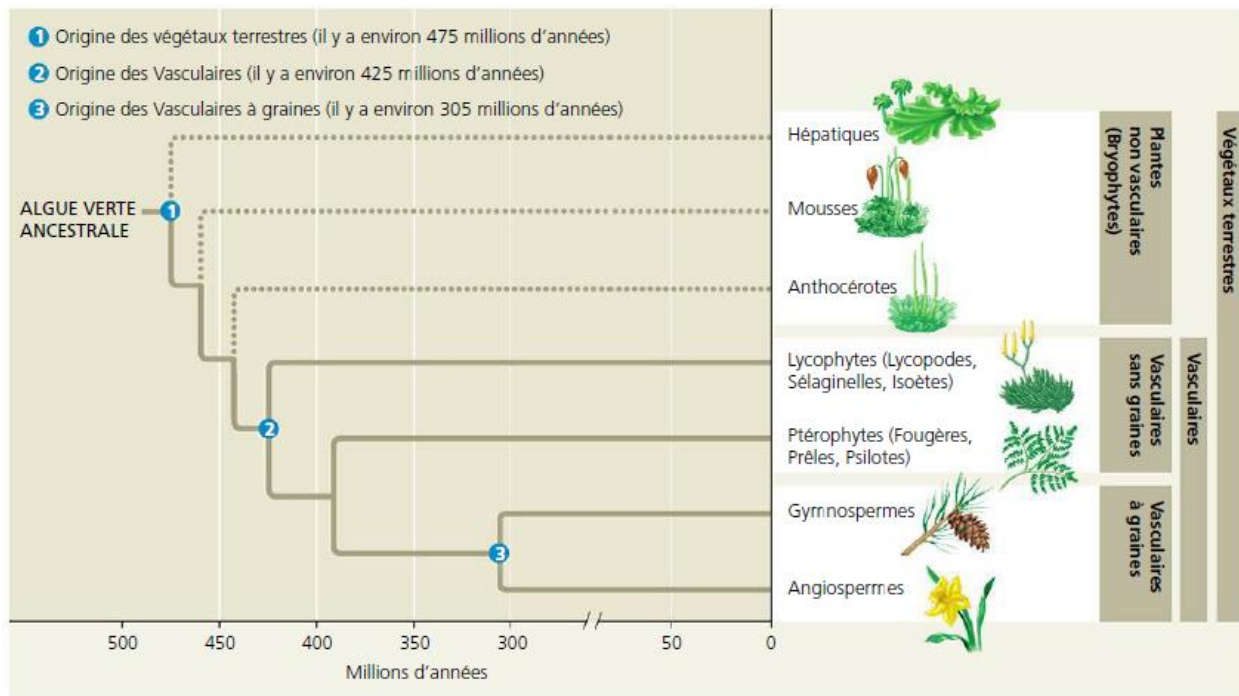


Illustration issue de Campbell Biologie 4ème édition

Figure 4. Arbre phylogénétique simplifié représentant quelques grands épisodes de l'évolution des végétaux. Cette phylogénèse représente une hypothèse sur les liens de parenté entre les groupes de végétaux. Les lignes pointillées indiquent que la phylogénie fait toujours l'objet de débats. Pendant la thèse nous nous intéressons aux plantes vasculaires à graines.

Notre étude portera plus particulièrement sur les clades des plantes vasculaires à graines modernes, les Gymnospermes et les Angiospermes, car elles représentent 90% des plantes terrestres.

Les Gymnospermes sont des plantes faisant partie d'un sous-embranchement paraphylétique des spermapytes. Gymnospermes vient du grec "gymnos" qui veut dire "nue" et "sperma" qui veut dire graine. Ils ne possèdent pas d'ovaire. Mais les Gymnospermes constituent un groupe en voie de disparition. En effet, ce groupe n'a fait que régresser et compte aujourd'hui moins de 1 000

espèces. La plupart des Gymnospermes sont des conifères, et nous avons choisi des représentants présents dans le sud de la France en raison de leur accessibilité qui sont le pin (*Pinus nigra*, *Pinus caribensis* et *Pinus brutia*) et le sapin (*Abies alba*).

La division des Angiospermes regroupe les plantes à fleurs, et donc les végétaux qui portent des fruits. Angiosperme vient du grec “angios” qui veut dire récipient. Ils regroupent la majeure partie des espèces végétales terrestres, avec environ 250 000 espèces. Les Angiospermes comprennent les dicotylédones et les monocotylédones. Ce groupe est particulièrement important pour les humains et leur survie. En foresterie et en agriculture, il constitue les sources essentielles de nourriture, de combustible, de produits du bois et de médicaments...

En raison de cette dépendance, il est indispensable de préserver la diversité de ces plantes (Angiospermes et Gymnospermes compris). Nous avons choisi de travailler sur les essences de chêne vert (*Quercus ilex*) et d'acajou (*Pycnanthus sp.*) comme représentant des Angiospermes dicotylédones. Nous expliciterons dans la partie ci-dessous pourquoi les monocotylédones ne sont pas étudiées.

1.1.2 La croissance et l'anatomie des végétaux

Les croissances primaire et secondaire se produisent simultanément dans les plantes présentant un double développement. La croissance primaire ajoute des feuilles, des nouveaux axes et allonge les tiges et les racines (la plante grandit), alors que la croissance secondaire augmente le diamètre des tiges et des racines (la plante grossit) là où (généralement) la croissance primaire est terminée. La figure 5 schématise la croissance d'un axe (tronc, branche, racine).

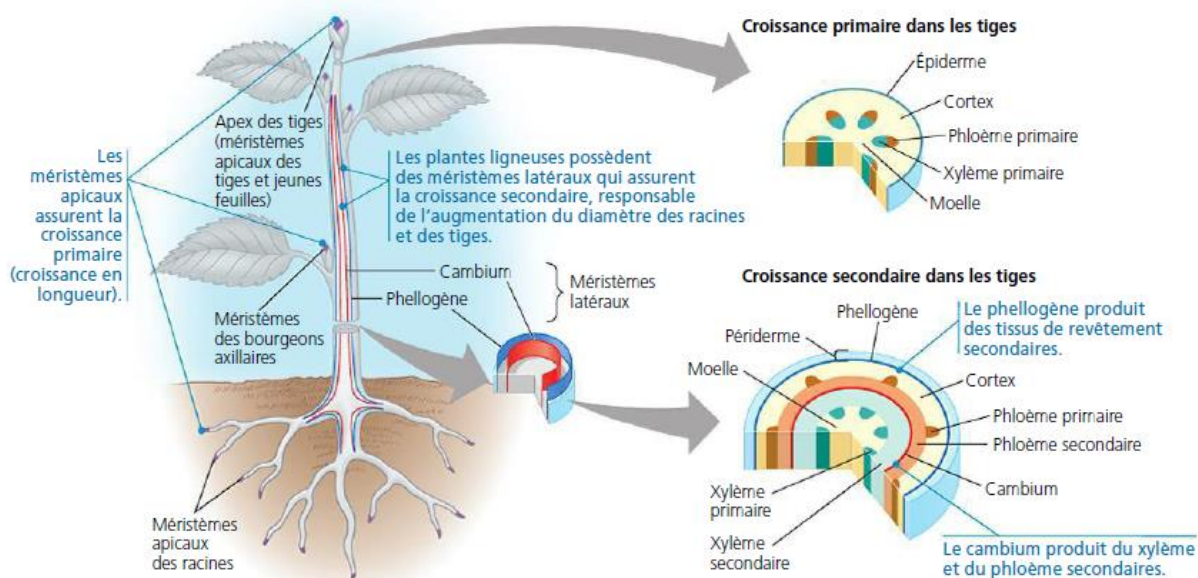


Illustration issue de Campbell Biologie 4ème édition

Figure 5. Vue d'ensemble des croissances primaire et secondaire.

Chez les Angiospermes monocotylédones, peu d'espèces ont une croissance secondaire. Ce groupe comporte notamment les orchidées, les palmiers, les bananiers et les graminées. Or le but de la détection des files cellulaires est de suivre l'évolution de la croissance secondaire. C'est pour cette raison que nous n'avons pas choisi de faire les expériences sur les monocotylédones.

La croissance secondaire s'effectue grâce aux méristèmes latéraux, plus précisément par le cambium (du latin *cambiare*, «changer») et par le phellogène (du grec *phellos*, «liège»). Le cambium produit vers l'intérieur des couches de tissus conducteurs supplémentaires appelées xylème secondaire (bois) et vers l'extérieur du phloème secondaire (le liber) (cf. figure 6). Ce sont les couches internes qui vont constituer les files, objet de notre étude (cf. figure 7).

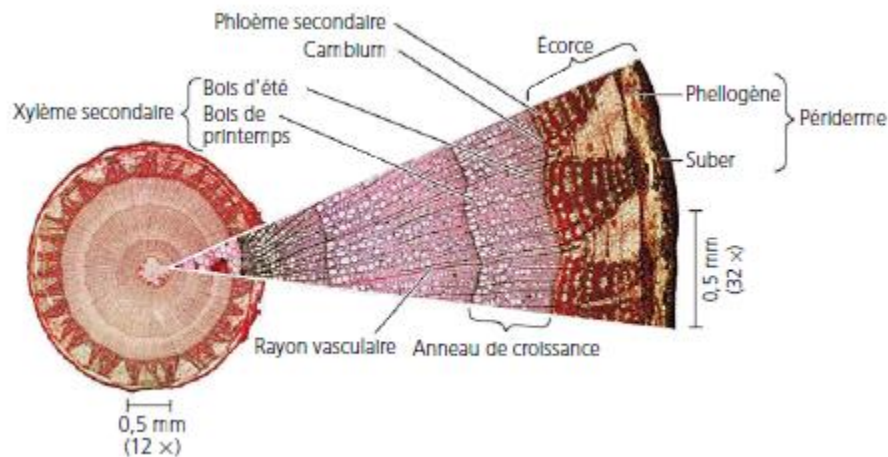


Illustration issue de Campbell Biologie 4ème édition

Figure 6. Coupe transversale d'une tige de tilleul (*Tilia sp.*) de trois ans. Vers l'intérieur, les nouvelles cellules créées se différencient en grandes cellules lignifiées de xylème secondaire ou bois. Vers l'extérieur, les nouvelles cellules se différencient en petites cellules formant le liber, certaines cellules contribuent à former l'écorce.

Le xylème secondaire mène à la formation de vaisseaux, de fibres, de parenchymes verticaux et horizontaux.

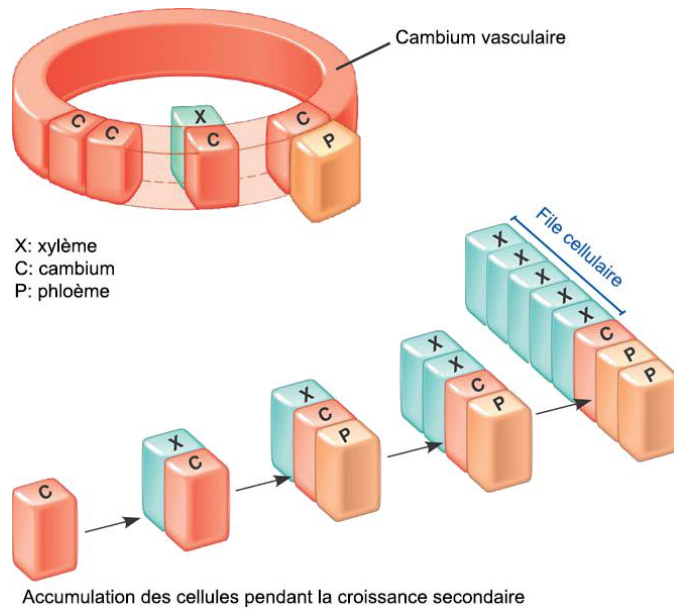


Illustration issue de Campbell Biologie 4ème édition

Figure 7. Schéma de l'accumulation de cellules filles de xylème à partir d'une cellule mère cambiale formant une file cellulaire.

Les deux types de cellules conductrices de sève brute, les trachéides et les éléments de vaisseaux, sont des cellules allongées tubulaires qui meurent à maturité. Les vaisseaux sont de plus grosse taille que les trachéides, et sont généralement absents chez les Gymnospermes qui possèdent, eux, des canaux résinifères.

Dans notre étude nous considérons 3 types cellulaires :

- Les fibres : faisceaux de cellules résistantes, disposées dans le sens axial (de la tige), assurant la rigidité et la résistance mécanique du bois.
- Les vaisseaux : cellules creuses, disposées dans le sens axial (de la tige), servant à conduire la sève.
- Les rayons : cellules à orientation transversale et rayonnante par rapport à l'axe longitudinal de l'axe (primaire ou secondaire), participant à la fonction de soutien.

Les autres types cellulaires sont considérés comme des fibres car avec un seul plan de coupe, il est difficile de pouvoir identifier correctement tous les types cellulaires, notamment les tissus de soutien (collenchyme et sclérenchyme) et le tissu assimilateur (parenchyme).

1.2 Protocoles de préparation et d'acquisition des échantillons

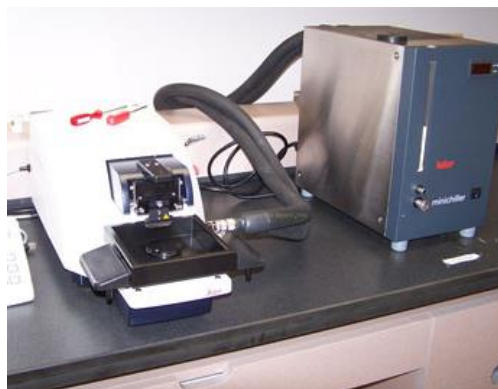
Le protocole complet d'acquisition des images comporte deux étapes : la première (préparation) vise à nettoyer et contraster au mieux les structures cellulaires, la seconde (numérisation) consiste à prendre un cliché de l'échantillon à l'aide d'un appareil photo numérique monté sur l'optique du microscope.

1.2.1 Les préparations du plan ligneux

Deux méthodes de préparation des échantillons pour obtenir des plans ligneux transversaux sont utilisées avant la numérisation qui consiste en une prise de photographie de l'échantillon sous microscope.

Le premier protocole (standard) consiste à effectuer des coupes histologiques fines des échantillons avec un microtome. De par son temps de préparation long et minutieux, ce protocole est mal adapté à de grandes zones d'études. De plus il induit des biais lors du pavage des images pour reconstruire une mosaïque du plan.

Le second protocole étudié consiste à poncer des rondelles de bois. Il est particulièrement bien adapté à l'étude de grandes zones et présente l'avantage d'être économe en temps. Pour l'instant le protocole n'introduit pas de coloration de l'échantillon, ce qui entraîne une différence entre les parois et les lumens moins marquée que sur les coupes histologiques. Des tests sont actuellement menés pour introduire une phase de coloration pendant ce protocole.



Clichés de l'auteur

Figure 8. A gauche, la polisseuse semi-automatique avec un plateau magnétique d'un diamètre de 300 mm (Tegramin 30 - Struers). A droite, le microtome automatique à lame vibrante (vibratome VT1200S-Leica).

1.2.2 Protocole 1. Les coupes histologiques colorées

L'obtention d'une coupe transversale de 20-25 μm d'épaisseur produite à l'aide d'un vibratome ou microtome (cf. figure 8) nécessite entre 5 à 15 minutes. Les coupes sont ensuite nettoyées et colorées pour augmenter le contraste paroi-lumen, en utilisant des colorants se fixant spécifiquement sur les parois pectocellulosiques tel que le carmin aluné. Cette phase de nettoyage et de coloration dure environ 30 minutes. Par conséquent il est judicieux de prévoir une heure pour un protocole complet par coupe. Cette méthode permet de réaliser des mosaïques de coupes histologiques, aux imprécisions des tissus de juxtaposition près (problème de calage et destructions ou déformations partielles lors de la coupe), sur quelques cernes selon les dimensions de la carotte.

N.B. : L'utilisation de colorant permet de faire ressortir des éléments désirés, par exemple le vert d'iode fixateur de la lignine fait ressortir les parois des cellules.

Nombreux sont les colorants et les protocoles de coloration, mais nous ne nous étendrons pas sur les différentes préparations (Bancroft and Gamble 2002).

1.2.3 Protocole 2. Ponçage de rondelles de bois

La préparation de rondelles de bois adaptées pour la numérisation est d'un grand intérêt pour l'étude de grandes zones du plan ligneux. Avec un tel procédé, il n'y a pas de problème de raccord. La taille d'acquisition est limitée par la taille du plateau de la polisseuse et par le déplacement maximal de la platine de 10x10 cm². Pour notre étude nous avons mis en place un protocole dont les grandes lignes sont résumées ci-dessous (pour plus d'information se référer à l'article II.AOB'14 §Method).

Les rondelles de bois prélevées au centre des unités de croissance présentent une hauteur maximum de 2,5 cm afin de passer sous le microscope. Séchées pendant 3 mois à l'abri du soleil et de l'humidité elles sont ensuite poncées et polies à l'aide d'une polisseuse semi-automatique (cf. figure 8) afin d'obtenir deux faces parfaitement parallèles et une face parfaitement polie sans destruction des structures pour la numérisation. La fragilité de la surface, notamment sa forte sensibilité à la poussière, nécessite des précautions particulières lors de la manipulation et de la conservation des rondelles. Dès le polissage effectué, il est nécessaire de réaliser la numérisation de la rondelle, car, sans contrôle très précis de l'hygrométrie et de la température de la zone de stockage, la rondelle se dessèche et sa surface se déforme. Ces micros déformations compromettent la netteté des images produites lors de la phase de numérisation.

1.2.4 La numérisation des échantillons

La numérisation consiste à produire une série ou succession d'images, communément appelée mosaïque d'images couvrant la zone d'observation.

Les plans transversaux sont numérisés par une caméra LCD Olympus DP71 montée sur un Microscope Olympus BX51 et équipée d'une platine mobile permettant par déplacement progressif de la lame de créer une mosaïque d'images.

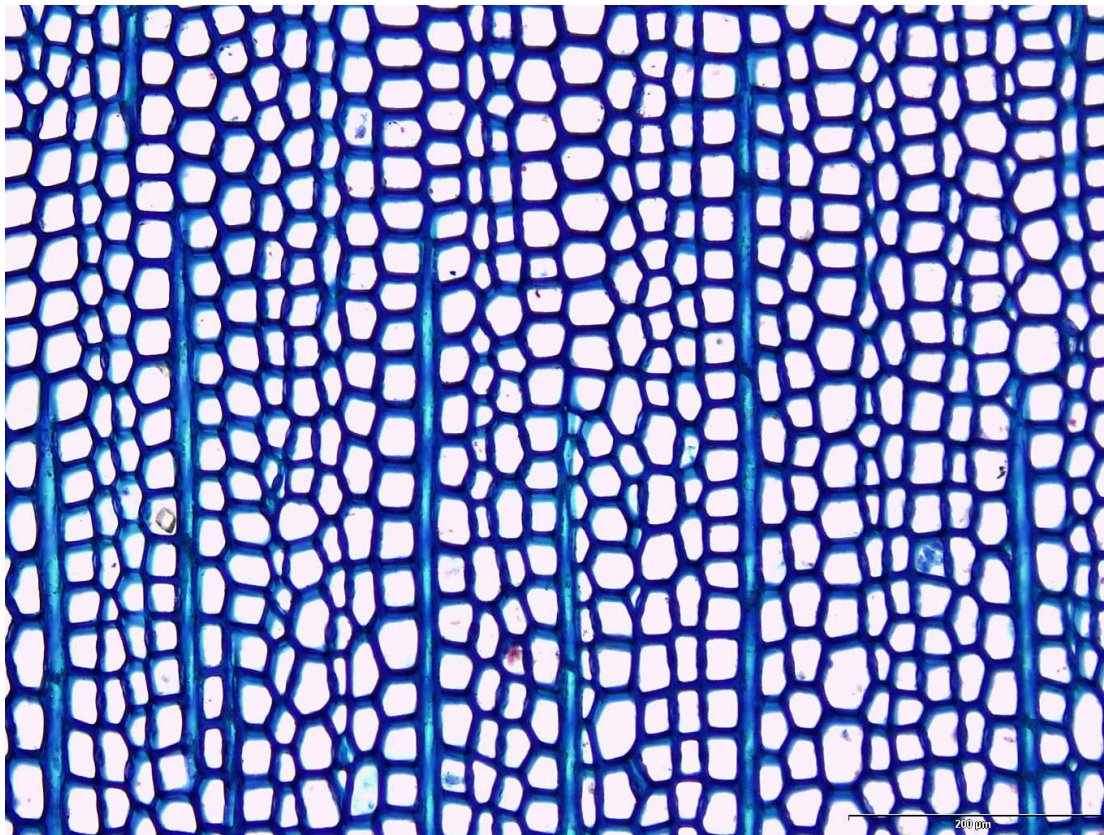
Un grossissement x200 est utilisé pour avoir des surfaces de cellules significatives sur des zones d'observations suffisamment grandes pour permettre le recalage des images avoisinantes.

La position de l'image dans la mosaïque est donnée par le déplacement en X et Y de la platine. Ici, la zone maximale d'observation est de 10x10 cm², soit une mosaïque d'environ 1500 images pour une zone d'environ 10x10 cm². Chaque image est codée dans un système RGB (24 bits) et a une définition de 1600x1200 pixels. Les images peuvent présenter un flou local ou périphérique lié à la déformation optique de la lentille du microscope ou à des variations sensibles d'épaisseur

de l'échantillon et ce malgré la précision du protocole de préparation. Ces variations peuvent provenir de l'utilisation du ruban double-face utilisé pour solidariser la rondelle à la platine lors du polissage ou d'un effet de contrepoids qui fait que la platine ne tourne pas parfaitement sur un plan orbital mais oscille légèrement.

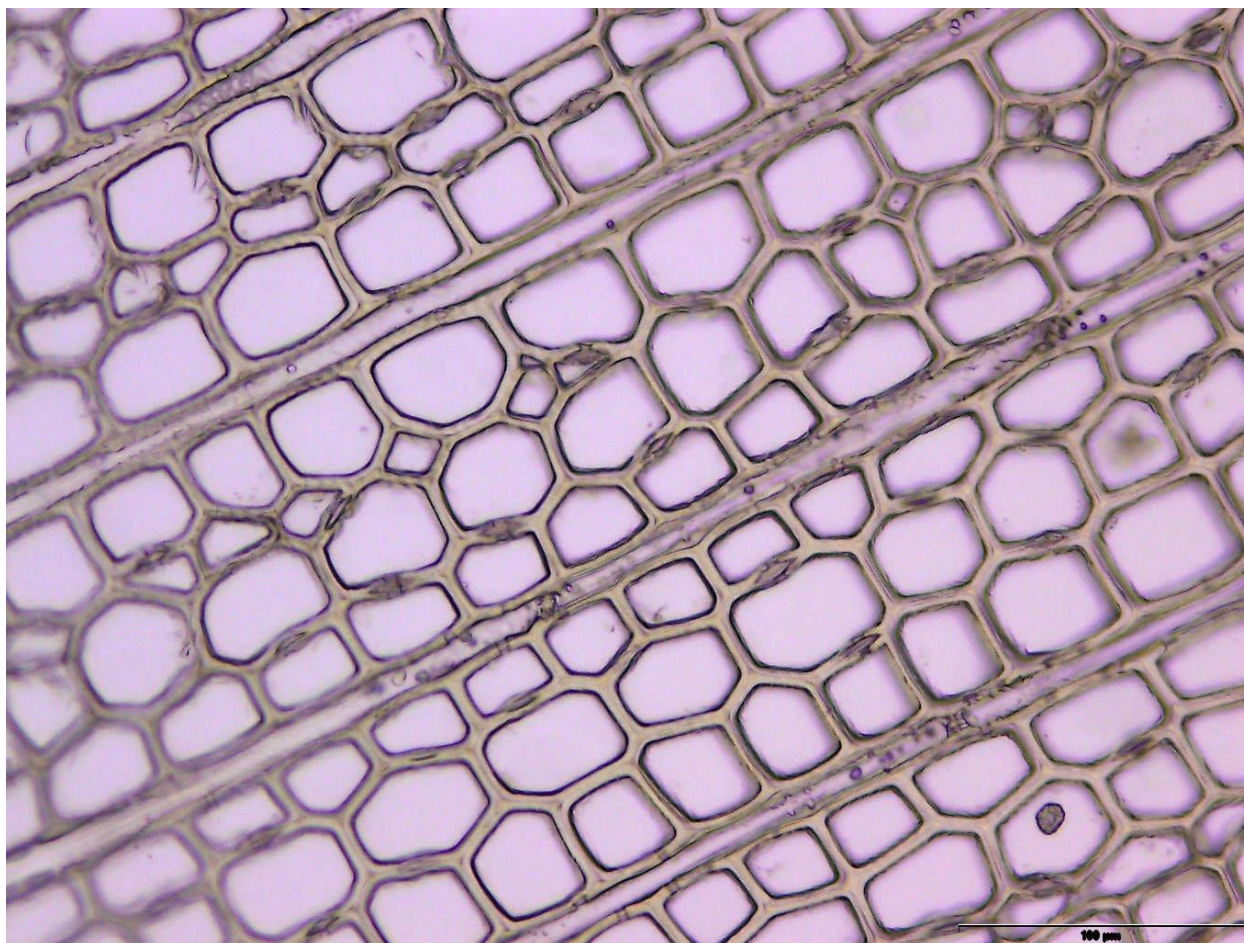
1.3 Résultats des acquisitions & discussions

Toutes les images sont produites en couleur (24 bits) avec une résolution de 1600x1200 pixels. Ci-dessous un exemple de quatre images acquises avec les différents protocoles de préparation et de numérisation.



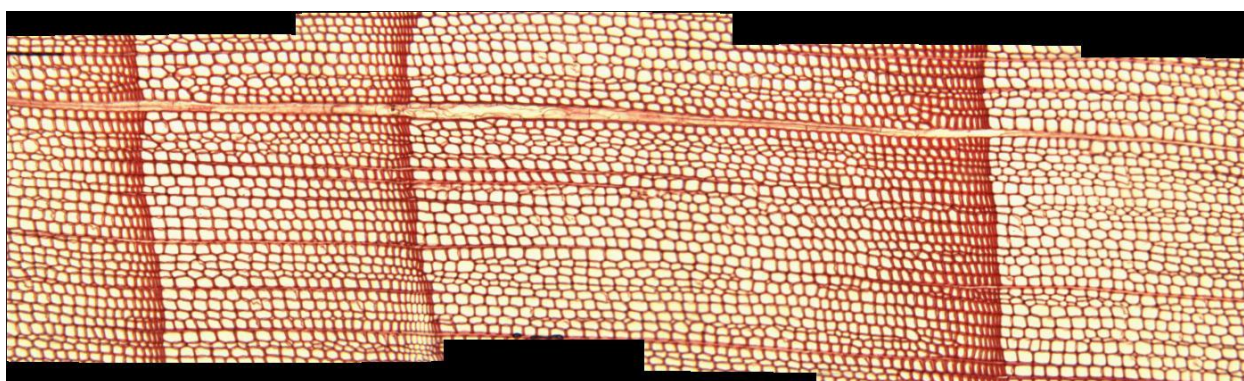
Cliché de l'UMR AMAP, 2010

Figure 9. Coupe histologique transversale de sapin (*Abies alba*) colorée au bleu de toluidine au grossissement x200. Le colorant se fixe sur les tissus celluloseux, donnant la couleur aux parois des cellules. Les lumens sont vides et apparaissent donc en blanc (lumière du microscope).



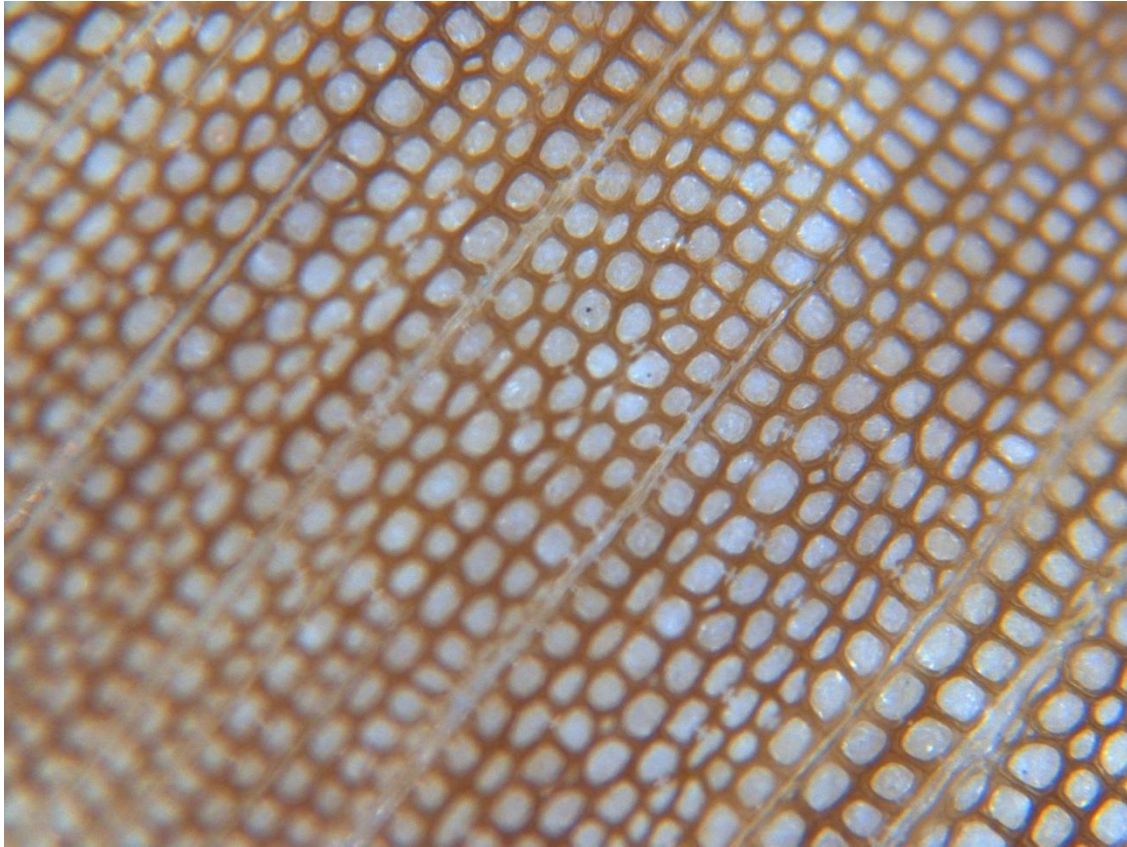
Cliché de l'UMR AMAP, 2010

Figure 10. Coupe histologique sans coloration de sapin (*Abies alba*) au grossissement x400. Le fond des lumens est dans des tons roses, car la balance des blancs n'a pas été effectuée avant l'acquisition de l'image. Les liserés noirs sur les contours de l'image sont dues à des effets optiques.



Cliché de l'auteur, 2013

Figure 11. Mosaïque de 4 coupes histologiques transversales de pin (*Pinus nigra*) colorées à la safranine au grossissement x200. Les images ont été raboutées à l'aide du logiciel libre Hugin.

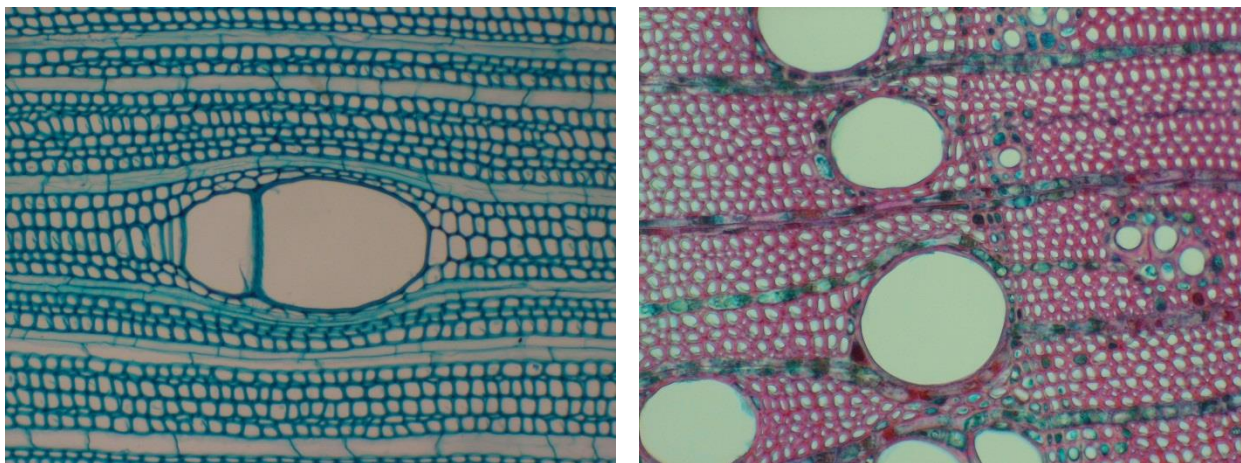


Cliché de l'auteur, 2014

Figure 12. Acquisition d'une image d'une rondelle poncée non colorée de pin (*Pinus nigra*) au grossissement x200 avec un éclairage épiscopique. Les lumens présentent une texture car on voit à travers le reste de matériel biologique et des résidus du ponçage. En périphérie des images, il peut y avoir un flou important provenant de la non planéité d'une des faces de la rondelle. Le contraste entre le lumen et la paroi est moins marqué que sur les coupes histologiques.

Les images présentent une très grande variabilité due à différents facteurs :

- La préparation : selon ce que veut observer le biologiste et selon le protocole d'étude, la coloration des tissus biologiques sera différente.
- Le grossissement : il dépend du champ d'étude que veut obtenir le biologiste. Pour observer un petit élément structurant, le champ pourra être faible pour pouvoir le comparer avec d'autres éléments, ou très fort pour le voir en détail.
- L'essence : selon les différentes essences, l'organisation cellulaire est différente. On peut ainsi différencier les résineux qui ont une organisation cellulaire simple, des feuillus qui ont une organisation plus complexe. Il existe aussi une forte variabilité au sein d'un même grand ensemble d'essences. Par exemple chez l'acajou (*Picnanthus sp.*) (figure 13, à gauche), les vaisseaux apparaissent comme une déformation continue des cellules au sein d'une file, alors que pour le frêne (*Fraxinus angustifolia*) (figure 13, à droite), ces derniers semblent apparaître anarchiquement entre les files.



Cliché de l'UMR AMAP, 2010

Figure 13. A droite, coupe histologique d'acajou (*Picnanthus sp.*). L'apparition des éléments de vaisseaux semble se faire graduellement, on arrive à voir à quelle file ils appartiennent. A droite, coupe histologique de frêne (*Fraxinus angustifolia*). Les vaisseaux semblent apparaître au milieu des files, on ne peut pas distinguer leur file d'appartenance.

L'hétérogénéité des protocoles, des essences et des éléments structurants au sein même des essences doit pouvoir être traitée par la méthode de détection des files cellulaires.

Le protocole de ponçage doit être repris et complété. Un échantillon de bois poncé a été produit par le constructeur de la polisseuse comme étalon. Actuellement, les images produites au laboratoire ne sont pas de qualité égale à celle du constructeur, rendant la détection des files moins fiable. La difficulté majeure du protocole est de fournir un échantillon où les deux faces sont parfaitement parallèles et planes. Les micro-déformations de la rondelle vont produire un flou pouvant être très important pendant la numérisation. La surface polie est très fragile, elle demande une précaution importante et ne doit pas être touchée sous peine d'être altérée. Les rondelles polies doivent être conservées dans un milieu contrôlé pour atténuer les déformations. Ainsi, nous avons opté pour une numérisation complète du plan directement après le polissage sans conservation des échantillons. Le protocole est théoriquement reproductible, mais les interventions opérateurs et les conditions de préparations (humidité, température...) ne permettent pas d'assurer une reproductibilité parfaite des acquisitions.

Le chapitre suivant va décrire la méthode mise en œuvre pour traiter automatiquement les images et d'en extraire les cellules et leur organisation. La méthode fut pensée pour autoriser les traitements sur les images des Gymnospermes et des Angiospermes. L'étude principe a été menée sur des images issues de coupes histologiques, car le protocole de ponçage a fini d'être mis au point en début de 3^{ème} année de thèse. Néanmoins des tests ont toujours été menés en parallèle pour s'assurer que la méthode soit transposable aux images issues du protocole de ponçage.

2 METHODES

2.1 Segmentation des images

La segmentation a pour but l'extraction d'objets structurants, ici des objets biologiques élémentaires de l'image, comme par exemple les cellules. La segmentation d'image définit un bloc de traitement, c'est un processus de rassemblement de pixels entre eux suivant des critères prédéfinis, qui va former un partitionnement de l'image. Ces caractéristiques peuvent être naturelles comme la colorimétrie ou plus complexe comme des critères géométriques (Shapiro 2001 ; Barghout and Lawrence 2003). Le but de la segmentation est de simplifier et/ou modifier la représentation d'une image pour la rendre plus lisible et facile à analyser. Les pixels ayant des critères semblables sont ainsi regroupés en régions, qui constituent un pavage ou une partition de l'image. Il peut s'agir par exemple de séparer les objets du fond ou des objets entre eux.

N.B. Les différentes étapes du processus sont illustrées, ici, à partir d'une portion d'une image de Sapin (*Abies alba*) colorée au bleu de toluidine et numérisée avec un grossissement de x200.

2.1.1 Généralités sur les méthodes de segmentation

À ce jour, il existe de nombreuses méthodes de segmentation, que l'on peut regrouper en trois catégories (Davies 2005) :

1. **La segmentation par régions** (en anglais : region-based segmentation) correspond aux algorithmes d'accroissement de régions. Ce type de segmentation consiste à prendre un ensemble de petites régions uniformes dans l'image et à regrouper les régions adjacentes de même critère (couleur, intensité, critères géométriques...) jusqu'à ce qu'aucun regroupement ne soit plus possible. On y trouve par exemple les algorithmes de *region-growing* (Verma et al. 2011) ou *split and merge* (Damian and Resch 2003).
2. **La segmentation par contours** (en anglais : edge-based segmentation) s'intéresse aux contours des objets dans l'image c'est-à-dire une transition détectable entre deux régions connexes. La transition est souvent un ajustement brutal de l'intensité au niveau des frontières des régions. On y trouve par exemple les algorithmes tels que le filtre de Canny (Canny 1986), le watershed ou encore les *snakes* qui produisent un modèle déformable local attiré ou repoussé par des points particuliers de l'image (maximum ou minimum locaux, ligne de crête...).
3. **La segmentation par classification** : la plus connue (et la plus simple) est sans doute le seuillage qui consiste à binariser l'image en la réduisant à deux régions à partir d'un seuil donné : la première est constituée des points de l'image dont la valeur est inférieure au seuil donné, la seconde des points dont la valeur est supérieure ou égale au seuil donné. Ce mécanisme peut être compliqué à souhait en utilisant simultanément plusieurs seuils (multi-seuillage). Mais dans les faits, le seuillage n'est qu'un cas particulier de

classification : ces méthodes, largement développées pour la fouille de données, sont utilisées autant par les banques pour repérer les « bons » clients sans facteur de risque ("scoring"), que pour l'analyse du génome pour repérer de nouveaux gènes ou qu'en image pour classer les pixels. Elles présentent l'avantage de regrouper les points en régions homogènes (selon un ou plusieurs critères donnés et un estimateur statistique ou mathématique prédéfini). Par exemple, la méthode des k-moyennes (Duda 2001) (ou ses adaptations à la binarisation d'images connues sous les noms de seuillage d'Otsu, Huang...) (Otsu 1979) minimise la variance intra-classe et maximise la variance inter-classe en utilisant le plus souvent un estimateur de distance (euclidienne, de Manhattan, de Mahalanobis...). La différence majeure : la classification est réalisée soit en fonction d'un nombre de classes donné a priori, soit le nombre optimal de classes évalué à partir d'analyses bayésiennes (C-means) (Dunn 1973) ou en fonction de limites préfixées d'étalement des classes (Paiva's classification) (Paiva and Tasdizen 2010). La segmentation de l'image est *de facto* obtenue par la classification des valeurs de ses points.

N.B. : Il existe d'autres types de segmentation, comme par exemple des algorithmes mobilisant les 3 catégories décrites. La profusion des méthodes illustre la difficulté à segmenter des images : il n'y a pas de consensus sur une méthode, car une bonne segmentation dépend du contenu, de la dynamique, du système de description (i.e. le système de couleurs) de l'image et de la nature des zones à segmenter et de ce que l'on cherche à obtenir (Busin et al. 2009).

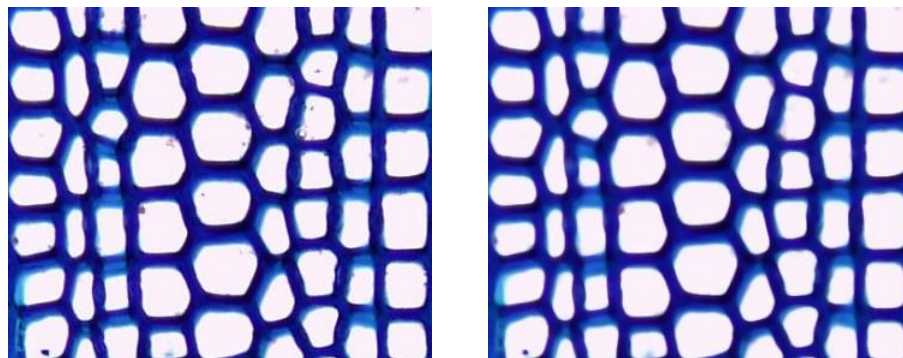
Dans les cellules végétales, la lamelle moyenne est l'espace intercellulaire, c'est-à-dire une séparation physique entre deux cellules adjacentes. Elle contient essentiellement des composés pectiques, ce qui lui donne une intensité différente de coloration par rapport à la paroi cellulaire. La segmentation de ces cellules doit être continue, c'est-à-dire produire des contours fermés et correspondre à la lamelle moyenne. L'algorithme du watershed (Najman and Schmitt 1994) possède toutes les caractéristiques pour produire une bonne segmentation des cellules végétales. Il existe 3 méthodes différentes pour faire un watershed : les algorithmes par inondation simulent une montée progressive du niveau d'eau à partir des minima du relief. Les algorithmes par ruissellement suivent, à partir de chaque pixel de l'image, la ligne de plus grande pente jusqu'à atteindre un minimum. Et finalement, les algorithmes topologiques proposent une déformation progressive du relief, préservant certaines caractéristiques topologiques, jusqu'à ce que le relief soit réduit à une structure fine correspondant à la ligne de partage des eaux. Ces trois méthodes donnent des résultats sensiblement différents, mais toujours avec les mêmes caractéristiques. La définition des images ne permet pas de dire expérimentalement quelle méthode est la meilleure car les différences sont de l'ordre du pixel et on ne peut pas déterminer quels sont le(s) pixel(s) de la lamelle moyenne. Aussi, nous avons choisi l'algorithme par inondation, car son initialisation par minima locaux permet d'envisager de les filtrer (voir plus bas).

Dans les chaînes de traitement d'images, la segmentation est une étape primordiale car elle permet d'extraire les zones d'intérêt dont la qualité influence directement les résultats finaux tels que les mesures des objets. Elle est souvent précédée d'une phase de prétraitement visant à nettoyer l'image (par exemple en supprimant le bruit ou en normalisant la dynamique du signal induit par les capteurs optiques), à rehausser les contrastes en ré-étalant ou compressant la dynamique du signal (par exemple en utilisant des combinaisons de différents lissages de l'image pour accentuer des traits caractéristiques), etc. En raison de la multitude de sources de bruits, et surtout de la multitude d'effets de ces bruits sur une image, il n'existe pas de technique générique, adaptée à toutes les situations. La segmentation est généralement suivie d'une phase de calcul ou de modélisation des objets. Par exemple en imagerie médicale, les neurones sont segmentés avant d'être mesurés.

2.1.2 Chaîne de traitement : de l'image à l'élément structurant

Dans Plant Modeling and Applications (papier I, PMA'12), nous décrivons la méthode de traitement d'image que nous avons mise en place, qui, à partir de l'image native produit une image binaire où les cellules sont détournées par un contour d'épaisseur d'un pixel. Les valeurs des paramètres sont automatiquement évaluées pour s'adapter au mieux à chaque image. La méthode est constituée de 4 étapes.

La première étape consiste à atténuer le bruit impulsif (vraisemblablement induit par un effet thermique de la lampe du microscope) à l'aide d'une combinaison d'un filtrage médian (cf. figure 14) et d'un lissage gaussien.



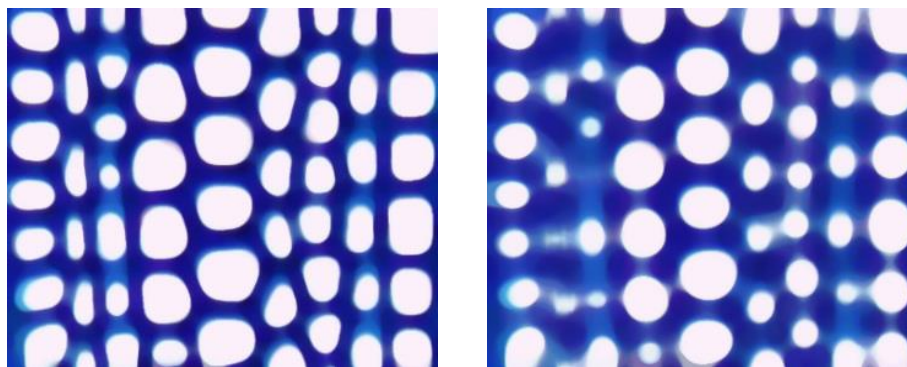


Figure 14. En (a) droite image native. En (b) application d'un filtrage médian avec un rayon de 3, le bruit impulsif est supprimé sans influencer la morphologie des bassins. Respectivement (c) et (d) application d'un médian avec un rayon de 10 et de 20. Avec des rayons élevés, le médian va "lisser" la morphologie des bassins entraînant une modification brutale des structures.

La seconde étape consiste à rehausser le contraste entre les différentes zones de l'image notamment entre parois et lumina des cellules. Nous utilisons la méthode de Différence de Gaussiennes (Einevoll and Plesser 2005) qui consiste à soustraire à l'image faiblement floutée une image fortement floutée. Avec un ratio de variance des Gaussiennes largement supérieur à 1.6 le filtre a un effet passe-bande après soustraction dont les intensités supérieures à 0 sont celles des lumens et les intensités négatives ou nulles celles des parois (cf. figure 15).

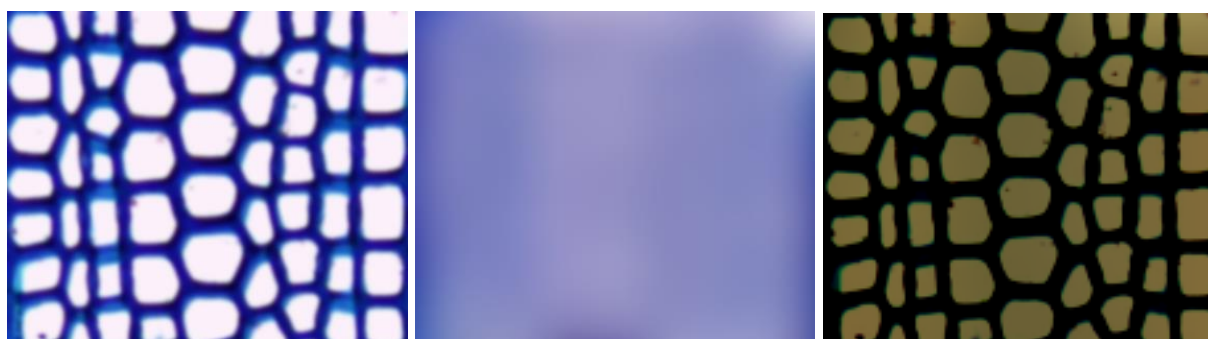


Figure 15. Rehaussement du contraste de l'image par la méthode de Différence de Gaussiennes (DoG). A gauche l'image résultante d'un faible DoG ($\sigma = 3$ pixels). Au centre, l'image résultante d'un fort DoG ($\sigma = (\text{largeur image})/10$). A droite, la soustraction des images A et B qui augmente le contraste lumen/paroi. La valeur de l'intensité des pixels de la paroi est nulle.

La troisième étape consiste à convertir l'image en niveaux de gris par moyenne des canaux car à partir de cette étape la couleur n'est plus significative. Le choix du système couleurs est un aspect important du traitement de l'image. Plusieurs études ont montré que le "meilleur système" dépend du contenu de l'image. Pour notre étude, nous avons choisi de conserver le système RGB (*Red Green Blue*) qui correspond aux longueurs d'onde qui stimulent les trois cônes de l'œil humain. Pour ne pas favoriser un canal, le passage en niveaux de gris est réalisé par moyennage de canaux (cf. figure ci-dessous) et non par la luminance comme généralement utilisée dans la littérature.

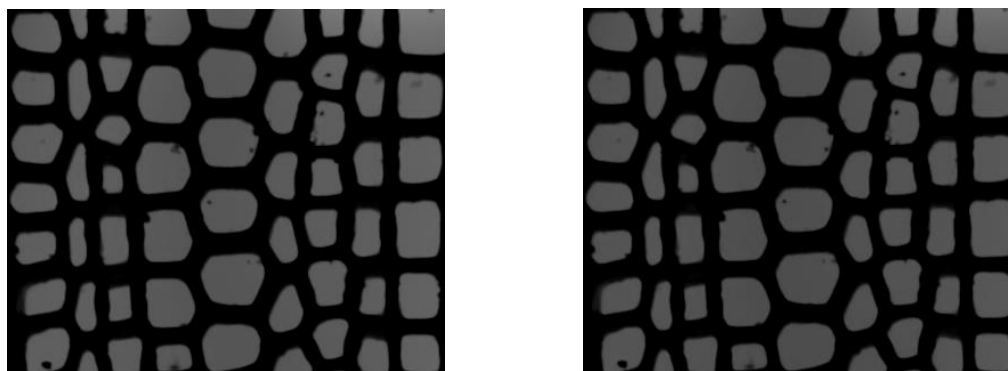


Figure 16. Passage en niveau de gris. A droite, l'image est convertie par moyenne de canaux. A gauche, l'image est convertie à l'aide d'une fonction de la luminance. Les différences ne sont pas perceptibles à l'œil et sont de l'ordre de 1% sur les intensités. L'impact sur la suite du traitement est de l'ordre du pixel, ne permettant pas de définir si une méthode est meilleure que l'autre. Nous avons choisi la moyenne de canaux pour rester cohérent avec le système de couleurs choisi.

La quatrième étape consiste à segmenter et individualiser les cellules à l'aide de l'algorithme du watershed. Avant son application, l'image est inversée (négatif de l'image) pour que les germes du watershed correspondent aux bassins de l'image. Les lignes de crêtes obtenues définissent les régions dont les limites correspondent aux contours des cellules. L'algorithme du watershed est connu pour produire une sur-segmentation qui se traduit par des lignes qui coupent anormalement des cellules. Ce phénomène est dû au fait que chaque minimum local (germe) dans l'image produit un bassin versant. Nous avons implémenté une solution qui consiste à étudier le profil d'intensité de chaque ligne de partage des eaux et à supprimer celles qui présentent une dynamique d'intensité importante, caractéristique des passages paroi-lumen (illustration ci-dessous).

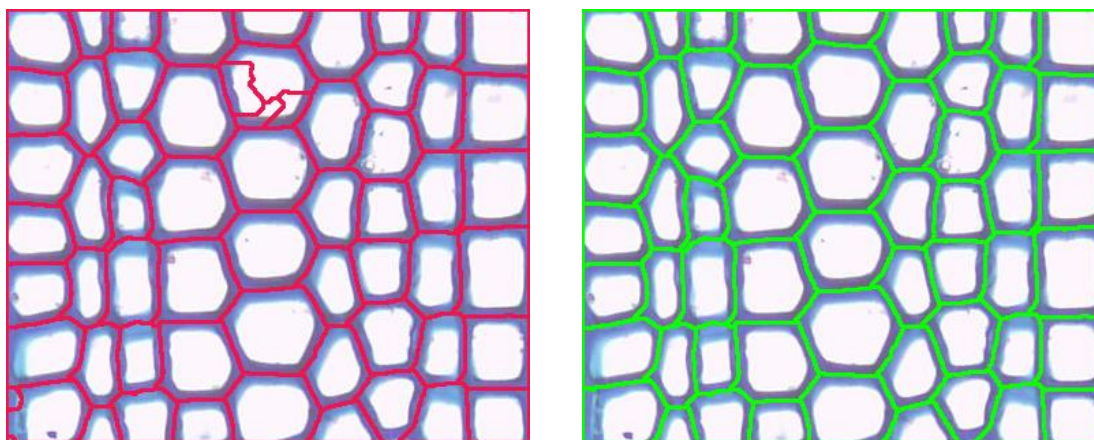


Figure 17. Individualisation des cellules à l'aide de l'algorithme du watershed. Les lignes de bassin versant correspondent à la lamelle moyenne de cellule. Dans la segmentation obtenue (image de gauche), certaines lignes de partage des eaux ne correspondent manifestement pas à l'espace intercellulaire. Des lignes fragmentent l'intérieur de certaines cellules. C'est un cas de sur-segmentation due à la présence de plusieurs minima locaux au sein d'une même cellule. Il s'agit vraisemblablement de bruit atypique qui n'a

pu être supprimé par le lissage appliqué globalement à l'image. Suite à l'étude du profil d'intensité de chaque ligne permet de supprimer la sur-segmentation comme illustré sur l'image de droite. A noter que les lignes sont artificiellement grossies pour accroître la visibilité.

2.1.3 Application de la segmentation aux cellules

Comme nous venons de le voir, le nettoyage du watershed permet d'éliminer les arêtes surnuméraires. La méthode proposée consiste à analyser le profil d'intensité de chaque ligne du watershed pour vérifier son homogénéité. Cette méthode convient bien aux images de files cellulaires car elles présentent un contraste marqué entre le lumen et la paroi. Mais pour d'autres images, cette méthode ne sera pas suffisante. Il existe des variantes du watershed permettant à la construction des lignes d'atténuer voire de supprimer la sur-segmentation. On peut citer la méthode du waterfall (Beucher 2012) qui va fusionner les bassins adjacents reversants. La difficulté de cette méthode est de déterminer d'une part les critères de fusion idéale et d'autre part les conditions d'arrêt, car si les critères permettent une fusion à chaque étape, il n'y aura à la fin qu'un bassin résultant comprenant toute l'image. Les études que nous avons effectuées sur le waterfall et sur le filtrage des minima locaux (germes) du watershed n'ont pas permis d'obtenir de meilleurs résultats que la méthode actuelle. Cela est dû à la complexité des critères et au nombre de niveaux de fusion. Nous pensons cependant que l'utilisation de variantes de watershed telles que le waterfall ou le P-algorithme (Beucher and Marcotegui 2009) serait une alternative satisfaisante pour la suppression de la sur-segmentation.

Une étude a été menée sur l'effet du flou optique souvent présent dans les images du fait des déformations périphériques dues à la lentille du microscope ou au défaut de planéité de l'échantillon. Des acquisitions d'une même zone de l'image avec des mises au point différentes ont été réalisées pour évaluer la stabilité algorithmique de notre méthode. Nous avons conduit une étude statistique sur la variabilité du détourage des cellules. Nous avons mis en évidence que les lignes intercellulaires obtenues par l'algorithme de partage des eaux sont globalement invariantes au floutage de l'image. En effet, elles correspondent aux courbes d'inversion des pentes d'intensité : le lissage de l'image produit par le flou optique réduit la dynamique d'intensités de l'image sans toutefois supprimer les changements de variation d'intensité et les lignes de crête restent globalement inchangées (cf. figure 18).

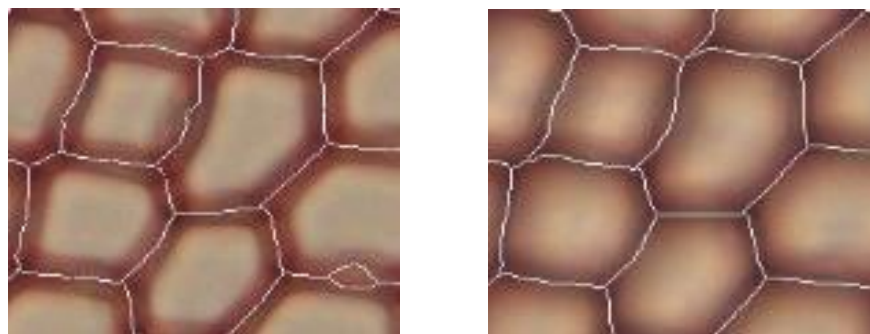


Figure 18. Résultat du watershed sur des coupes transversales de pin colorées à la safranine. A gauche, les lignes de crêtes sur une image nette. A droite, les lignes de crêtes sur une image floutée optiquement. Le contour des bassins est quasiment identique entre une image nette et une image floue. Sur l'image nette, on peut apercevoir un bassin surnuméraire dû à un décollement de la paroi lors de la coupe. Ce bassin sera enlevé lors de la phase de suppression de la sur-segmentation au profit de la cellule du haut.

La figure 19 montre la comparaison d'une soixantaine d'aires de bassins normalisés obtenues à partir d'une image faiblement et fortement floutée. Le coefficient de détermination tend vers 1 montrant que les bassins sont bien corrélés. La pente de la droite est légèrement inférieure à 1 indiquant une faible sous-évaluation sur les images floues. Le procédé donne des résultats quasiment identiques quelle que soit la netteté de l'image.

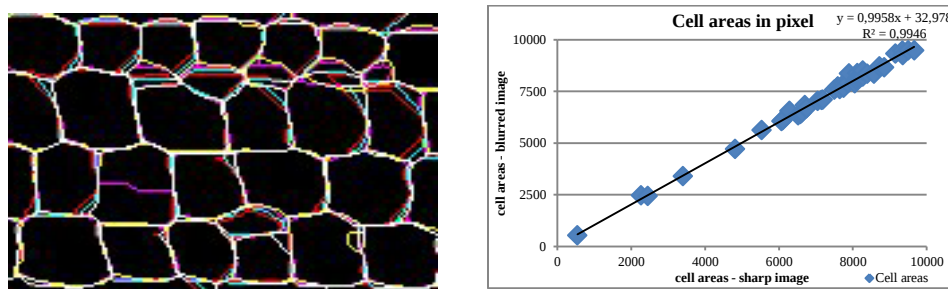


Figure 19. A gauche, étude de la limite des bassins par la superposition des lignes de crêtes produite par un watershed appliqué à une image progressivement floutée (avec un rayon de gaussienne allant de 1 à 15). Les lignes de crête sont affichées en différentes couleurs et en cas de stricte superposition ces couleurs sont additionnées ; les points communs à toutes les lignes de crêtes sont affichés en blanc ; les différences majeures sont dues à la sur-segmentation de l'algorithme du watershed. A droite, étude de la surface des bassins. En abscisse, l'aire des bassins obtenus à partir d'une image nette. En ordonnée, l'estimation de l'aire des bassins à partir d'une image artificiellement floutée. Le coefficient de détermination tend vers 1 montrant que les bassins sont bien corrélés. La pente de la droite de régression est faiblement inférieure à 1 indiquant probablement une faible sous-évaluation des zones de bassins dans les images floues.

2.1.4 Bilan de mise en œuvre de la segmentation

La segmentation permet l'extraction des éléments structurants de l'image. La méthode proposée est composée de deux étapes : une phase de préparation de l'image et une phase de segmentation de l'image.

La première étape consiste à nettoyer l'image de ces défauts (réduire le bruit impulsionnel, etc.) et à augmenter le contraste entre les zones à extraire. Il existe une multitude de filtres et de combinaisons de filtres pour réaliser cette étape. Nous proposons une chaîne de (pré)traitement avec des filtres standard peu coûteux en temps de calcul et donnant des résultats satisfaisants. Selon le contenu et la nature des images, chaque étape du (pré)traitement peut être revue et complétée par des opérateurs plus adaptés : ainsi dans certains cas le filtre médian peut avantageusement être remplacé par un filtre MeanShift ; ce dernier remplace l'intensité de chaque point de l'image par celle du - du maximum local le plus proche (et non pas par celle de la valeur médiane de l'ensemble des pixels de son voisinage). Pour notre étude nous n'avons pas retenu ce filtre car il est très coûteux en temps de calcul et n'apporte pas un gain suffisant par rapport au filtre médian.

La seconde étape consiste à individualiser les éléments à l'aide d'un opérateur de segmentation. Dans le cas des files cellulaires en microscopie optique, la segmentation de ces cellules doit être continue, c'est-à-dire produire des contours fermés, elle doit correspondre à la lamelle moyenne (structure biologique séparant deux cellules) et doit être insensible au flou (présent en périphérie des images). Pour ces raisons, nous avons choisi d'utiliser l'algorithme du watershed implémenté par inondation, il réalise une segmentation des cellules satisfaisante : seules les cellules comportant des inclusions, des décollements de paroi ou étant partiellement obstruées posent problème.

A cette étape chaque cellule est individualisée mais il n'y a aucune relation entre elles. La prochaine étape du traitement va permettre d'introduire les relations entre les cellules et d'identifier automatiquement les organisations.

2.2 Modélisation des arrangements

2.2.1 Définition de la modélisation dans le cadre des files cellulaires

Cette section décrit la conception d'un modèle visant à définir les objets d'étude et les règles permettant d'identifier et de caractériser l'organisation des éléments structuraux précédemment segmentés à un niveau d'observation donné. En général, la modélisation consiste le plus souvent à définir des mécanismes complexes, qu'ils soient géométriques, physiques ou mathématiques dont le but est de reproduire au mieux un comportement naturel expérimentalement observé et foncièrement admis (ou assimilé). Le modèle ainsi conçu permettra de générer des instances représentant potentiellement des expressions de la réalité (i.e. des réalisations « virtuelles » présentant l'ensemble des caractéristiques des échantillons observés). Ces instances seront utilisées par confrontation avec les observations annotées par les experts pour valider ou invalider la cohérence et l'exactitude du modèle. Une fois validé, le modèle permet de proposer des outils de mesure et de caractérisation des productions du modèle.

L'étude porte sur la caractérisation de l'organisation cellulaire dans le plan ligneux. Ces organisations peuvent être de différents types : alignement pour les files cellulaires dans le plan ligneux, circulaire pour les anneaux de croissance, en fuseau dans le plan tangentiel vertical pour les rayons ligneux, en spirale pour les cellules de méristème racinaires...

Cela revient donc à chercher des arrangements particuliers de cellules qui suivent un comportement connu a priori : la droite pour les files cellulaires, le cercle pour les anneaux de croissance, la spirale pour les méristèmes racinaires de riz... Ces différents arrangements peuvent être modélisés par des fonctions.

La construction et l'identification demandent de définir des règles permettant de trouver dans le plan ligneux les arrangements locaux qui se rapprochent le plus des formes recherchées, et plus généralement des propriétés (géométriques ou biologiques) attendues sachant que les distributions spatiales des cellules présentent par essence même de nombreuses perturbations rendant cette tâche plus ou moins complexe. Ces perturbations ont plusieurs origines, elles peuvent venir de défauts de préparation (présence d'intrusions ou de déchirures), ou de réactions internes de la plante (e.g. compression des cellules au pourtour des vaisseaux) ou correspondre à un stress ponctuel ou permanent.

Classiquement, un modèle (d'identification des organisations cellulaires) comprend deux niveaux : (i) un niveau d'exploitation exclusive des informations inhérentes à l'image (basée sur l'étude des arrangements des pixels ou des structures élémentaires homogènes de l'image) et (ii) un niveau d'apport d'une connaissance "*a priori*" caractérisant les organisations recherchées. Le modèle doit de plus reposer sur des mécanismes devant être réversibles et robustes.

La robustesse signifie que les organisations doivent être reconnues même en présence de perturbations et la réversibilité signifie que les organisations doivent être reconnues quel que soit le point de départ. A titre d'exemples, pour la spirale, on doit pouvoir partir du centre et arriver à l'extrémité, ou l'inverse ; pour les alignements cela signifie pouvoir partir indifféremment d'un bord de l'image ou de l'autre et avoir le même résultat.

L'exploitation de règles de voisinages dans un graphe d'adjacence va permettre de bâtir les arrangements. Ces règles sont modélisées par des fonctions géométriques génériques. La spécialisation de la méthode se fait par le biais de ces fonctions, dans lesquelles on va injecter des connaissances a priori. Ainsi, une fonction de droite sera utilisée pour les alignements, une fonction de cercle pour les anneaux de croissance, etc. L'analyse du graphe permet d'une part d'extraire, si elle existe, l'orientation des arrangements et d'autre part de parcourir de proche en proche les nœuds pour créer les arrangements.

2.2.2 Processus d'identification des arrangements

Dans notre étude nous nous intéressons aux files cellulaires distribuées selon des alignements rectilignes.

La première étape consiste à extraire les objets structurants à savoir les cellules et ensuite décrire leur topologie. Un graphe d'adjacence des bassins est alors introduit pour décrire les relations de voisinage entre les différents bassins. C'est un graphe simple classiquement défini par un ensemble de sommets et un ensemble d'arêtes. Plus précisément, chaque arête du graphe relie les centres géométriques de deux bassins voisins, i.e. incidents à une même ligne de crête (cf. figure ci-dessous).

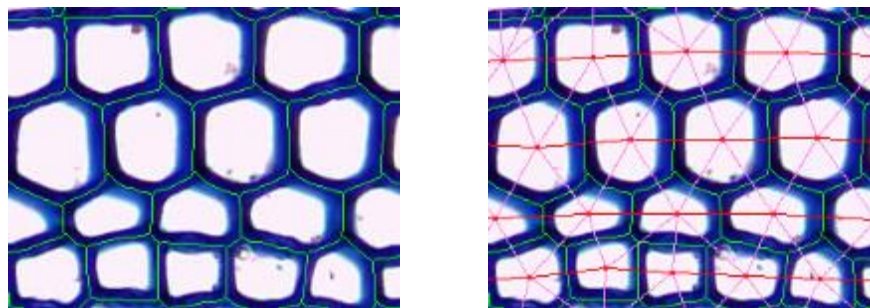
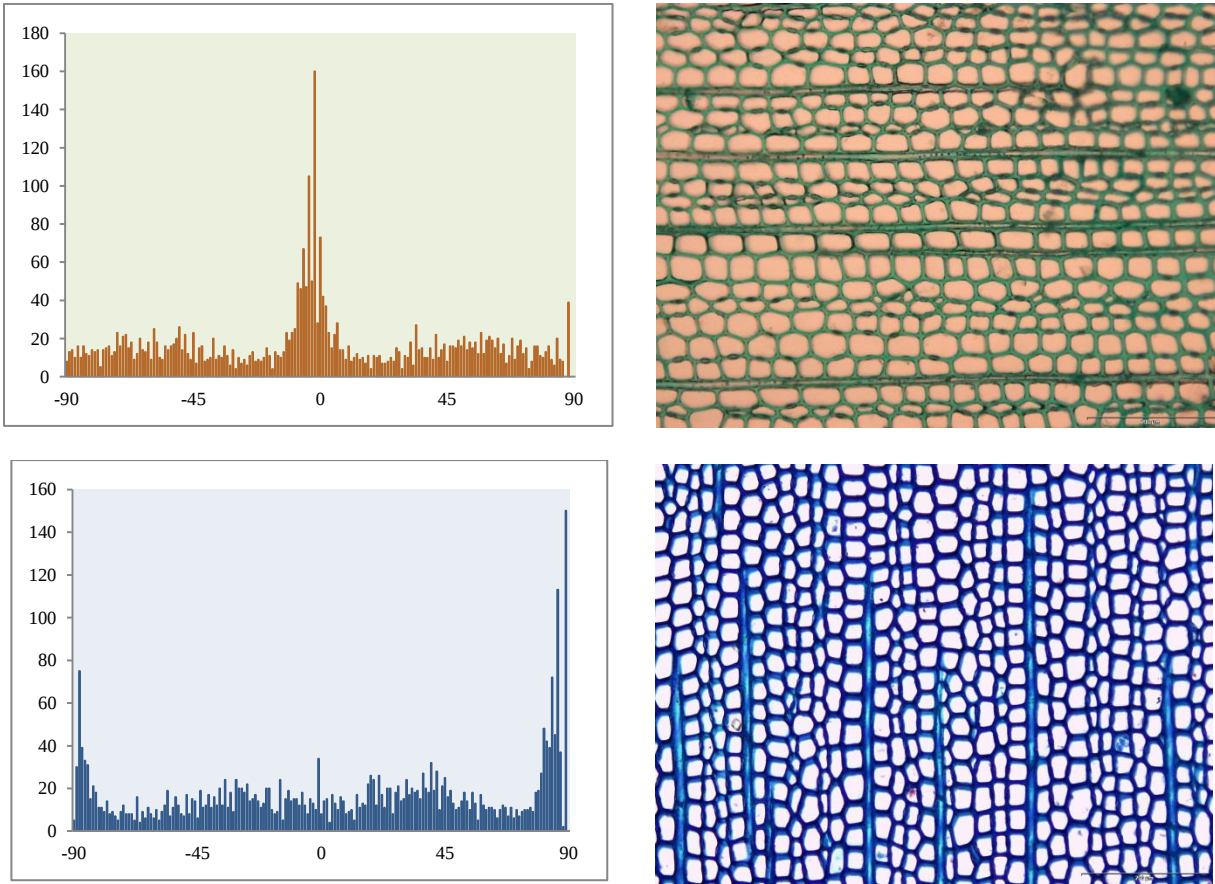


Figure 20. A gauche : les lignes de crête, représentées en vert, produites par le watershed correspondant globalement à la lamelle moyenne, l'espace entre les parois cellulaires ; les bassins sont une bonne approximation des cellules. A droite : le graphe d'adjacence des cellules est construit à partir des lignes de crête ; les chemins correspondant aux files cellulaires sont affichés en rouge.

La seconde étape consiste à trouver l'orientation des lignes. Un histogramme des arêtes du graphe est construit à partir de l'angulation formée par chaque arête avec l'axe des abscisses. Le maximum du profil sera assimilé à l'orientation principale de l'image (Jones and Bischof 1996). Le profil peut aussi permettre de déterminer les orientations secondaires. Du fait de l'agencement

en quinconce des cellules, les arêtes de ce graphe sont orientées selon trois directions : la plus représentée correspondant aux alignements cellulaires. On parlera de direction principale. Le mode de plus grande amplitude indique l'orientation la plus représentée dans le graphe d'adjacence, c'est-à-dire l'orientation principale des files (figure 21).



Clichés de l'UMR AMAP

Figure 21. Histogrammes représentant la distribution de l'angle entre les arêtes du graphe d'adjacence et l'axe des abscisses. A gauche, le mode principal centré sur 0° correspond à des files horizontales. A droite, l'histogramme semble montrer deux modes, un centré sur -90° et l'autre sur 90° ; ces valeurs sont données sur l'intervalle $[-90^\circ, +90^\circ]$, les deux modes ne forment par conséquent qu'un seul mode centré sur 90° (ou -90°), qui correspond à des files verticales.

La troisième étape consiste à bâtir les arrangements, le processus comprend les règles de voisinage sous la forme d'une fonction à minimiser. Pour les files cellulaires, nous proposons une règle géométrique (fonction de droite) et une règle morphologique (fonction de similarité). La fonction de similarité permet de lever les ambiguïtés quand il y a plusieurs bassins candidats (cf. figure 23).

Le processus d'identification des files cellulaires dans une image est progressif et repose sur les principes suivants : les files sont des alignements de cellules deux à deux similaires, i.e. proches

en termes de taille, de forme et les alignements cellulaires sont indépendants de l'orientation de l'image.

La construction des files est progressive et repose sur les 3 étapes suivantes : (i) la recherche d'alignements sous une double contrainte spatio-géométrique, (ii) la gestion des recouvrements et (iii) le chaînage des tronçons isolés.

i. *La recherche d'alignements sous double contrainte spatio-géométrique*

Elle permet d'extraire du graphe les plus longs chemins rectilignes de sommets « géométriquement similaires », i.e. une succession de sommets dont la surface des bassins sous-jacents varie de manière graduelle. Ces alignements sont construits l'un après l'autre, par agrégations successives de sommets. Plus précisément, il s'agit de trouver quel sommet w du graphe doit être rajouté à l'extrémité v de la file en construction pour la compléter.

On définit l'ensemble $N_k(v)$ des voisins de rang k du sommet v , ce qui correspond à tous les sommets du graphe situés à une distance topologique k de v .

Le sommet solution w , s'il existe, est choisi parmi les sommets w_k de l'ensemble $N_k(v)$ et est le sommet géométriquement le plus similaire et le moins éloigné de l'axe de la file. Dans les faits, on ne garde que 2 candidats, les deux plus proches sommets situés de part et d'autre de l'axe formé par le sommet v et le sommet précédent dans la file en construction (illustration figure 22).

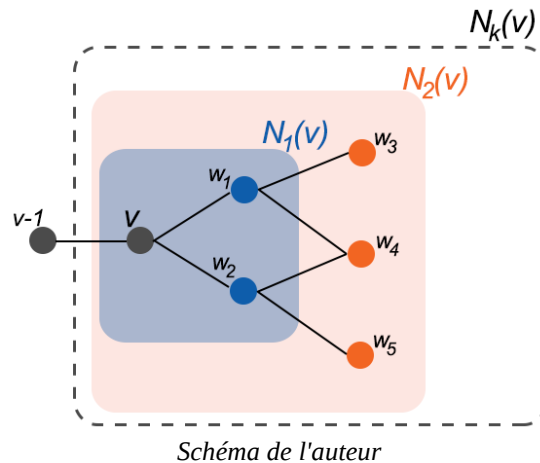


Figure 22. Schéma de la recherche des sommets candidats pour l'agrégation au sommet v courant. L'exploration se fait en premier lieu au voisinage $k=1$ (en bleu). Si un bassin est choisi, il devient le sommet courant et le processus réitère. Sinon la méthode cherche d'autres sommets sur un voisinage étendu $k=2$ (en orange). Le processus de voisinage étendu itère jusqu'à $k=5$. Si aucun sommet n'est choisi le processus s'arrête et la file (ou tronçon de) est créée.

L'indice de similarité géométrique (formule au §2.3.2) permet de choisir le candidat géométrique le plus similaire du sommet v , i.e. dont le bassin sous-jacent est géométriquement le plus proche du bassin sous-jacent au sommet v . Cet indice est bien adapté aux résineux dont la forme et la

surface des trachéides varie peu. Pour les Angiospermes, l'indice est un très bon indicateur de ruptures brutales de taille tout en autorisant des variations continues et progressives de surface.

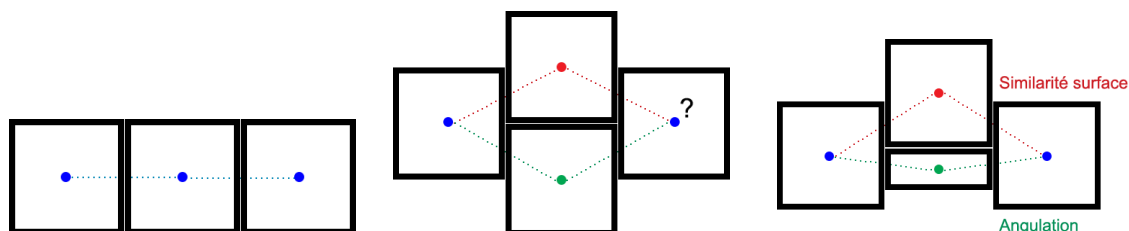


Figure 23. Comment choisir les sommets à rajouter à la file ? A gauche : choix trivial avec une angulation et une similarité identique. Au centre : choix indéterminé, les bassins ayant une angulation et une similarité de surface très proches. A droite : le choix du chemin dépend des critères que l'on veut privilégier.

Si l'indice de similarité géométrique des sommets est supérieur à 0.5, alors aucun candidat du rang considéré n'est retenu ; le processus est répété au rang suivant. La recherche des candidats est ainsi progressivement élargie jusqu'au 5^{ème} rang. Au-delà de cette distance la construction de la file est arrêtée si aucun candidat n'est retenu. La largeur du voisinage étudié est une limite expérimentale, arbitrairement fixée par les anatomistes du bois impliqués sur le projet.

Quand un sommet w est retenu, on vérifie si la réciprocité est vraie, i.e. si, à partir du sommet w , on retombe sur le sommet v courant en utilisant comme direction de référence l'orientation principale des files. S'il ne vérifie pas la condition de réciprocité, la construction de la file s'arrête. Dans le cas contraire, les sommets situés sur le plus court chemin reliant v à w sont rajoutés à la file par l'algorithme de Dijkstra (Cormen et al. 2001) ; dans le cas où w appartient au voisinage de rang 1, aucun sommet intermédiaire n'est rajouté à la file. Enfin le sommet w devient la nouvelle extrémité de la file et le processus de recherche est réitéré.

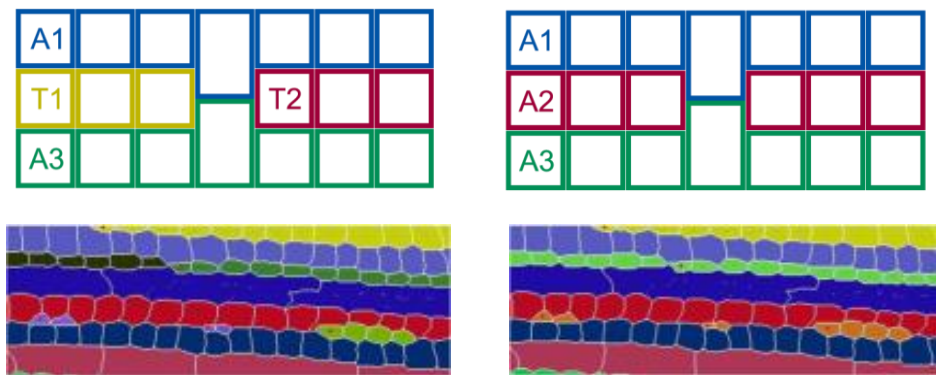
De manière pratique l'angulation maximale entre le sommet v et les sommets candidats v_k est une fonction décroissante qui dépend du rang du voisinage. Elle est de 45° au rang 1 et de 10° à l'ordre 5. Ces seuils ont été fixés expérimentalement d'après les observations faites par les anatomistes du bois. Dans la première étape de construction, l'axe de la file est donné par l'orientation principale des files de l'image.

ii. La gestion des recouvrements

Pendant le processus de construction, la présence de déformations peut donner lieu à des sommets appartenant à plusieurs arrangements. On choisit l'arrangement qui va minimiser la dérive de la file, grâce au résultat de l'algorithme de Dijkstra. Les autres arrangements se verront destitués des sommets en commun.

iii. Le chainage des tronçons isolés

Du fait de la présence d'intrusions, de déformations ou des limites algorithmiques du watershed, le processus peut faire apparaître des tronçons isolés (cf. figure 24). Un tronçon est un arrangement qui ne parcourt pas l'image de part en part. Le principe de la fusion permet de concaténer plusieurs tronçons, en utilisant les règles topologiques simples et ainsi d'établir des lignes entières. L'idée repose sur une double hypothèse : (i) les arrangements traversent l'image de part en part, (ii) et ils ne se croisent pas. Soient deux tronçons T2 et T2' adjacents aux files A1 et A3 ; il est fortement probable que T1 et T2 soient deux parties d'un seul et même arrangement A2. Actuellement seuls les cas typologiquement simples sont traités pour ne pas induire de faux positifs dans la détection.



Schémas et clichés de l'auteur

Figure 24. Chainage des tronçons isolés. En haut, schéma théorique de la fusion des tronçons. Les tronçons T1 et T2 sont compris strictement entre les files A1 et A3, ils seront fusionnés et identifiés comme appartenant à la même file A2. En bas, application sur un cas réel. A gauche, résultat de la chaîne de traitement sans chainage des tronçons. A droite, résultat après application de la méthode de chainage.

Les tronçons sont identifiés comme appartenant à une même file (même couleur).

Cette méthode permet de créer des files ou des tronçons de file sûrs. A noter que des cellules isolées ou des artéfacts (dus à une erreur de segmentation ou à une résolution trop faible de l'image ne permettant pas alors une bonne segmentation) peuvent n'appartenir à aucune file. Nous ne cherchons pas à les raccorder à une file. Ces cas complexes sont laissés à l'appréciation de l'utilisateur.

2.2.3 Mise en œuvre de l'identification des files cellulaires

Dans cette partie nous allons montrer l'application de la chaîne complète de traitement (segmentation et modélisation) et d'identification des files cellulaires (cf. illustration ci-dessous).

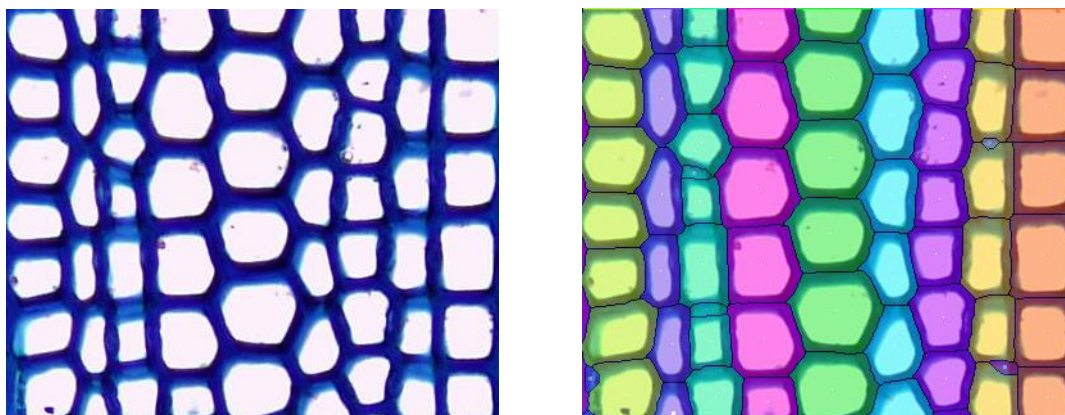


Figure 25. Identification automatique des files de cellulaire avec une coloration aléatoire des files.

La représentation visuelle permet d'identifier facilement chaque file. On peut voir apparaître sur la droite de l'image des petits bassins supplémentaires due à l'épaisseur de la coupe entrainant un écrasement de la paroi. L'algorithme du watershed va identifier ces structures comme des bassins qui ne seront pas supprimés lors de la phase de traitement. Une gestion des cellules peut permettre de fusionner les bassins entre eux.

Nous envisageons la formalisation d'une nouvelle méthode d'identification des files cellulaires indépendante de l'ordre de parcours. La méthode teste toutes les combinaisons à partir d'un nœud fictif de référence (Nœud Père N_p) relié à tous les nœuds du bord de l'image. A partir de N_p , on calcule les scores pour tous les chemins permettant de revenir sur N_p . Les scores sont déterminés par la différence d'angulation entre les arêtes, l'orientation des files et la similarité entre 2 bassins. Une fois tous les parcours réalisés, on dépile les files en fonction des scores, en supprimant au fur et à mesure les trajets sélectionnés ainsi que les trajets contenant au moins un nœud du trajet sélectionné. Cette méthode doit permettre d'identifier toutes les files ainsi que les tronçons en une seule étape.

Nous avons vu que les files sont créées uniquement à partir de règles géométriques et topologiques sur les bassins. Il est important de s'assurer de la robustesse de la détection des bassins, c'est-à-dire que la méthode utilisée soit précise dans le détournage des cellules et insensible au flou de l'image.

Sur l'évaluation de la précision de la mesure, la figure 26 présente la comparaison d'une soixantaine d'aire de bassins normalisés obtenu à partir de mesures manuelles et automatiques. Le coefficient de détermination tend vers 1, montrant que les aires des bassins sont bien corrélées. La pente de la droite de régression est faiblement supérieure à 1 ce qui indique une faible surévaluation de la méthode automatique. La méthode automatisée semble surévaluer les mesures (ou l'expert les sous-évalue). La méthode automatique est cependant répétable.

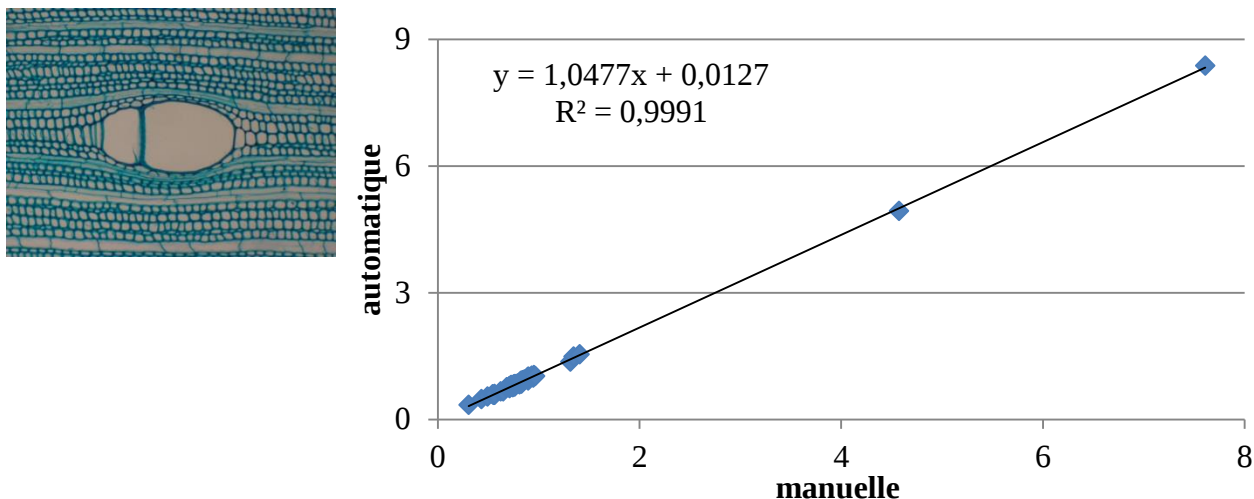


Figure 26. Etude de la surface de bassin chez l'acajou (*Pycnanthus sp.*). En abscisse, les surfaces normalisées obtenues à partir d'une segmentation manuelle. En ordonnée, celles obtenues à partir de la méthode automatique. Le coefficient de détermination est proche de 1, montrant un bon ajustement des surfaces.

Ici, seule la précision géométrique est étudiée : le décalage géométrique n'est volontairement pas pris en compte car estimé négligeable au regard de la surface des cellules. Mais il va de soi que l'étude pourrait être élargie au décalage, notamment pour évaluer le recouvrement géométrique entre annotations expertes et segmentation algorithmique.

2.2.4 Conclusion sur la mise en œuvre de la modélisation

La modélisation des données est indispensable pour l'étude d'éléments possédant une organisation, telle que les files cellulaires. Elle permet de suivre l'évolution des variations au sein de ces organisations. Pour cela nous avons mis en place une méthode de parcours sous contraintes de graphe. Le processus comprend d'une part des règles de voisinage sous la forme d'une fonction paramétrable à minimiser (droite, cercle, spirale...) et d'autre part l'apport expert d'une connaissance "*a priori*" caractérisant les organisations recherchées.

La méthode permet de détecter automatiquement les files cellulaires et les tronçons probablement issus de la même file. Le processus est capable d'absorber des déformations biologiques telles que celles produites par les vaisseaux et de concaténer des portions de files grâce à des règles topologiques simples.

La méthode traite les images à la volée, c'est-à-dire une à une. Pourtant pour réaliser les analyses statistiques, il faut pouvoir suivre les organisations sur une large zone d'étude. Le prochain chapitre décrit la méthode de suivi d'organisations sur le plan ligneux complet.

2.3 Fusion des données

2.3.1 Mosaïques des résultats

L'assemblage d'images (ou stitching) est un procédé consistant à combiner plusieurs images qui se chevauchent partiellement en vue de produire des images de très haute résolution (THR) ou des panoramas. C'est un problème complexe, mais largement traité dans la littérature. La majorité des méthodes d'assemblage d'images décrites nécessite un chevauchement exact entre les images (Brown and Lowe 2007) et des radiométries identiques pour produire des résultats sans déformation périphérique (Vercauteren et al. 2006 ; Clienti and Beucher 2010). D'autres méthodes, telle que SIFT (Scale-Invariant Feature Transform) (Lowe 2004 ; Toews and Wells 2009) ou SURF (Speeded Up Robust Features) (Bay et al. 2006 ; Oyallon and Rabin 2013), permettent de fusionner les images sans connaître a priori leur chevauchement grâce à des points d'intérêt identifiés dans les deux images à apparier. Cette étape est souvent suivie d'un étalonnage des images. Cette opération vise à minimiser les différences entre les images par la correction des défauts optiques tels que les distorsions, l'exposition, le vignettage ou les aberrations chromatiques. Les images THR restent complexes à manipuler du fait de leur taille importante (environ 3 milliards de pixels pour des poids de 30 Go) et ce malgré l'augmentation de la puissance des ordinateurs. En acquisition satellitaire, les images produites en THR sont souvent découpées ou réduites pour être plus facilement traitées. Mais, même avec ce procédé, les images ne peuvent être traitées sur des ordinateurs tous publics.

Pour s'affranchir de ces différentes difficultés techniques, notamment des corrections radiométriques et du redécoupage de l'image THR, nous avons opté pour une approche combinant le traitement à la volée des images composant la mosaïque et la fusion progressive des résultats produits (Wyant and Schmit 1998 ; Sjö dahl and Oreb 2002). Pour faciliter leur manipulation, les résultats sont décrits à l'aide d'un graphe. Un tel modèle est bien adapté aux organisations cellulaires souvent caractérisées par un ensemble de règles connues a priori qui servent à l'identification des structures à fusionner. La fusion des résultats de deux images adjacentes se ramène donc à la fusion de deux graphes.

Dans notre cas, la limitation de la course de la platine du microscope réduit la zone d'observation à une surface de 10x10 cm², décrite par une mosaïque d'environ 1500 images couleurs de 1600x1200 pixels. Les images peuvent montrer un flou local ou périphérique en raison des déformations optiques ou de la non planéité locale des échantillons observés. Par contre, le microscope est équipé d'une platine, dont le positionnement est connu et enregistré à chaque prise de vue.

2.3.2 Présentation d'une méthode générique de fusion de graphe

Nous présentons une méthode générique pour l'agrégation des résultats par la fusion unique de graphes disjoints. La méthode est basée sur une définition s'appliquant à des graphes simples (i.e. non orientés et sans boucle). L'agrégation des (graphes de) résultats est réalisée à partir d'un noyau de fusion, c'est-à-dire à partir d'un ensemble de couples de nœuds à fusionner. Cet ensemble est obtenu à partir d'une fonction bijective définie par un produit de règles de fusion.

Il est rappelé qu'une fonction est bijective si et seulement si tout élément de son ensemble d'arrivée possède un et un seul antécédent, c'est-à-dire qu'elle est image d'exactly un élément de son ensemble de départ.

Dans notre cas d'étude un graphe simple est utilisé pour décrire l'organisation cellulaire. Un graphe est dit simple s'il n'a pas de sommets en double, ni d'hyper-arcs (i.e. des arcs différents joignant les mêmes sommets), ni de boucles. C'est-à-dire qu'il n'existera au plus qu'une seule arête reliant le sommet X au sommet Y . La fusion des graphes se fera par la fusion des sommets communs aux deux graphes, c'est-à-dire des sommets portant les mêmes caractéristiques. Le noyau de fusion est décrit par un produit de termes normalisés décrivant uniquement les aspects géométriques – position et taille – des cellules potentiellement fusionnables (papier IV, SCIA'13). Un sommet n'est retenu que si sa position est dans un cercle dont le rayon correspond à celui du cercle inscrit du bassin. La similarité des surfaces est définie par l'indice de similarité géométrique noté GS . Défini entre deux sommets p et q , il est donné par la différence normalisée des surfaces S_p et S_q .

$$GS_{pq} = \frac{|S_p - S_q|}{(S_p + S_q)}$$

Figure 27. Différence normalisée des surfaces S_p et S_q : ce critère est une adaptation de l'indice de dissimilarité de Bray-Curtis (Bray and Curtis 1957) proposé pour comparer deux échantillons de même taille.

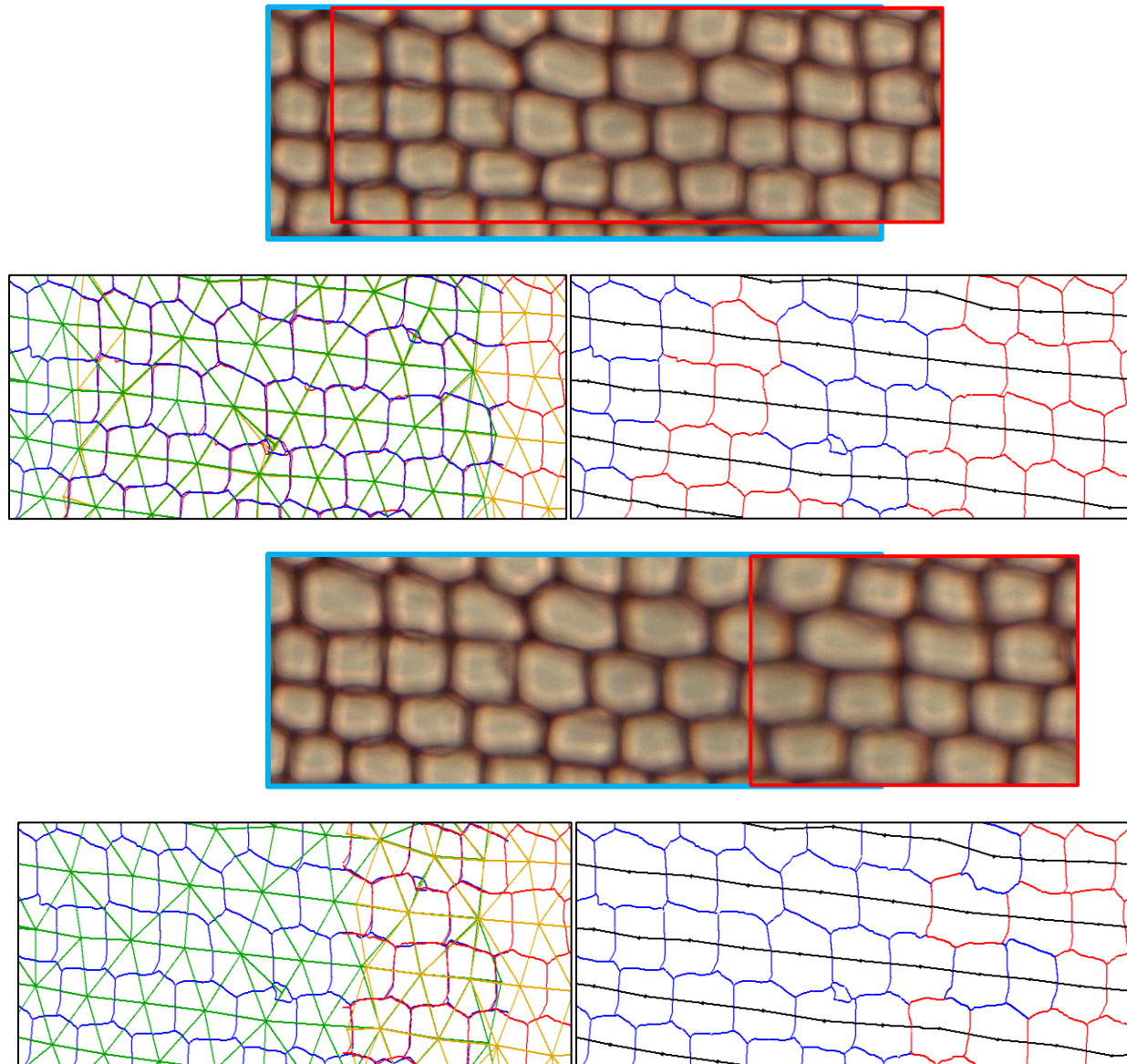
Plus l'indice tend vers 0 (figure ci-dessus), plus les surfaces des cellules comparées sont proches donc similaires ; en pratique, deux sommets sont considérés comme « identiques » dès que leur indice de similarité géométrique est inférieur à 0,1.

N.B. : Les cellules du bord de l'image étant tronquées, elles sont supprimées du graphe avant le processus de fusion.

2.3.3 Agrégation des résultats de files cellulaires

La méthode de fusion est indépendante du taux de recouvrement et du flou des images (cf. figure 28). En effet la méthode de détection du contour des cellules est globalement insensible au flou. Les critères utilisés dans le noyau de fusion (position et taille) sont de ce fait invariants au flou

local de l'image. De plus, la connaissance du déplacement du microscope est un avantage pour identifier les cellules potentielles à fusionner car on définit au préalable un référentiel commun à toutes les images qui correspond aux coordonnées [0,0] de la platine.



Clichés et illustrations de l'auteur

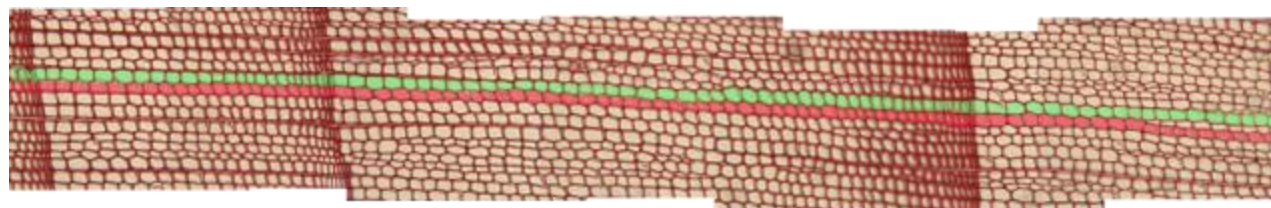
Figure 28. En haut, chevauchement fort de deux images de Pin (*Pinus sp.*). En dessous, superposition des graphes des deux images (en bleu et en rouge) et résultat de la fusion des graphes, les lignes sont en noir.

Au centre, chevauchement faible de deux images de Pin, avec le résultat de la fusion. Les résultats des fusions sont sensiblement identiques : la différence provient des échanges locaux de cellules : lors de la fusion, la cellule retenue est celle la moins floutée.

Le sommet issu de la fusion est représenté par le meilleur candidat des deux sommets, à savoir celui associé à la cellule la moins floue. L'estimateur de flou local, basé sur la relation entre les

dynamiques locales et les différences d'intensité (cf. §2.4.2.2 - *Niveau des composants*), est utilisé pour finaliser le choix.

La fusion des différents graphes d'adjacence des images de la mosaïque permet de suivre les files radiales sur de grandes zones d'observation, ici deux cernes pleins (cf. figure 29).

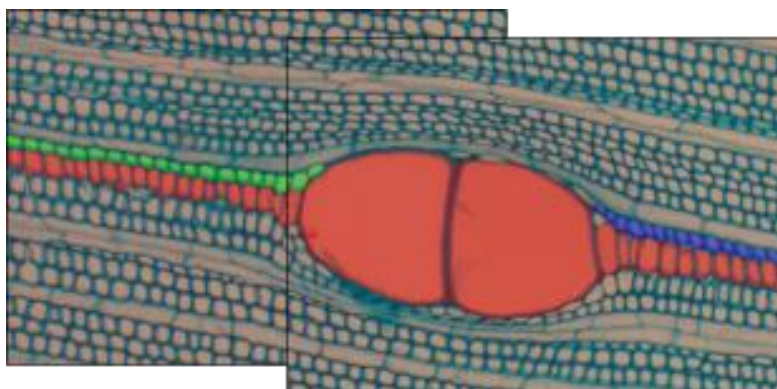


Cliché de l'auteur

Figure 29. Files radiales traversant une mosaïque de 7 images. La méthode ne fusionne pas les images, mais les graphes obtenus pour chaque image. Ici, nous avons fait une recomposition automatique des deux files du graphe mais un pavage manuel des images à des fins de visualisation.

La transition entre deux cernes successifs est caractérisée par une rupture morphologique des cellules. Cette rupture n'affecte pas le processus de fusion des graphes car le processus de fusion des graphes est basé uniquement sur une agrégation locale des cellules périphériques. L'approche est stable en ce qui concerne l'évolution des propriétés des cellules.

Le taux de recouvrement des images doit être au moins d'une cellule. Dans les faits, il est expérimentalement fixé à 3 cellules. Il prend toute son importance dans le cas des Angiospermes qui présentent de fortes singularités du fait de la diversité (de forme et de taille) des différentes structures biologiques (cf. figure 30). La présence de vaisseaux entraîne des déformations cellulaires progressives ou brutales importantes. Les images contiennent des cellules de tailles et de formes très variables : il n'est pas rare de voir des vaisseaux 10 à 20 fois plus gros que des fibres. Dans ce cas-là, le recouvrement des images est délicat : il est actuellement laissé à la discrétion du biologiste lors de l'acquisition.



Cliché de l'auteur

Figure 30. Cas des Angiospermes (ici de l'acajou, *Pycnanthus sp.*) : la fusion des graphes fonctionne dans le cas de recouvrement total du vaisseau. La file centrale (en rouge) est bien assemblée, du fait du recouvrement total des vaisseaux. Mais l'écrasement des fibres produit par le développement des

vaisseaux provoque des discontinuités dans les files périphériques (verte et bleue), que la méthode de fusion ne peut corriger

Il est nécessaire de mettre en place une approche de concaténation de files à partir de l'étude du voisinage des files, similaire à l'approche présentée dans (Article I.PMA'12) et sommairement décrite dans le chapitre §2.2.2 « Processus d'identification des arrangements ».

2.3.4 Conclusion

L'étude des tendances biologiques le long des files est nécessaire pour comprendre la mise en place des différentes organisations et maturations de cellule. Elle ne peut être faite qu'à partir d'une grande zone d'observation du plan ligneux. Les images produites par le microscope sont données sous la forme d'une pile d'images « spatio-localisées » à traiter. Les résultats produits devront être compilés afin de pouvoir suivre les structures sur tout le plan. Pour cela nous avons mis au point une fonction générique d'agrégation de graphe simple, basée sur une involution. Elle permet la fusion à la volée des résultats produits par chaque image. La fonction de fusion doit être spécialisée en fonction de l'objet d'étude. Pour les files cellulaires, cette spécialisation passe par l'ajout de règles géométriques insensibles au flou de l'image. Cette contrainte est d'autant plus importante car les images prises au microscope présentent des déformations périphériques. La méthode permet de reconstituer une mosaïque entière du plan de coupe.

Il pourrait être intéressant de remplacer dans la fonction de fusion les termes de position par un terme relatif à la dynamique de la cellule : ce point nécessite de caractériser le flou local de l'image pour normaliser la signature de chaque cellule. Un tel indicateur est présenté dans la définition des coefficients de pertinence (Article III, FSMP'13), et est utilisé pour valider les résultats quantitatifs produits. L'utilisation de coefficient pour corriger les mesures effectuées est sérieusement envisagée, notamment pour homogénéiser la signature d'intensités des cellules.

Bien que non testée, la méthode de fusion des graphes devrait pouvoir s'étendre aux organisations en spirales et en cercles (figure ci-contre) du moment qu'au moins un des sommets d'extrémité des arcs est commun aux 2 images.

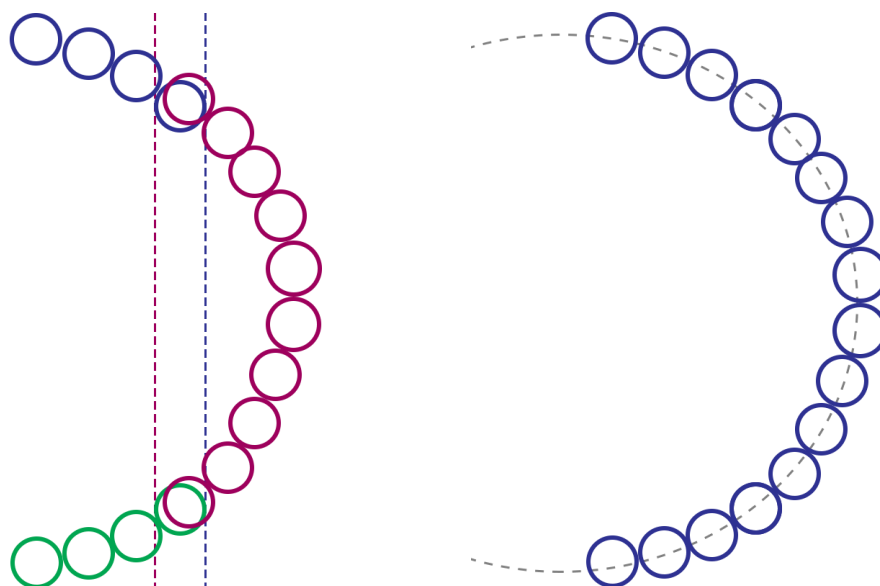


Schéma de l'auteur

Figure 31. Schéma de la fusion théorique d'un cercle ou d'une spirale. A gauche, les arcs bleu et vert appartiennent à la même image et ont un sommet en commun avec l'arc rouge de la seconde image. Les cellules chevauchantes sont décalées pour une meilleure visibilité. Le processus de fusion va tout d'abord fusionner l'arc bleu et l'arc rouge pour ne former qu'une seule structure. Il va ensuite fusionner l'extrémité de l'arc vert avec la structure pour ne former plus qu'un seul arc bleu représenté dans le schéma de droite.

Par définition la méthode de fusion nécessite de tester pour chaque bassin l'ensemble des candidats, ce qui représente une complexité algorithmique en $O(n^2)$. Pour optimiser cette méthode, chaque sommet à traiter est trié selon une double clé sur les coordonnées $[x ; y]$ et mis dans une liste. Les sommets candidats seront les points suivant de la liste où la clé $[x ; y]$ est comprise dans le rayon du cercle inscrit du sommet à traiter. Trouver les sommets candidats revient à parcourir une portion infime de la liste. La complexité de l'algorithme est donc ramenée à $O(n \log(n))$. Au grossissement utilisé, il y a en moyenne 1500 cellules par image, ce qui représente 2 millions de cellules à traiter pour créer la mosaïque entière. Avec un temps d'exécution estimés sur la base d'un accès mémoire de 10 nanosecondes par étape, il faudrait 11,11 heures de calcul pour une complexité en $O(n^2)$ et seulement 3,1 minutes en $O(n \log(n))$. Cette optimisation permet donc de traiter l'ensemble des sommets en un temps raisonnable.

La validation de la méthode nécessite une étude comparative des files identifiées par l'assemblage d'images et par fusion de graphes résultats. Cette étude à venir, devrait permettre de vérifier que les résultats produits sont rigoureusement identiques, et de comparer les temps de calculs associés et les ressources nécessaires.

Suivre une file le long du plan ligneux n'est pas suffisant pour les études statistiques, il faut pouvoir donner des mesures fiables caractérisant aussi bien les files, les cellules que les composants des cellules. Le prochain chapitre sera dédié à l'étude des mesures et à l'importance de la nature de celles-ci.

2.4 Mesures et fiabilités

2.4.1 Résultats numériques

Les utilisateurs attendent un ensemble de résultats numériques caractérisant la forme, la taille, la nature des éléments structurants (Igathinathane et al. 2006) tant pour caractériser que pour typer les structures cellulaires et leur arrangement. En anatomie du bois, ces éléments sont des structures biologiques plus ou moins complexes (parois, lumina, cellules, files...). Leurs caractéristiques sont définies par des paramètres automatiquement évalués.

Notre méthode produit plusieurs types de résultats correspondant à différents niveaux d'observation.

- Niveau global : les files cellulaires sont classées en fonction de leur longueur et de la continuité des hauteurs des cellules les composant ;
- Niveau de l'élément : les cellules sont caractérisées par les dimensions géométriques (diamètre, forme ...) ;
- Niveau des composants : les composants cellulaires (paroi, lumen) sont décrits par des paramètres géométriques (taille, épaisseur, diamètre...), photométriques, d'entropie ou d'homogénéité pour caractériser l'aspect des éléments.

Dans la pratique, seuls les paramètres géométriques sont utilisés car ils décrivent suffisamment bien la cellule, l'apport d'autres informations étant rarement utilisé par les experts.

2.4.1.1 Mesure des paramètres

Pour chaque mesure, il faut trouver et proposer la méthode la plus adaptée à l'objet d'étude. Un ensemble d'outils de calcul est déjà fourni par la plateforme ImageJ (Igathinathane et al. 2008), mais ces outils ne sont pas adaptés à la caractérisation de structures ou de notions complexes.

2.4.1.1.1 La circularité

La circularité (Žunić and Hirota 2008) est un indice normalisé classiquement exprimé par le rapport pondéré entre la surface et le périmètre de la cellule. Elle est calculée à partir de l'expression suivante :

$$\text{circularité} = 4\pi * \frac{[\text{aire}]}{[\text{périmètre}]^2}$$

La circularité est donc donnée par une valeur variant entre 0 et 1, 1 étant la circularité du cercle parfait. Bien que connu, ce paramètre reste délicat à utiliser : il est d'une part difficile de déterminer le seuil significatif au-delà duquel l'objet peut / doit être considéré comme plus rond qu'allongé, et d'autre part d'apprécier l'incidence réelle de la taille de l'objet sur la pertinence de l'indice.

2.4.1.1.2 La surface

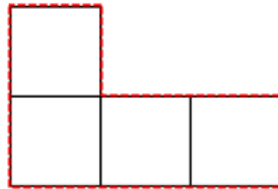
La surface (ou aire) est calculée à partir de la somme des pixels composant la cellule. La mesure est donnée en pixels au carré. Lorsque nécessaire, la calibration de l'image est réalisée à partir des grossissements optiques et numériques utilisés lors de la numérisation.

2.4.1.1.3 Le périmètre

Le périmètre est défini comme une approximation de la longueur réelle du bord de la cellule. C'est un problème complexe largement étudié dans la littérature. Différentes définitions du périmètre de régions discrètes existent (Sabri and Macdonald 2012), l'idée étant de fournir des approximations les plus proches possible des mesures réelles attendues en fonction de la forme des régions traitées (Igathinathane et al. 2008).

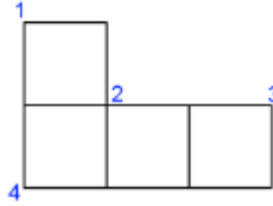
Dans les faits, le périmètre en tant que tel ne présente pas vraiment d'intérêt pour le typage ou la caractérisation des cellules. Il intervient uniquement dans l'évaluation de leur circularité, potentiellement discriminante pour différencier les rayons des cellules. De ce fait, c'est au final plus le comportement que la valeur « vraie » du périmètre qui nous intéresse. A l'instar de l'étude d'Igathinathane, nous avons recherché et comparé trois méthodes de calcul de périmètre simples plutôt adaptées à la description de formes à dominante convexe et ovale.

La première méthode consiste à compter le nombre d'arêtes de contour. Elle se prête bien à l'évaluation des périmètres de régions filiformes, mais moins à des formes "patatoïdales" présentant des cotés obliques pour lesquelles la longueur des pentes du bord, décrites par des marches d'escalier, sont surévaluées.



$$Perim = Nb \text{ arête bordure} = 10$$

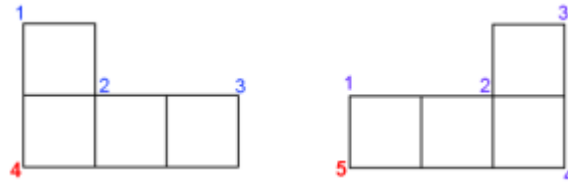
Dans ce cas-là, le périmètre doit être calculé en introduisant la prise en compte des diagonales. Par exemple, l'algorithme implémenté dans ImageJ comptabilise le nombre de coins non adjacents pour moduler le périmètre ; plus précisément le périmètre est défini par le nombre d'arêtes du bord, auquel est retranchée la valeur $(2 - \sqrt{2})$ pour chacun des coins.



$$Perim = Nb \text{ Arête bordure} - Nb \text{ coin non adjacent} * (2 - \sqrt{2})$$

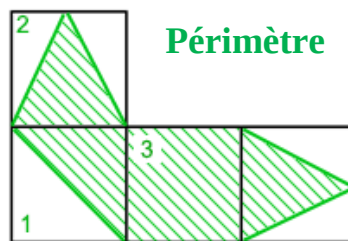
$$Perim = 10 - 4 * (2 - \sqrt{2}) = 7.656$$

Le défaut majeur de cette méthode est que le résultat va dépendre de l'initialisation du point de départ. La figure ci-dessous illustre le problème ; pour le premier cas on va trouver 4 coins et pour le second 5 coins.



On aura au plus un écart maximal de $(2 - \sqrt{2})$ entre deux structures identiques selon la façon de parcourir le contour.

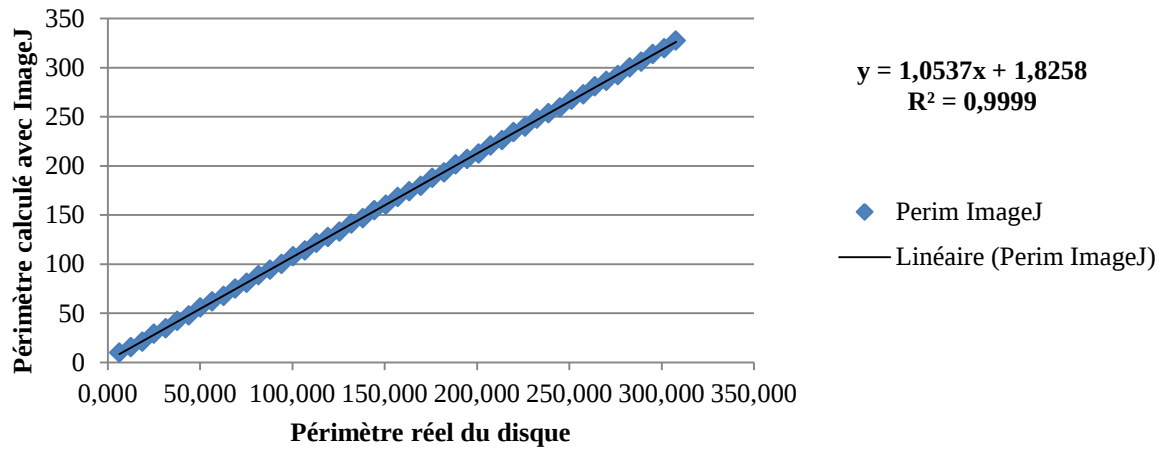
La troisième méthode (que nous appellerons New) consiste à donner un poids à chaque pixel en fonction du nombre de pixels voisin et de la pente à décrire. En 4-connexité, les pixels d'angle (en 1 sur l'illustration ci-dessous) ont un score de racine carré de 2 représentant la pente. Les pixels d'extrémités (en 2 sur l'illustration ci-dessous) ont un score de deux fois racine carré de 1,25 représentant la double pente. Enfin les pixels de contours (en 3) ont un score de 1.



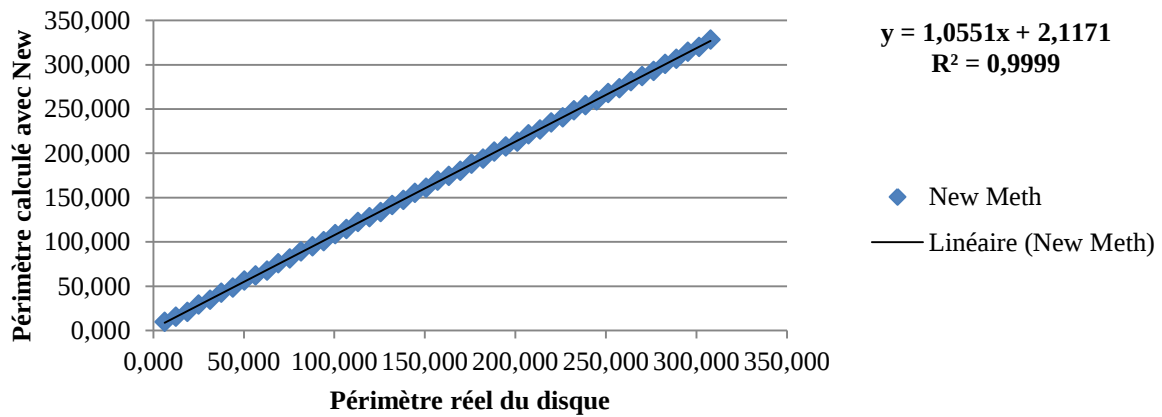
$$Perim = \sum (Pixel \text{ d'angle}) + \sum (Pixel \text{ d'extrémité}) + \sum (Pixel \text{ de contour})$$

Les figures ci-dessous montrent l'évolution des méthodes de calculs par rapport au périmètre attendu réel, avec une variation du rayon du disque de 0 à 50.

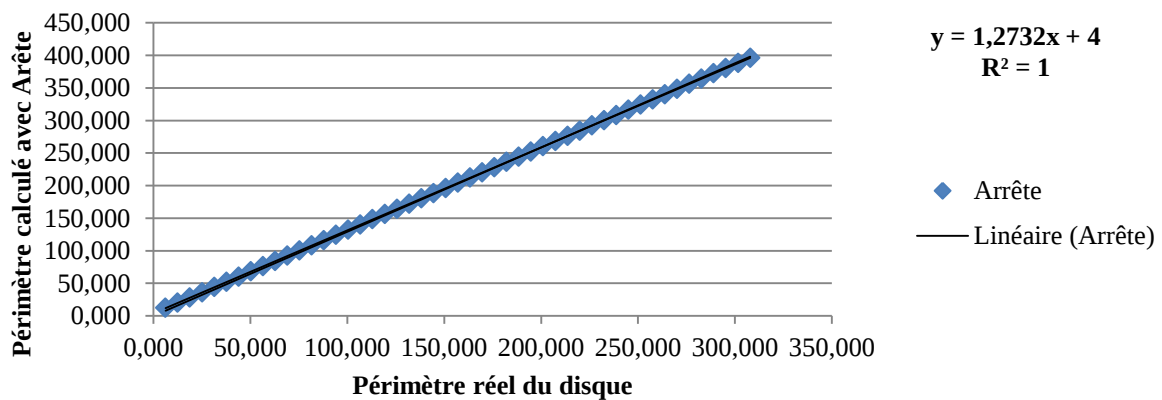
Périmètre Réel / ImageJ



New



Arête



Le coefficient de détermination R^2 représente la fraction de la variance de y "expliquée" par la corrélation de y avec x . Le R^2 est proche de 1, ce qui signifie que les périmètres sont fortement corrélés et que presque 100% de la variance de y est expliquée.

- Courbe 1, on surévalue le périmètre et on a un décalage de 1,8 à l'origine.
- Courbe 2, on surévalue le périmètre et on a un décalage de 2,1 à l'origine.
- Courbe 3, on surévalue le périmètre et on a un décalage de 4 à l'origine.

Avec les trois méthodes on surestime le périmètre du cercle. La méthode d'ImageJ a la pente la plus proche de 1 ($y = 1,0537x$) et le plus faible décalage à l'origine. Elle donne donc une meilleure approximation du périmètre que les autres méthodes.

2.4.1.2 Séparations des composants

Chaque cellule est composée de deux structures anatomiques (lumen et paroi) qu'il convient de différencier avant toute mesure : cette différenciation conditionne la nature des mesures effectuées. Le lumen est plutôt caractérisé par sa surface, la paroi par son épaisseur (Guo et al. 2011). Par ailleurs, la définition de paramètres même apparemment simples n'est pas partagée : l'épaisseur de la paroi en est une très bonne illustration, aucun consensus n'ayant été trouvé auprès de la communauté des anatomistes du bois. Ce paramètre peut être exprimé de différentes manières : une distribution de valeurs, la valeur moyenne ou médiane de ladite distribution, des valeurs prises selon l'orientation de la file ou sa normale.

Pour notre part, les lumens et les parois sont identifiés par répartition des points de la cellule en deux groupes selon leur intensité : d'une part les points clairs correspondant au lumen, et d'autre part les points sombres correspondant à la paroi. La limite sombre / clair est automatiquement évaluée par l'algorithme des 2-moyenne qui présente l'avantage d'être un bon compromis entre efficacité et généralité. En l'absence de formalisation précise, nous avons choisi d'exprimer l'épaisseur de la paroi non pas par une valeur mais par l'ensemble de toutes les épaisseurs mesurables à partir de son contour externe. Cette distribution est calculée à partir d'une carte de distances orientées. La façon dont elle sera utilisée est laissée à la convenance de l'utilisateur.

2.4.2 Fiabilité

Dans le cadre d'un traitement de masse il est intéressant de pouvoir qualifier l'exactitude de ces évaluations. Un indice de certitude est donc attribué à chaque paramètre calculé, qu'il porte sur les cellules, ses constituants ou leurs alignements.

Intuitivement, les mesures seront plus ou moins proches de la valeur réelle en fonction du cumul des erreurs liées à la qualité de l'image, la configuration biologique, les approximations algorithmiques,... Il est donc important de donner à chaque type de résultat une fiabilité estimée.

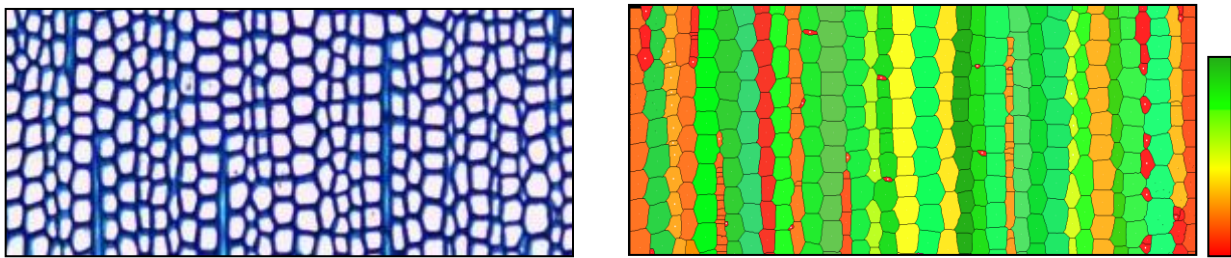
2.4.2.1 Niveau global

Une file f doit être suffisamment longue pour être significative : sa longueur L_f est donc comparée à un seuil L automatiquement obtenu à partir de la distribution des longueurs calculées. De même, une file est normalement composée de cellules proches, en termes de forme, de taille et d'aspect. Pour simplifier, ces trois critères sont ramenés à la seule hauteur des cellules prises en tout point dans la direction normale à la file : cette simplification est une modélisation de la conséquence du mécanisme de maturation des trachéides. Ce mécanisme n'est pas transposable aux feuillus : mais dans la pratique, ce seul critère semble être suffisant pour caractériser la variation continue observée au sein des files cellulaires des Angiospermes. En conséquence, les hauteurs H_j et H_{j+1} des cellules consécutives de la file f sont comparées. Le coefficient de fiabilité R_f de la file f est donc défini par un produit de termes normalisés variant entre 0 à 1.

$$R_f = \left(1 - \max\left(\frac{L - L_f}{L}, 0\right)\right) \prod_{j=0}^{n-1} \left(1 - \frac{|H_j - H_{j+1}|}{H_j + H_{j+1}}\right)$$

Plus le coefficient est proche de 1, plus la file f est dite fiable.

Le coefficient de fiabilité globale permet de caractériser les files, tant en termes d'expression significative que de comportement géométrique : une « bonne » file est une file suffisamment longue et dont la taille des cellules varie de manière graduelle. Ce coefficient permet de filtrer directement les files (cf. figure 32). Les files ayant un bon coefficient sont représentées en vert, celles ayant un mauvais coefficient en rouge. Les teintes rouges-orangées traduisent des configurations présentant des coupures ou de fortes discontinuités. Cet estimateur est bien adapté pour les résineux présentant des arrangements cellulaires très organisés et continus. Pour les Angiospermes, ce coefficient permet de visualiser rapidement les éléments de vaisseaux et les files adjacentes. Ces files sont intéressantes pour l'analyse des contraintes géométriques endogènes des files causées par les vaisseaux.



Cliché de l'auteur

Figure 32. Le coefficient de fiabilité R_f permet un affichage intuitif de la fiabilité des files cellulaires : les files qui sont moins fiables (discontinues ou hétérogènes) apparaissent dans des tons chauds (jaune à rouge).

Cet indicateur est finalement une note très intégrative des files identifiées. Plus complets que les indices de performance évoqués en première partie des résultats, ils sont également plus délicats

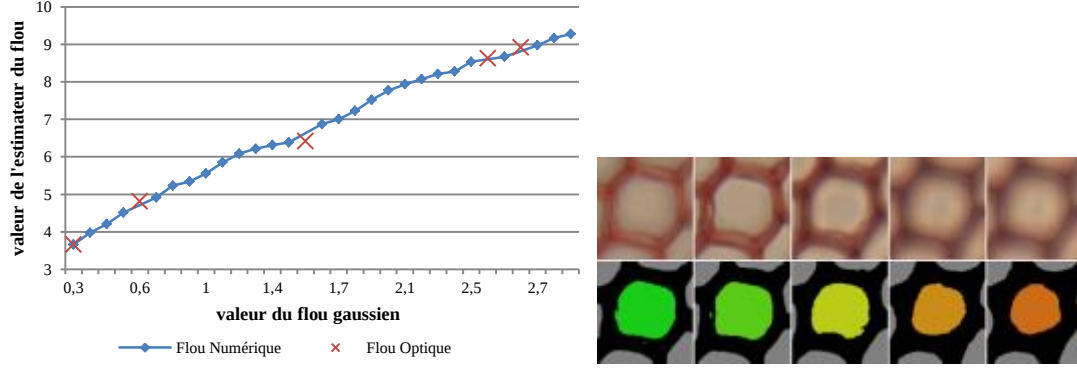
à expliquer : au-delà de la fiabilité de l'identification même de la file, ils reflètent sa qualité. Par exemple, certains alignements, représentés en jaune sur l'image de droite, semblent « visuellement corrects » alors que leur coefficient est plutôt mauvais : ce sont des variations de surface trop fortes qui les pénalisent. Ainsi, la 3^{ème} file orange en partant de la gauche paraît bonne, mais on peut voir de petits bassins introduits dans la file, ils sont issus d'une sur-segmentation due à un décollement de la paroi. Ils peuvent être filtrés pour des analyses statistiques grâce aux coefficients des cellules.

2.4.2.2 Niveau des composants

Par analogie, un coefficient de fiabilité peut être défini pour toute mesure géométrique, à condition de savoir à quels critères elle est sensible. Ainsi, contrairement à la surface des cellules, la surface du lumen dépend du flou local de l'image. Cela provient des algorithmes utilisés pour segmenter l'image ou les portions de l'image. Les cellules sont obtenues par watershed : cette méthode n'est pas sensible au flou car même si l'étalement de l'histogramme est réduit, la ligne de fusion des bassins est préservée tant que les minima locaux ne disparaissent pas. Les minima locaux sont juste décalés sous l'effet du flou. Le lumen est, quant à lui obtenu par l'algorithme des k-moyennes : les points de la cellule sont répartis en deux classes, la classe des faibles intensités et la classe des fortes intensités. L'évaluation du seuil est laissée à l'algorithme. Cette classification est sensible à la dynamique des intensités car la réduction de la dynamique supprime les faibles différences, changeant ainsi la répartition des points dont l'intensité est proche du seuil de segmentation. Le coefficient de fiabilité pour la surface des lumens est donc directement lié au niveau de flou. Il est donc nécessaire de caractériser ce flou pour estimer l'erreur commise sur la mesure des surfaces (ou de l'épaisseur) des éléments constituant la cellule. Nous utilisons un indicateur de flou local introduit dans (Ladjal 2006) qui reflète les deux propriétés suivantes. L'image est d'autant moins floue que la variation d'intensité est rapide et, elle est d'autant plus floue que la variation totale d'intensité est grande. La rapidité de variation des intensités est caractérisée par le maximum local de la norme du gradient, et la variation totale dépend de l'amplitude locale de l'image. Le flou $\mathfrak{F}(p)$ est défini par l'expression suivante où $p(x)$ représente l'intensité de pixel d'un bassin et $p'(x)$ la variance.

$$\mathfrak{F}(p) = \frac{\max_{x \in \mathbb{R}} p(x) - \min_{x \in \mathbb{R}} p(x)}{\max_{x \in \mathbb{R}} |p'(x)|} = \frac{\text{amplitude}(I)}{\max(|\nabla I|)}$$

La figure 33, illustre le comportement cohérent de l'estimateur de flou, calculé à partir de cellules prises sur des images numériquement perturbées par un flou gaussien incrémental.



Cliché de l'auteur

Figure 33. A gauche, relation entre un flou gaussien appliqué à une cellule et la valeur de l'estimateur de flou local. L'estimateur de flou (en bleu) décrit bien l'augmentation du flou. Les croix en rouge représentent la valeur de l'estimateur pour les cinq cellules de droite optiquement floutées. Elles sont positionnées sur le graphe en fonction de leur variance. À droite, les cinq cellules avec en couleur la correspondance de la valeur de l'estimateur. Les lumina les plus flous sont colorés dans les tons rouges, les plus nets dans les tons verts.

Lorsque le coefficient est élevé, le flou est fort. Le coefficient de fiabilité de la surface du lumen (I_{LS}) est défini ci-dessous. Le seuil de T_s est défini par l'étude de la variation des lumina en utilisant un flou gaussien simulé. Les surfaces supérieures au seuil ne sont donc pas significatives.

$$I_{ls} = \begin{cases} 0 & \text{si } \mathfrak{F}(p) > T_s \\ 1 - \frac{\mathfrak{F}(p)}{T_s} & \text{sinon} \end{cases}$$

Cet indicateur de flou permet de filtrer les résultats pour ne garder que les cellules les plus nettes. Il est notamment utilisé dans la fonction d'appariement des cellules pour la création de mosaïques d'images (Article IV, SCIA'13). A terme, il pourrait également fournir un facteur correctif aux mesures évaluées sur les zones plus ou moins assujetties à un flou optique.

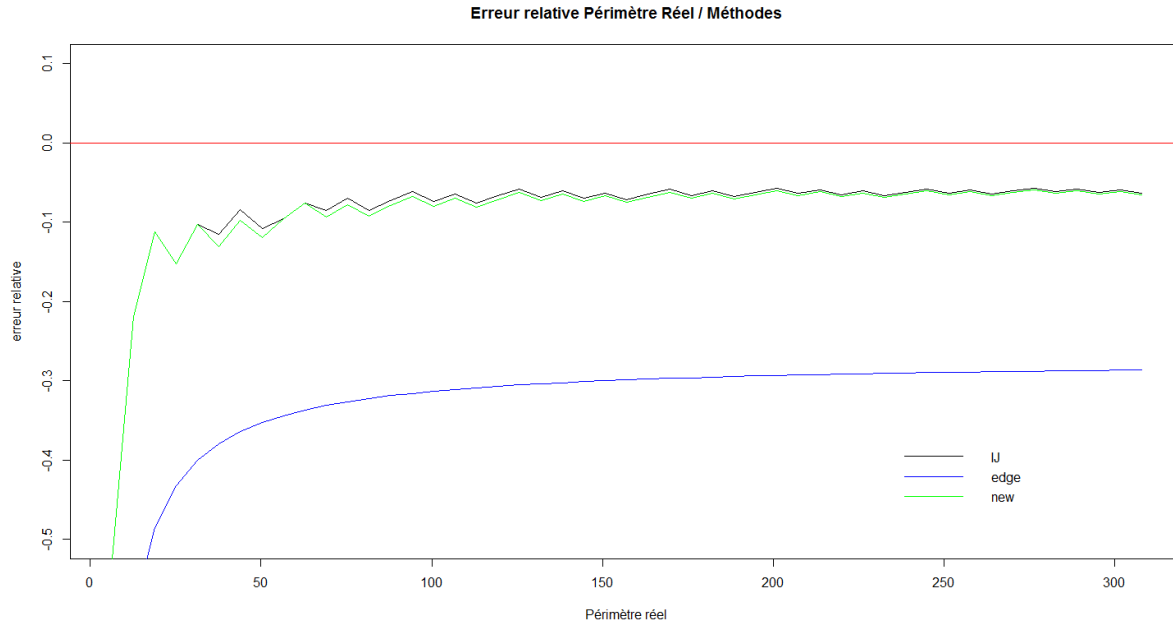
2.4.2.3 Niveau de l'élément

A chaque mesure, une étude des facteurs d'influence est effectuée. Par exemple la précision des surfaces, du périmètre, de la circularité sont sensibles à la taille des objets mesurés. Nous cherchons à mieux appréhender les variations de précision d'une mesure en fonction d'une taille. La mesure retenue est la surface, a priori la moins sensible aux variations locales périphériques des objets. L'objet d'étude est le cercle discret d'Andrès (Andres 1994) car il se rapproche le plus de la forme d'une cellule théorique ; défini par une double inéquation diophantienne, il présente les mêmes propriétés de symétrie que le cercle euclidien. Le rayon du cercle est progressivement augmenté et les valeurs réelles sont comparées avec différentes méthodes de calcul.

2.4.2.3.1 Le périmètre

La figure ci-dessous représente l'erreur relative de chaque méthode par rapport au périmètre réel.

$$Courbe_i = \frac{(Perim_{réel} - Perim_i)}{Perim_{réel}}$$



Les méthodes 2 et 3 convergent vers un même palier à -0.1. Ce qui représente une erreur de 10% par rapport à la valeur théorique. Le palier est atteint pour un rayon 20. La méthode des arêtes tend vers un palier, avec une erreur de 30%. La méthode d'ImageJ donne les meilleurs résultats et est optimisée. Seul problème, elle est sensible à la rotation. Elle peut varier d'une valeur de $2 - \sqrt{2}$. Pour avoir une erreur maximale de 1%, Il faut que le nombre de pixels soit de 60. Ce qui représente un rayon de 10 pixels.

$$2 - \sqrt{2} < 0.01 * Perim$$

$$Perim > 100 * (2 - \sqrt{2}) \approx 60$$

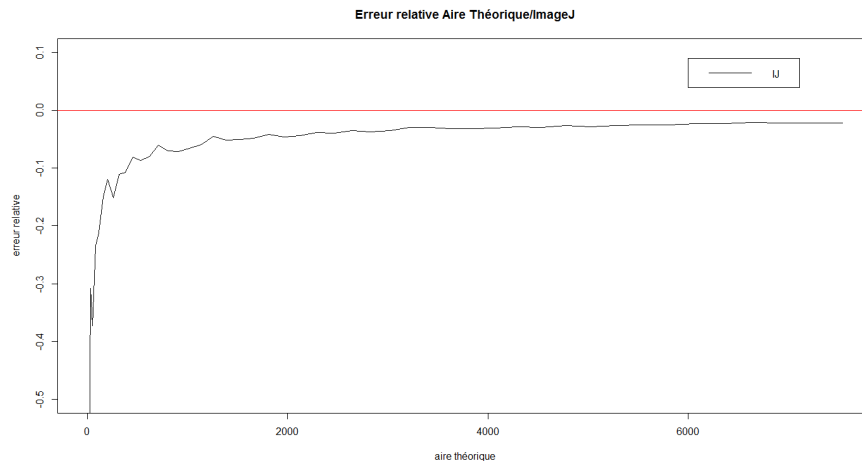
$$Perim = 2 * \pi * r$$

$$r = \frac{Perim}{2 * \pi} \approx 10$$

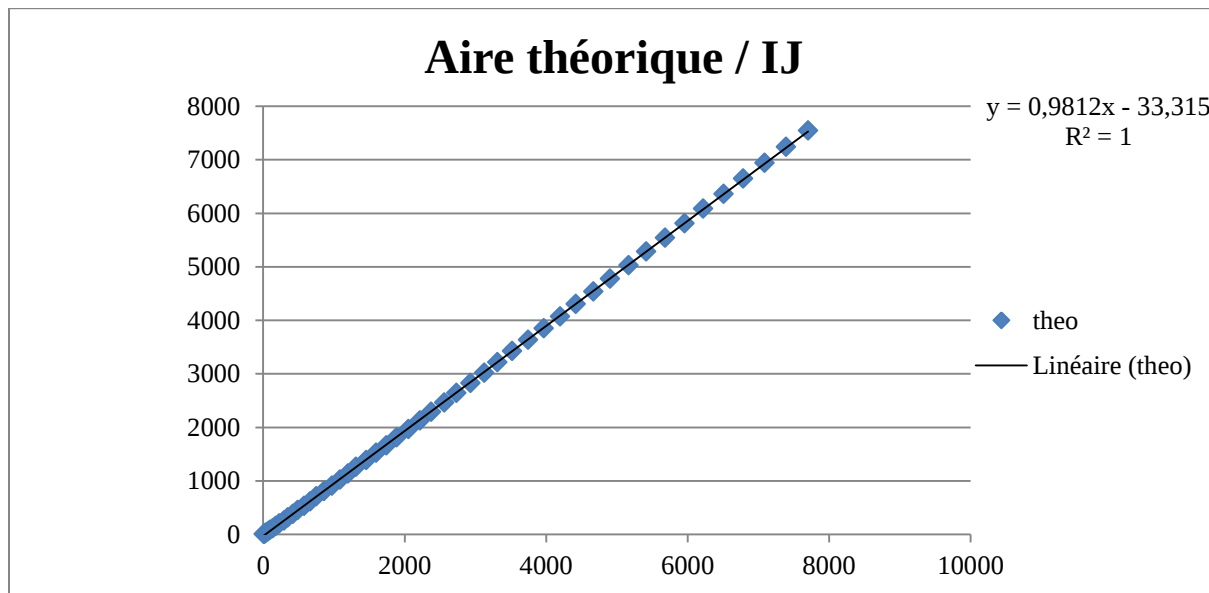
NB : Les cellules ont en moyenne un rayon de 21,4 pixels. Les bassins avec des rayons faibles correspondent souvent à des cellules d'insertions ou des erreurs de segmentation. Un filtrage sur la taille des bassins peut donc permettre à l'utilisateur de supprimer ou d'identifier ces cas.

2.4.2.3.2 La surface

L'étude a aussi été menée pour l'aire avec les cercles discrets et la méthode de comptage des pixels disponible sous ImageJ. La figure ci-dessous montre l'évolution de la mesure par rapport à la valeur théorique.



La taille des cellules varie de 500 à 4000 pixels. Par rapport à un cercle discret, on commet une erreur d'environ 10%.



Le R^2 vaut 1, ce qui signifie les valeurs estimées sont fortement corrélées aux valeurs attendues. On sous-évalue l'aire d'environ 9.8%. La méthode produit une erreur constante de 10%, pour des cellules ayant une aire supérieure à 500 pixels, due un effet de discrétisation du cercle (marche d'escalier). Pour des cellules avec une aire inférieure à 500 pixels, l'erreur devient importante.

2.4.2.3.3 *La circularité*

Nous avons vu que l'aire et le périmètre sont sensibles à la taille des objets calculés. Pour que la circularité soit fiable, il faut que le périmètre de l'objet soit supérieur à 10 pixels. Les tests menés sur la circularité confirment ce seuil. Par exemple, un cercle discret avec un rayon de 1 sera représenté comme un carré ou une croix et sa circularité sera maximale du fait qu'on ne puisse pas représenter mieux un cercle si petit. Les valeurs en dessous du seuil ne seront plus significatives.

2.4.3 **Conclusion sur la mise en œuvre de la fiabilité des mesures**

L'étude statistique des mesures des éléments biologiques est indispensable pour l'étude des tendances et des invariants, comme par exemple la mise en place des canaux résinifères ou des vaisseaux. Pour réaliser ces études statistiques, il faut produire des résultats en très grandes quantités. Produire en grande quantité nécessite de mettre en place des outils de contrôle qualité afin d'assurer les résultats produits pour les études statistiques. Pour garantir la pertinence des mesures, nous avons mis en place des indices de fiabilité basés sur les méthodes mises en œuvre et sur la réalité biologique (longueur et continuité des files). L'indice est un produit de différents termes normalisés. Ces termes sont la transcription d'un effet (facteur d'influence) sur la mesure : le flou va être pondérateur de la mesure des composants cellulaires, tandis que la taille des éléments va être un facteur de pondération de la mesure des cellules. Ces fonctions de fiabilité combinent indépendamment différentes dimensions (ou échelles d'observation) : de la file aux composants de la cellule. L'approche est générique, les termes peuvent être complétés/modifiés en fonction des méthodes de mesure utilisées ou des propriétés des organisations à détecter. Ces indices permettent ainsi de retenir les files les plus significatives ou de mettre en avant des perturbations ou configurations biologiques particulières.

En utilisant trois dimensions (cellules, files, composants), l'étude a montré que les résultats peuvent être qualifiés avant d'être filtrés par l'utilisateur. La méthode donne la distribution des indices de fiabilité, mais ne fixe pas de seuil de caractérisation. Ce seuil est laissé à l'attention de l'utilisateur. La mise en place de ces indices a été confrontée à l'expérimentation et la validation sur des formes de référence (mesures réelles et théoriques). Les conclusions de ces expérimentations ont permis l'extrapolation des fonctions de fiabilité pour l'étude des files cellulaires.

Ces indices caractérisant les facteurs d'influence peuvent être utilisés pour proposer un correctif sur les mesures. Une courbe d'étalonnage du flou pourrait être mise en place pour proposer un coefficient de correction à appliquer à la valeur évaluée. Les indices de fiabilité permettraient ainsi de définir la pertinence d'un résultat et aussi de proposer un correctif quand l'utilisateur le souhaite.

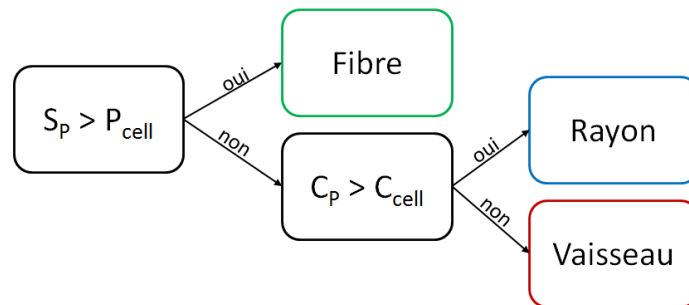
2.5 Typage cellulaire

Le typage cellulaire est l'ultime étape : il permet de classifier les différentes cellules (fibres, trachéides, vaisseaux, rayons,...). Il est réalisé à partir de la caractérisation géométrique et densitométrique des bassins versants associés aux sommets du graphe d'adjacence des cellules. Contrairement aux travaux de (Kennel et al. 2010) et (Jones and Bischof 1996), la classification supervisée a été écartée du fait de la difficulté avérée de constituer des jeux d'apprentissage suffisamment complets et représentatifs de la variabilité biologique existante.

2.5.1 Description de la méthode

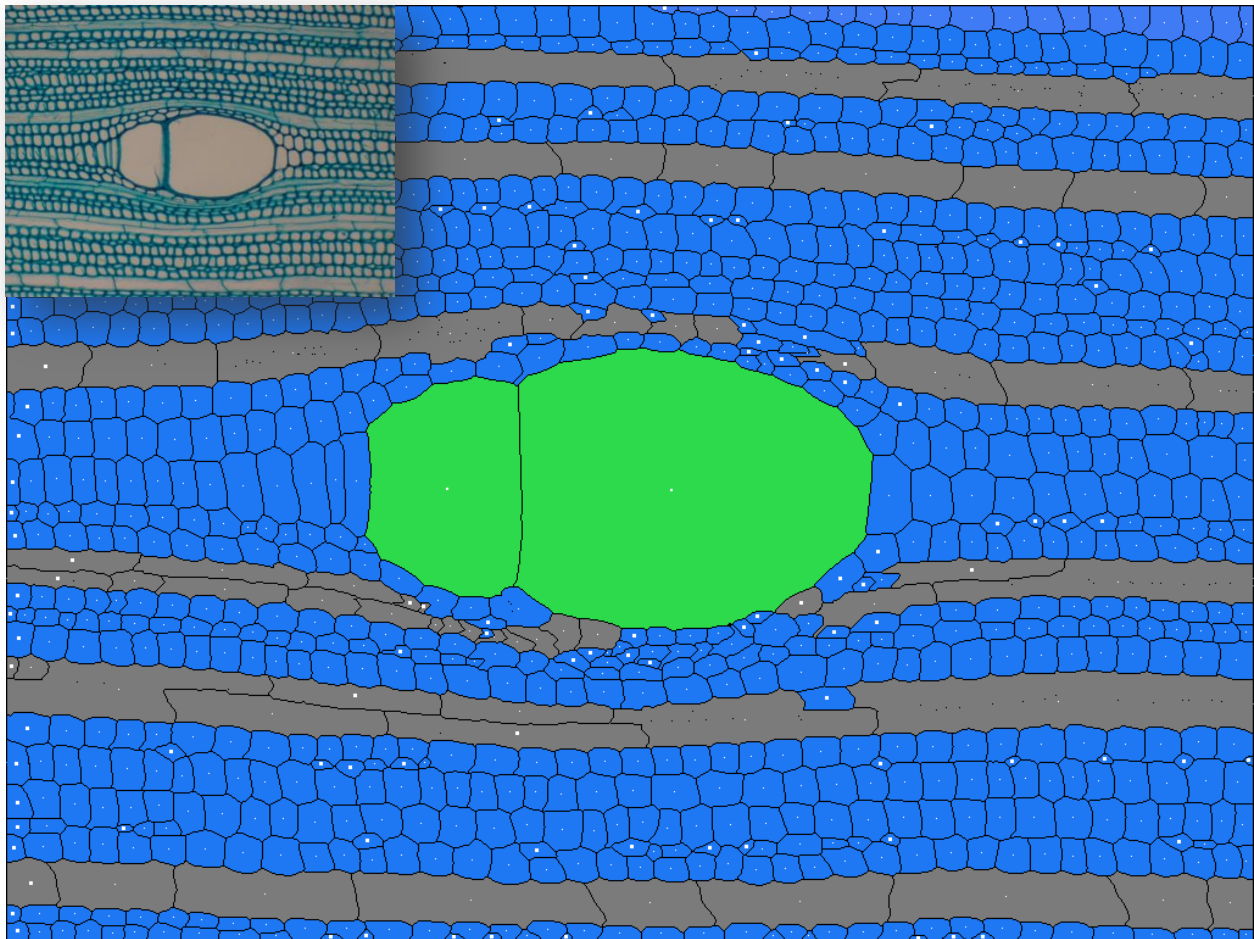
Un arbre de décision permettant de discriminer chaque structure anatomique a été établi avec les anatomistes du bois à partir des deux hypothèses suivantes :

- Le périmètre des cellules, noté P , permettant de différencier les « grosses » structures des cellules globalement « petites ».
- La circularité des cellules, noté C , permettant de différencier les rayons globalement allongés des vaisseaux globalement circulaire.



Les seuils (S) utilisés dans l'arbre de décision (ci-dessus) sont automatiquement réévalués pour chaque image par une classification des valeurs numériques produites réalisées par une classification par 2-moyenne.

Un résultat du typage est présenté ci-dessous sur une image de Gymnosperme.



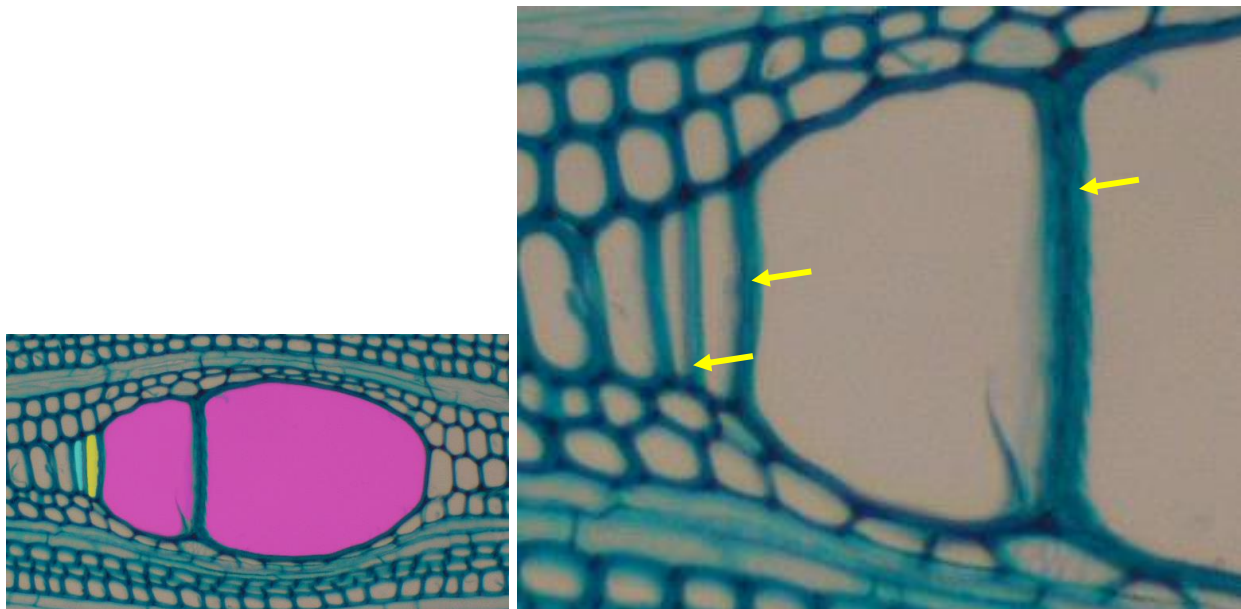
Cliché de l'auteur

Figure 34. Typage automatique des cellules sur une coupe transversale d'acajou (*Pycnanthus sp.*). Les vaisseaux sont colorés en vert, les fibres en bleu et les rayons en gris. Hormis le typage des cellules hybrides (cf. §2.5.2), le typage avec des règles simple est bon.

Nous appelons "cellule hybride", des cellules en voie de spécialisation ou de transition, ne répondant plus tout à fait aux critères classiques de fibres et pas tout à fait aux critères des vaisseaux.

2.5.2 Etudes prospectives

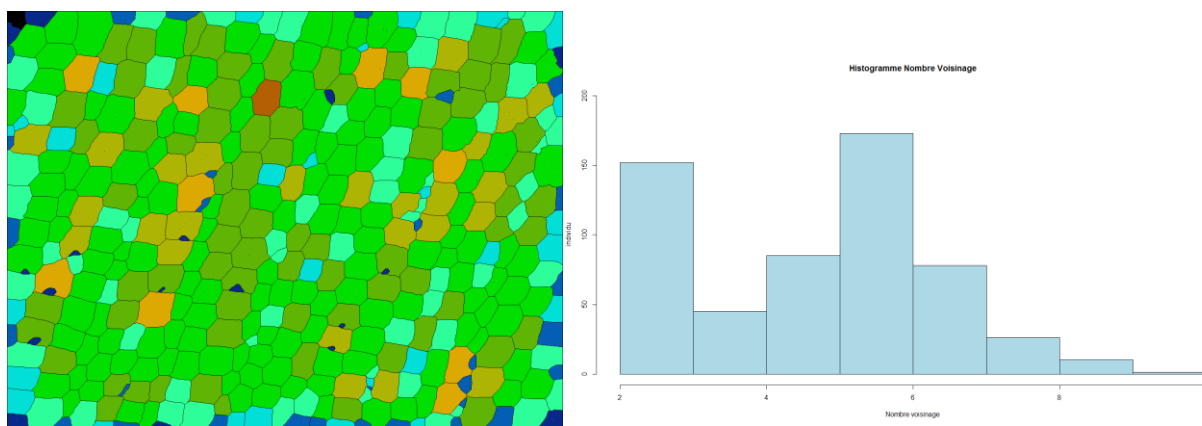
Actuellement il n'y a pas de consensus des botanistes sur les critères réellement discriminants pour typer les cellules hybrides. Les cellules en violet sont identifiées par tous comme étant des vaisseaux. Après discussion la cellule jaune serait aussi un élément de vaisseau car la texture de la paroi est plus diffuse (cf. figure 35, flèches en jaune). La cellule bleue devrait être pour les mêmes raisons assimilée à un vaisseau bien que les experts la classent comme une fibre à cause de sa taille.



Cliché de l'auteur

Figure 35. Reconnaissance des vaisseaux chez une Angiosperme (acajou - *Pycnanthus sp.*). Les cellules colorées en rose et en jaune sont des éléments de vaisseaux du fait de leur taille, alors que la cellule colorée en bleue serait une fibre. La transition fibre – vaisseaux est difficile à apprécier : les cellules prennent des critères géométriques qui ne sont ni totalement ceux des fibres, ni totalement ceux des vaisseaux. La caractérisation porte sur d'autres critères, ici la granularité de la paroi.

Une étude prospective a été menée pour caractériser les types cellulaires à partir de leur valence (cf. figure 36). La valence d'une cellule est définie par le nombre de cellules voisines qu'elle possède. Théoriquement les fibres possèdent entre 4 et 6 voisins selon la position sur le cerne. Tandis qu'un élément de vaisseau ou un rayon en possèdera beaucoup plus.



Cliché de l'auteur

Figure 36. A gauche, colorisation en fonction de la valence sur une coupe de sapin (*Abies alba*). Les cellules sont colorées graduellement en fonction du nombre de voisin. En bleu les cellules avec une faible valence, en vert les cellules avec une valence proche de la valeur théorique et en marron les cellules avec une forte valence.

Les tendances théoriques sont bien suivies. Des éléments perturbateurs (artéfacts de segmentation, intrusions, ponctuations, etc.) peuvent faire augmenter le nombre de voisins d'une cellule, mais les rayons et vaisseaux (facilement reconnaissables) sont bien discriminés. La valence peut être utilisée pour identifier et supprimer les ponctuations, dont leur valence est toujours de 2. Pour les typages complexes, la valence n'apporte pas d'information supplémentaire permettant d'effectuer un choix. Seule l'étude de la texture des parois semble pouvoir à terme de lever l'ambiguïté.

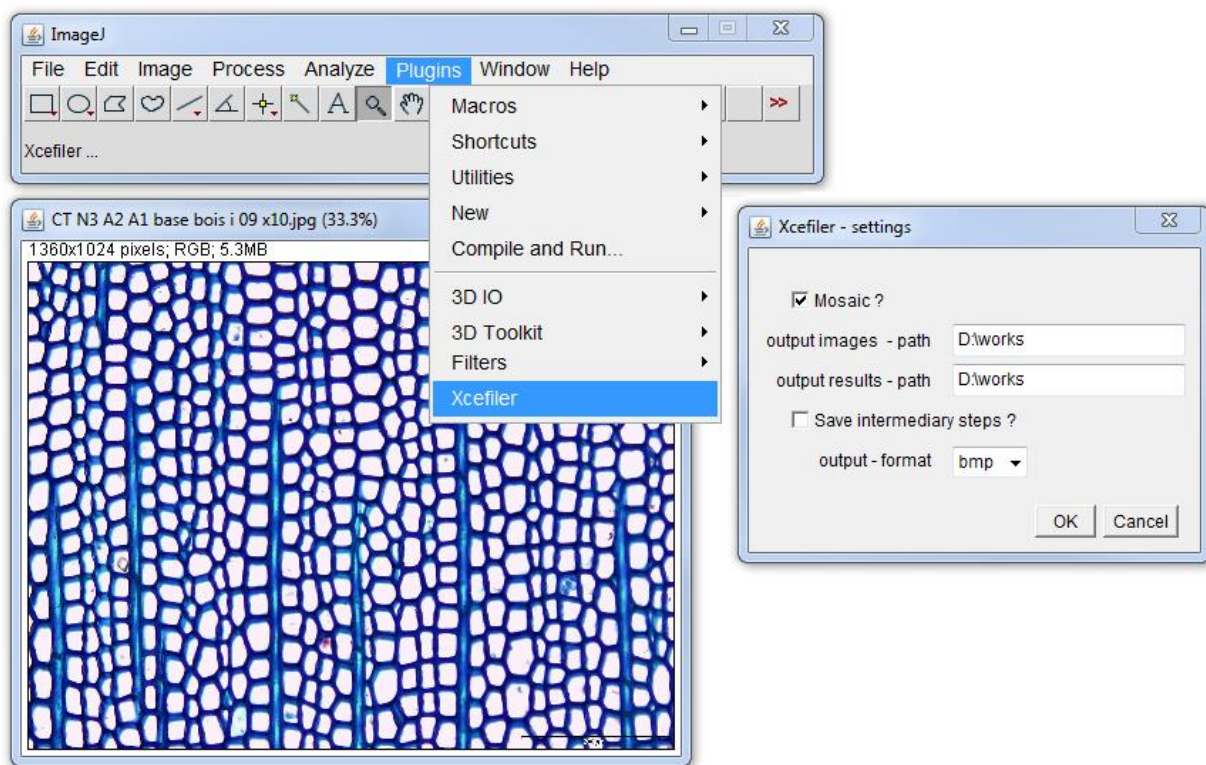
2.5.3 Conclusion sur la mise en œuvre du typage cellulaire

Le typage cellulaire consiste à classifier les cellules en trois catégories : les fibres, les rayons et les vaisseaux (cf. §1.2.2). Il est souvent réalisé à partir d'au moins deux plans de coupes pour trouver les indices permettant la classification. Nous proposons une méthode simple, basée sur un arbre de décision expert, pour typer les cellules à partir d'un seul plan de coupe. Les discussions avec les experts et les études prospectives montrent que le typage est un problème très complexe sur un seul plan de coupe. Des pistes de travail ont malgré tout été identifiées, comme par exemple l'ajout d'indices texturaux de paroi pour différencier les éléments.

3 APPLICATION ET RESULTATS

3.1 Implémentation

La méthode a été implémentée en Java et intégrée sous forme de plugin à la plateforme ImageJ. ImageJ (Schneider et al. 2012) est un logiciel multi-plateforme et open source de traitement et d'analyse d'images développé par le National Institute of Health ; il peut être utilisé sur n'importe quel ordinateur capable d'exécuter les plates-formes Java (Mac OS X, Linux x86 ou Microsoft Windows), ce qui en fait un produit idéal à utiliser dans la plupart des laboratoires. Des bibliothèques libres, tel que *Java Universal Network/Graph Framework* (<http://jung.sourceforge.net/doc/index.html>) pour la gestion des graphes sont utilisées pour avoir une structure de données efficace. A titre indicatif, le plugin permet de traiter des images de 1600x1200 pixels en moins de 20 secondes sur un ordinateur doté d'un processeur Intel Q720 cadencé à 1.6 GHz.



Cliché de l'auteur

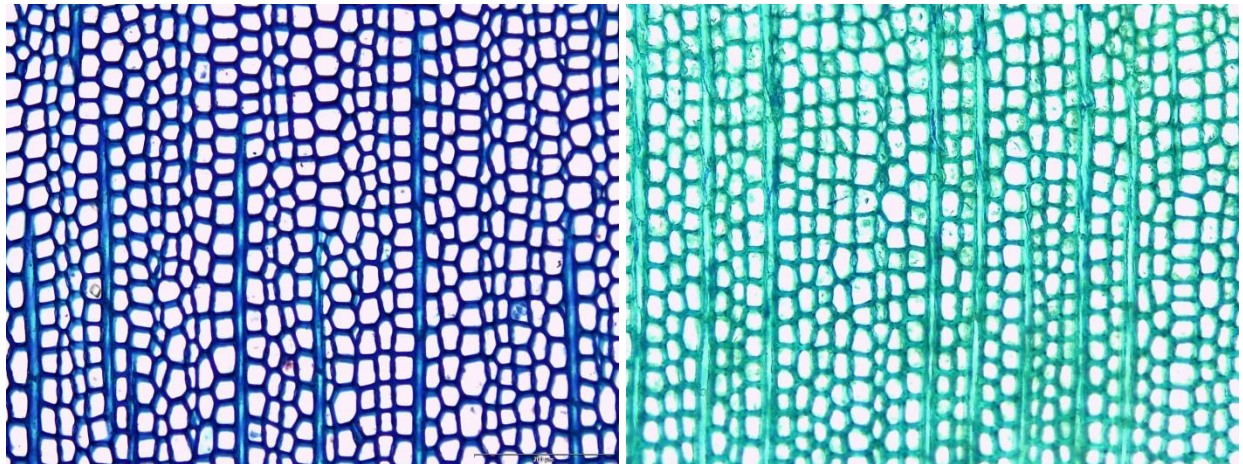
Figure 37. Interface d'ImageJ avec la fenêtre du plugin ouverte. Le plugin permet de choisir si l'on veut traiter une mosaïque ou juste une image et de sauvegarder les étapes intermédiaires du traitement.

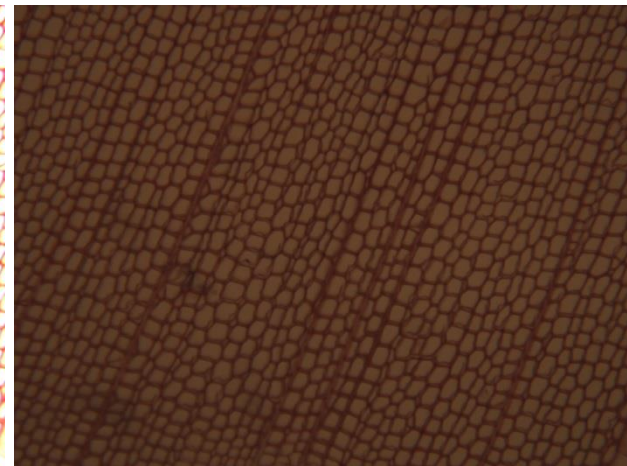
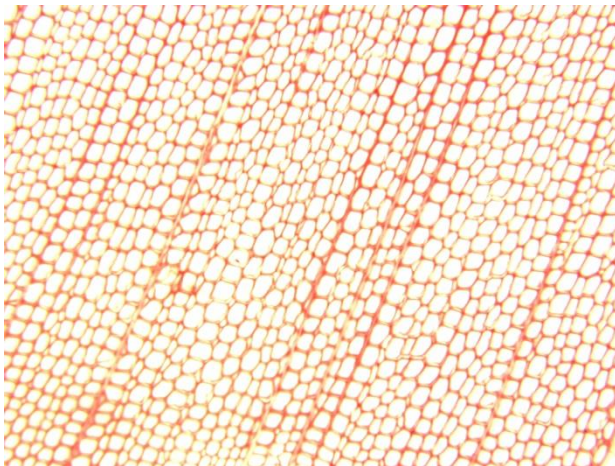
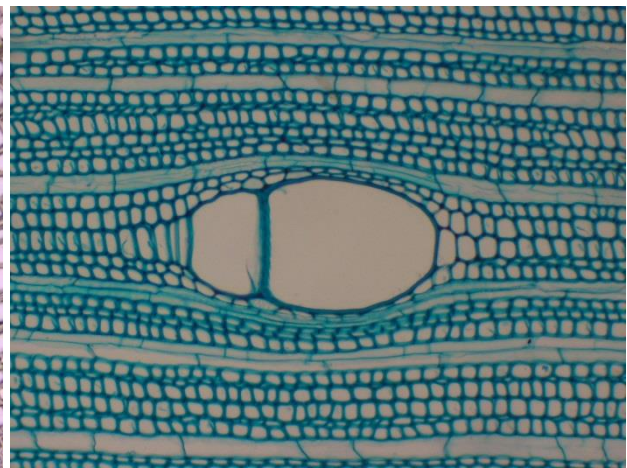
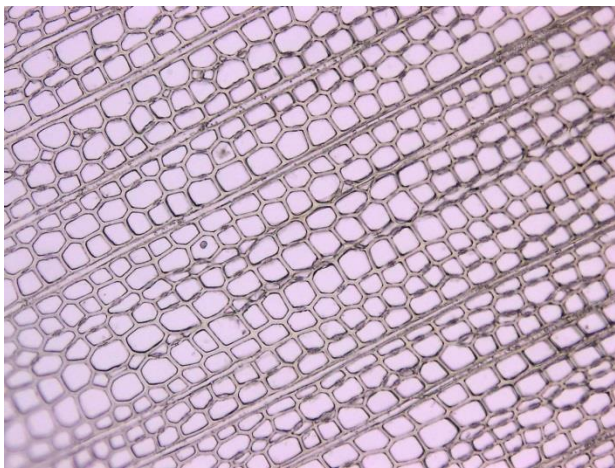
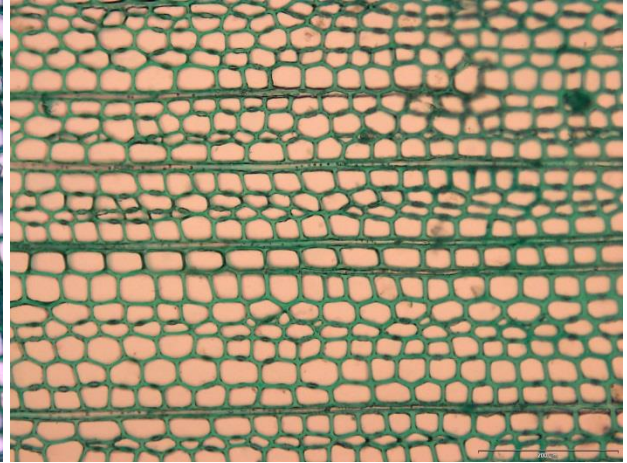
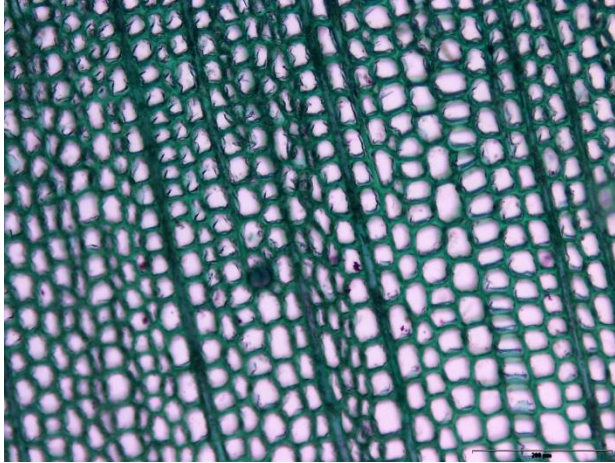
3.2 Application aux files cellulaires

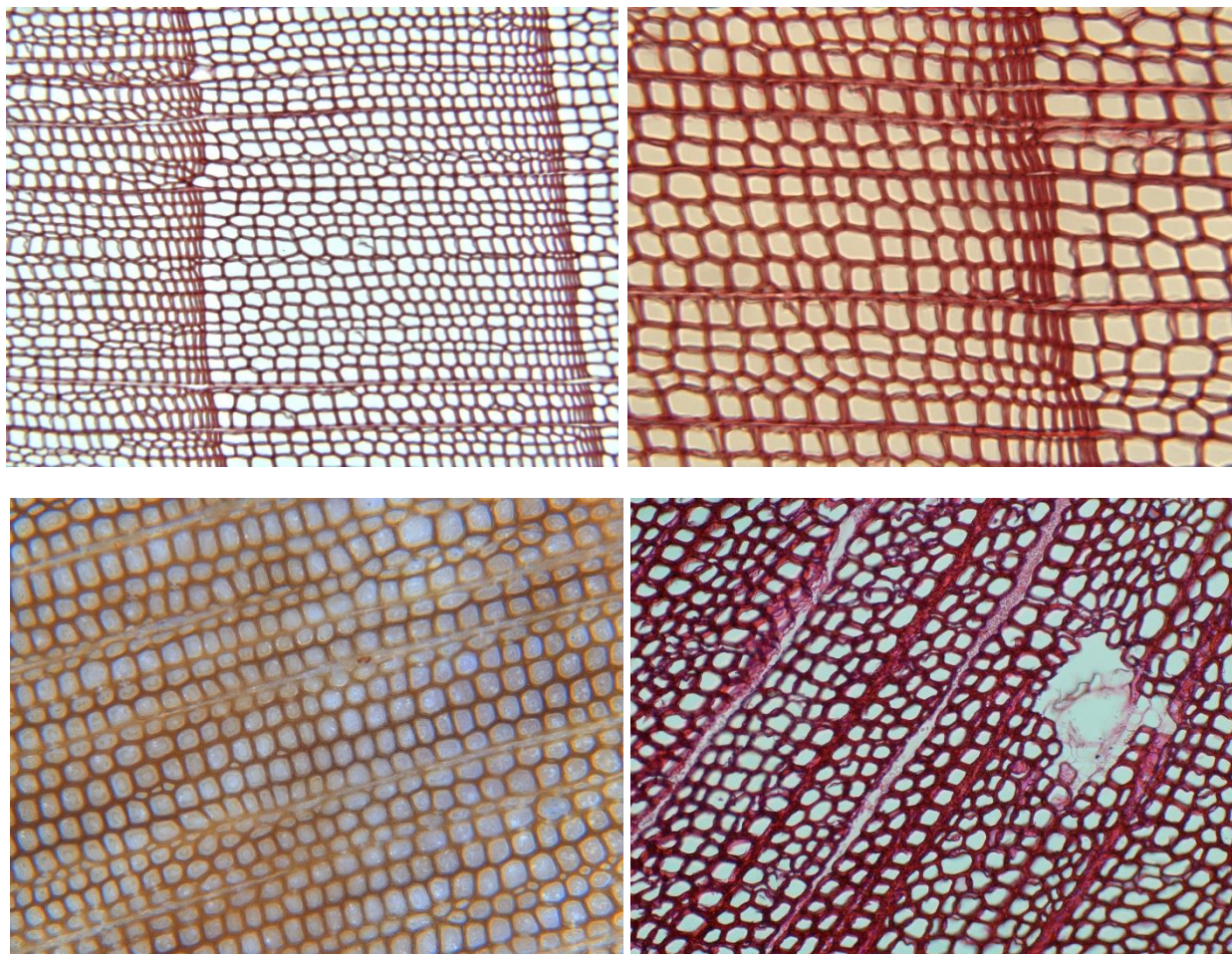
Les résultats obtenus au cours de la thèse sont présentés et discutés dans ce chapitre. Dans un premier temps, nous présentons le jeu d'essai sur lequel la méthode a été testée pendant son développement. Ensuite, nous montrons un exemple de détection automatique de file cellulaire. Les limites de la méthode sont discutées et illustrées en troisième partie. Et pour finir, des études (réduites) de performance de l'ensemble de la méthode et des perspectives à court terme sont présentées.

3.2.1 Jeu d'essai

L'étude expérimentale consiste à comparer les résultats produits par notre méthode à ceux issus d'un jeu de données expert constitué de 12 images (cf. figure 38) dont les cellules ont été manuellement détournées par des anatomistes du bois. Les différentes expérimentations permettent d'évaluer les limites de notre méthode notamment à travers des contextes biologiques, des protocoles de préparations ou des conditions d'acquisition divers et variés. Les expérimentations sont essentiellement motivées par des questions simples : notre méthode est-elle adaptée aux Angiospermes et aux Gymnospermes, sensible aux protocoles de préparation, à la présence de flou optique, au niveau de luminosité, etc. ?





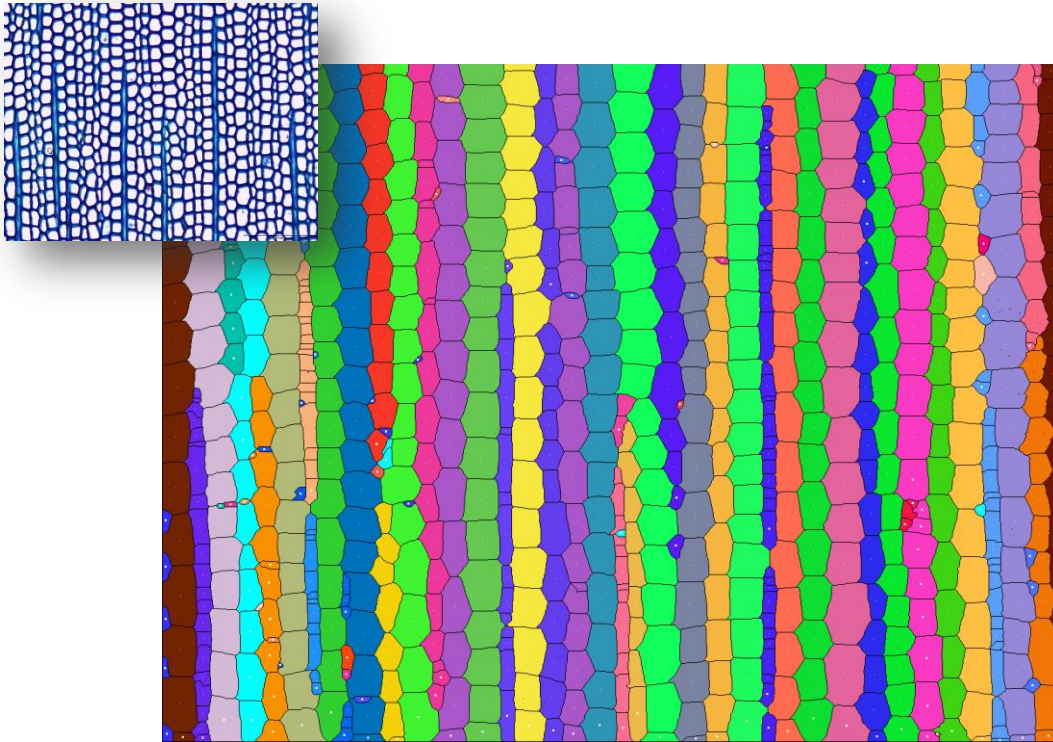


Clichés de l'UMR AMAP

Figure 38. Panel d'images du jeu d'essai. Les images proviennent de 5 espèces différentes d'arbres (cf. §1.2.1) issues de différents protocoles de préparations. Pour plus d'information sur les essences et protocoles d'acquisition par image, se référer à l'annexe 1.

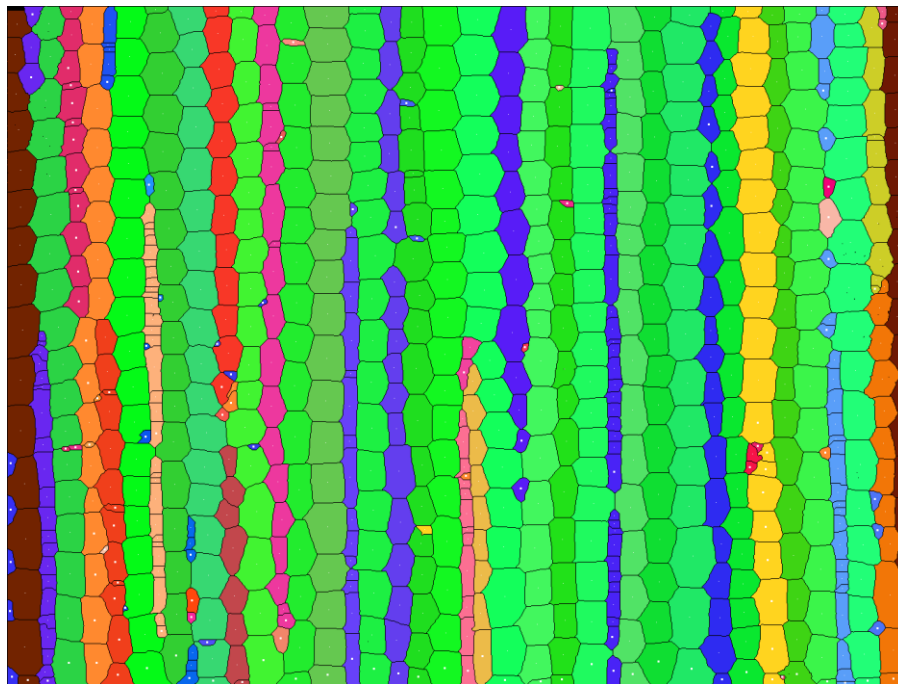
3.2.2 Mise en œuvre de la détection automatique des files cellulaires

Nous avons choisi d'illustrer et détailler le résultat de toute la chaîne de traitement sur une image. Les résultats des images du jeu d'essai (ci-dessous) sont présentés en annexe. Les résultats sont illustrés avec l'image de Sapin (*Abies alba*) colorée au bleu de toluidine et numérisée avec un grossissement de x200. La figure 39 présente le résultat de la détection automatique des files (cf. §2.2). Les files sont colorées aléatoirement. La figure 40 montre la colorisation des files en fonction de leur score (cf. §2.4). Comme la fusion ne donne pas de représentation visuelle, pour illustrer son principe et le résultat, nous avons découpé l'image en 4 sous-parties de taille et de chevauchement inégaux (cf. figure 41 et 42).



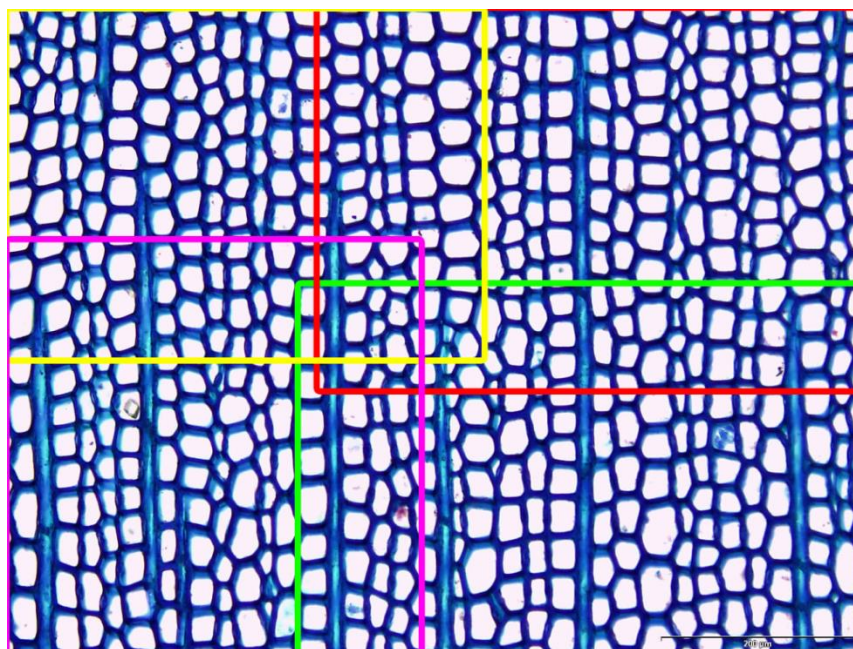
Cliché de l'auteur

Figure 39. Résultat de la modélisation des files cellulaires. Chaque file est aléatoirement colorée. 88% des files sont reconnues par la méthode avec un temps d'exécution de 10,4 sec.



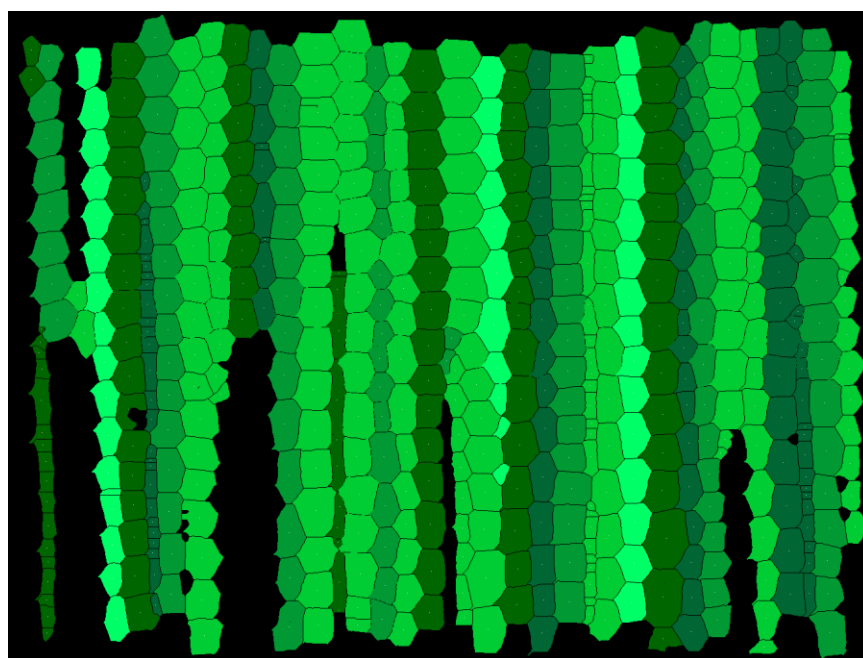
Cliché de l'auteur

Figure 40. Résultat du processus de fiabilité des files (cf. §2.4.2.1). La colorisation des files cellulaires est fonction de leur fiabilité. Les files qui sont moins fiables (discontinues ou hétérogènes) apparaissent dans des tons chauds (jaune à marron).



Cliché de l'auteur

Figure 41. Découpage de l'image en 4 sous-parties de taille et de chevauchement inégaux. La méthode de fusion doit être indépendante de la dimension de l'image. Seul le chevauchement compte pour la fusion. Les cellules du bord de l'image ne sont pas prises en compte pour la fusion, car elles sont tronquées et donc erronées.



Cliché de l'auteur

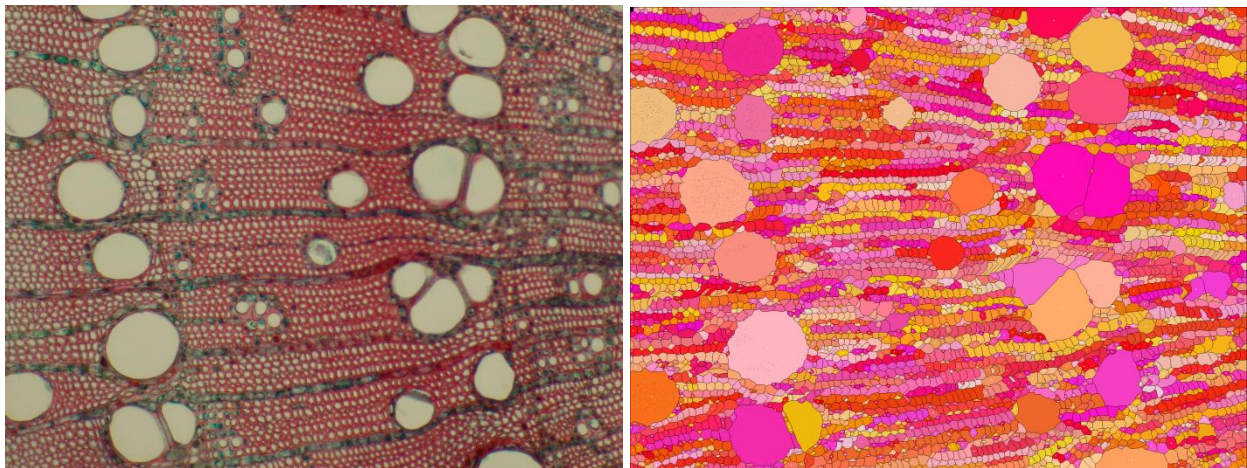
Figure 42. Les quatre images sont traitées l'une à la suite de l'autre. Les zones noires de l'image sont les portions de file qui n'ont pas de sommet en commun avec au moins une des images. Les cellules du bord de l'image ne sont pas prises en compte pour la fusion car elles sont biaisées. Nous préconisons un chevauchement minimal de 3 cellules pour s'assurer qu'au moins une cellule éloignée du bord soit en commun.

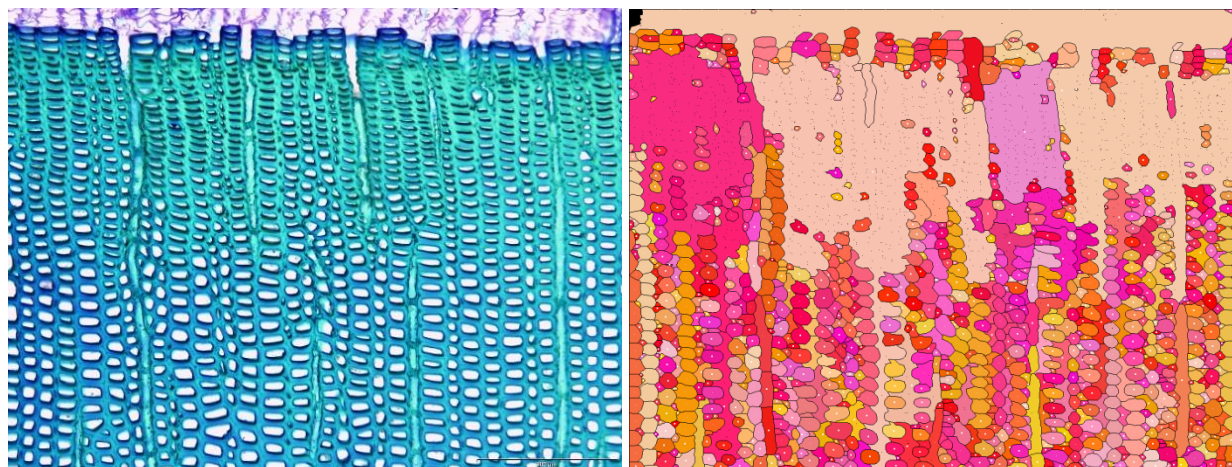
La méthode de reconnaissance automatique d'organisation permet la détection des files cellulaires sur des mosaïques d'images numériques. La représentation des files permet d'une part de contrôler visuellement leur parcours, et, d'autre part de choisir les plus pertinentes pour une étude statistique. En moyenne 80% des files sont automatiquement reconnues. Les 20% de files cellulaires non reconnues sont compensées par le nombre d'images que peut traiter la méthode. La fusion des résultats permet de suivre les files sur l'ensemble du plan ligneux. L'intégration des résultats dans un outil de visualisation et de correction des files, est un plus pour l'utilisateur.

3.2.3 Limites

La limite majeure à l'identification automatique des files est inhérente au contenu des images, c'est-à-dire soit aux caractéristiques photométriques de l'image, soit aux configurations biologiques. Les images présentant une dynamique d'intensités localement inversée - par exemple la jonction bois d'hiver / bois d'été (figure 43, image du bas), restent difficile à traiter. Notre méthode reposant sur le contraste de l'image, les bassins versants correspondant aux cellules sont alors mal détectés et de ce fait, la détection des files est erronée et incomplète. De la même manière, les images de rondelles de bois poncé (non colorées) sont traitées en réflexion : la lumière ne traverse pas l'échantillon. De ce fait, les lumens apparaissent d'une couleur quasiment semblable à celle des parois, en tout cas, pas suffisamment contrastée pour garantir une bonne reconnaissance des bassins versants et par là-même une bonne identification des files. A contrario, les coupes de bois non colorées traitées en transmission c'est-à-dire traversées par la lumière, sont moins contrastées que les coupes colorées. Toutefois la différence paroi/lumen est suffisamment prononcée pour permettre à l'algorithme d'identifier correctement les files.

Pour les organisations cellulaires complexes (cf. figure 43, image du haut), c'est-à-dire pour lesquelles les alignements cellulaires ne sont pas intuitifs, de nouvelles règles des parcours et de reconstruction devront être établies avec les biologistes.





Cliché de l'auteur

Figure 43. En haut, frêne coloré à la safranine. La détection des cellules est bonne, la méthode ne produit que des tronçons à cause de la complexité de l'organisation biologique. En bas, sapin coloré au bleu de méthylène ; l'identification est biaisée par la présence d'inversion locale de contraste au niveau de la jonction bois été/hiver.

3.2.4 Performances

A ce stade de notre travail, la performance globale de la méthode de traitement est essentiellement une appréciation qualitative, i.e. une évaluation qui compare de manière assez globale les résultats produits aux annotations expertes. Les temps « de traitement », le nombre de cellules détectées et les pourcentages de files identifiées sont présentés au tableau 1. Le tableau montre que la méthode automatique détecte un nombre cohérent de cellules en environ 80 à 100 fois moins de temps que l'expert en fonction des préparations, des configurations anatomiques ou des grossissements utilisés et que globalement notre méthode fonctionne bien (88% de files bien détectées en moyenne) sur des images présentant une organisation structurée avec un contraste paroi/lumen marqué.

Tableau 1. Résumé de quelques résultats significatifs : la taille des images, le nombre de cellules extraites, le temps CPU obtenu sur un ordinateur équipé d'un processeur Intel Xeon à 2,3 GHz et le nombre de files détectées en pourcentage par rapport au nombre de files identifiées par l'expert. Pour l'expertise, les cellules ont été détournées manuellement sur le logiciel ImageJ. Les résultats sont donnés pour un représentant de chaque essence.

Espèce	Taille (pixels)	Nombre cellule (expert)	Temps expert (sec)	Nombre cellules (plugin)	Temps plugin (sec)	% de files détectées
<i>Pycnanthus sp.</i>	1024x768	1302	2520	1359	14,3	86
<i>Abies sp.</i>	1360x1024	787	1300	800	12,4	88
<i>Pinus nigra</i>	1600x1200	1794	2750	1873	23,2	92
<i>Pinus caribensis</i>	1360x1024	819	1500	828	11,5	91

<i>Pinus brutia</i>	1600x1200	1411	2450	1458	16,1	83
<i>Quercus ilex</i>	1600x1200	1623	3460	1724	16,2	62

L'appréciation de la performance est loin d'être triviale. Nous pourrions mettre en place un indice de performance basé sur un rapport entre les effectifs des classes « vrai positif », « faux positif » et « faux négatif » (Pavez et al. 2011). A l'échelle locale de la cellule, il s'agirait d'évaluer le taux de recouvrement entre cellules annotées par l'expert et cellules segmentées par l'application ; la classe « vrai positif » serait définie par l'intersection de la segmentation experte avec la segmentation automatique, et représenterait les pixels de l'image communs aux deux annotations ; les classes « vrai positif », « faux positif » (les pixels annotés par l'expert non présents dans l'image segmentée par l'application) et « faux négatif » (les pixels de l'image segmentée par l'application non annotés par l'expert) seraient définies par la réunion de la segmentation experte avec la segmentation automatique ; le rapport des effectifs des deux ensembles ainsi définis exprimerait une « performance » du recouvrement des annotations expertes et algorithmiques. A l'échelle globale, il s'agirait d'évaluer le taux de similitude dans la succession des cellules des files décrites par l'expert et celles identifiées par l'application. La limite actuelle à la mise en place de ces indicateurs est finalement le manque de données expertes car la segmentation manuelle est longue et fastidieuse, et souvent incomplète, i.e. l'image n'est pas entièrement traitée.

3.3 Autres applications

De nombreuses questions n'ont pas pu être approfondies du fait du nombre restreint de jeux de données cellulaires à disposition. Nous avons choisi de les aborder à travers des expérimentations menées sur d'autres objets d'étude.

3.3.1 La robustesse de la méthode de segmentation

La méthode de segmentation a été développée pour traiter des images couleurs de cellule de bois. Les images couleurs peuvent être vues comme des images multi-spectrales simples décrites par la réponse de trois longueurs d'onde du spectre visible.

Mais qu'en est-il pour le traitement d'images composées de plus de trois longueurs d'onde prises sur un spectre élargi ?

On s'intéresse à des images de champ de crevettes prise par satellite qui devraient être bien adaptées à une segmentation par watershed car les champs peuvent être assimilés à des bassins de type "cellule". L'idée est de segmenter les limites des champs pour dresser un plan cadastral. Les images sont acquises à partir d'images multi-spectrales satellitaires en Inde. Les tests sont menés sur une sous-partie (1720x1416 pixels) de l'image native pour accélérer les traitements

(cf. figure 44). La taille est néanmoins suffisante pour vérifier les points clés dans la méthode de traitement et la méthode de comparaison mise en place.

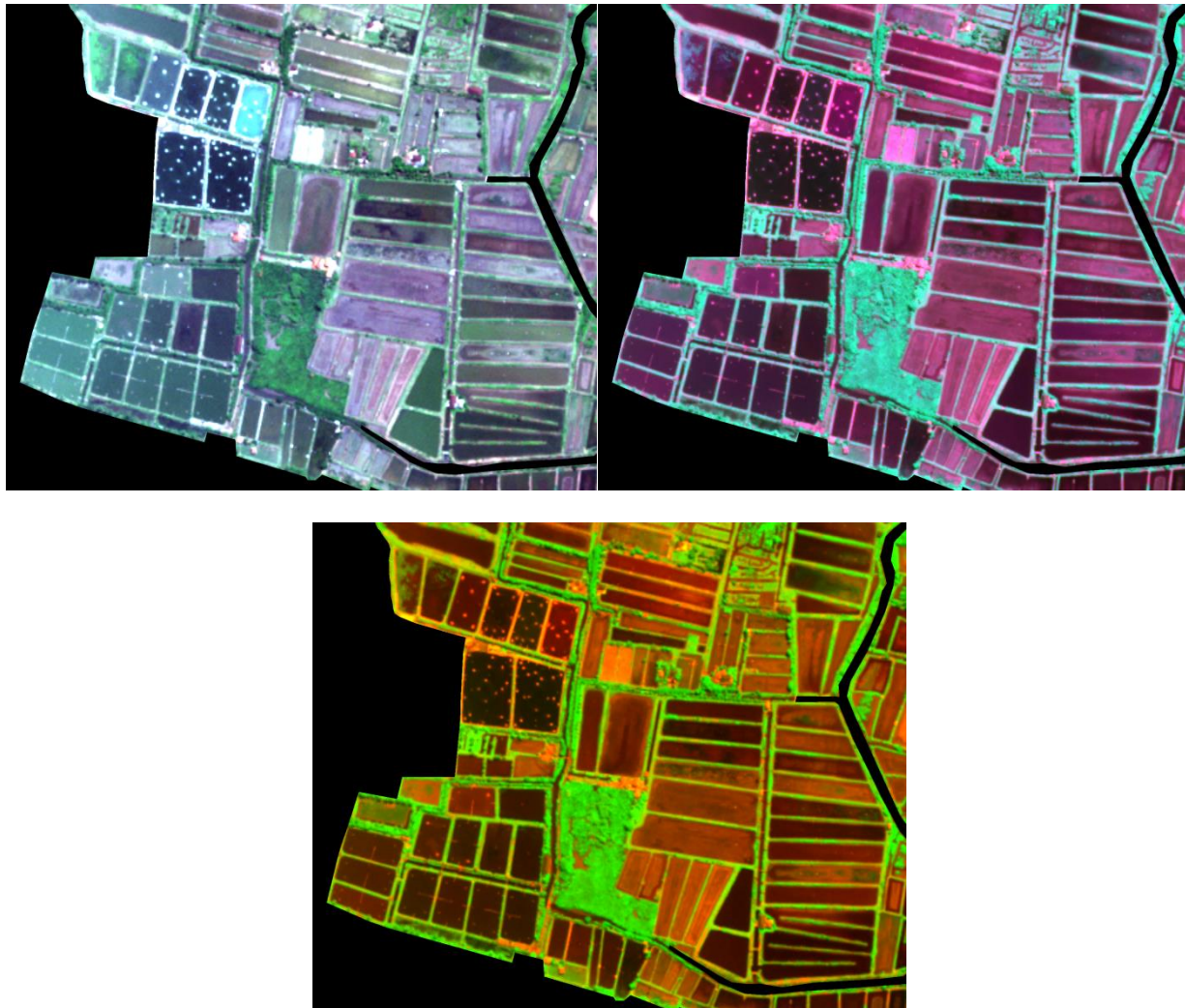


Clichés de l'UMR AMAP

Figure 44. Segmentation automatique d'une image satellitaire de champs de crevettes. A gauche, l'image native, la rivière au centre de l'image a déjà été filtrée par l'opérateur. A droite, le résultat de la segmentation en surimpression. Certains champs sont bien segmentés, mais on observe une sur-segmentation assez importante pour les zones texturées.

Plusieurs auteurs proposent des stratégies différentes pour traiter de telles images : l'une d'entre elles consiste à segmenter séparément chaque canal puis à fusionner le résultat de la segmentation (Frontera-Pons et al. 2013). Une autre consiste au contraire à combiner au mieux par ACP les différents canaux et ne travailler que sur les premières composantes obtenues (Tarabalka et al. 2010).

Du fait du nombre peu élevé de bandes spectrales dont nous disposons, nous avons opté pour des combinaisons idoines de certains canaux (cf. ci-dessous), classiquement utilisées en télédétection.



Clichés de l'UMR AMAP

Figure 45. En haut à gauche, recomposition avec les canaux [Red, Green, Blue]. En haut à gauche, recomposition [Red, Nir, Pan]. En bas, recomposition [Red, Nir]. Le canal Nir correspond au proche infrarouge (*Near infrared*), sa longueur d'onde est comprise entre 0,78 et 3 μm . Le canal Pan correspond au Panchromatique qui ne discrimine pas les couleurs, c'est-à-dire dont la sensibilité à la longueur d'onde de la lumière est similaire à celle de la vision humaine.

Les images sont traitées par le processus de segmentation mis au point pour l'identification des files cellulaires (voir §2.1.2).



Clichés de l'auteur

Figure 46. Lignes de crête respectivement obtenues à partir des images RGB (en haut à gauche), RedNirPan (en haut à droite) & RedNir (en bas à gauche). Quelques différences sont observables mais globalement, le résultat est largement sur-segmenté par rapport à l'image de référence (en bas à droite)

Les images de watershed sont sur-segmentées du fait des textures des structures étudiées (cf. ci-dessus). Les champs de crevettes non entretenues sont progressivement recolonisés par la mangrove et donc envahis par un couvert végétal.

Ce cas illustre :

- La difficulté à réduire la texture par filtrage préalable à la phase de segmentation ou par homogénéisation des zones (quantification de couleur, meanshift...). En clair, il est difficile de filtrer les minima locaux de l'image.
- La nécessité de bien traiter la réduction de la sur-segmentation : a priori, la solution serait plutôt d'approfondir des méthodes de fusion de bassins adjacents homogènes. Le watershed est globalement un bon algorithme à condition de pouvoir bien préparer

l'image, c'est-à-dire bien réduire le bruit, l'effet texture,... De ce fait, il ne peut seul conduire à un résultat satisfaisant dans la majorité des cas.

3.3.2 Evaluation quantitative de la segmentation

Les images de cellules ne nous pas permis d'évaluer la précision de la segmentation du Watershed. Les lignes de crête se positionnent relativement bien par rapport à l'espace intercellulaire et l'algorithme est indépendant du flou local. Mais en l'absence de segmentation de référence, il nous a été impossible de quantifier l'erreur de segmentation. Là encore, les champs de crevettes ont pu être utilisés pour mener ce type d'expérimentation : il existe en effet des jeux d'images annotés qui définissent les cadastres actuels de référence.

La comparaison d'images segmentées ou labellisées est en soi assez complexe. De nombreuses approches sont présentées dans la littérature (McCune et al. 2002) et ont été discutées lors de réunions avec des télédéTECTEURS :

- **Indice de Kappa** (de Cohen) qui exprime la réduction proportionnelle de l'erreur obtenue par une classification, comparée à l'erreur obtenue par une classification complètement au hasard. Un Kappa de 0,75 signifie que 75% des bonnes classifications ne sont pas due au hasard.
- **Indice de Jaccard** qui indique le recouvrement de deux régions : il est défini par le rapport entre l'intersection et la réunion des deux surfaces à comparer.
- **Indice de Dice** qui exprime la similarité entre deux échantillons. Il se définit comme le double de l'intersection de deux lots (échantillons de valeurs), divisé par l'union de ces deux lots.
- **Indices de sensibilité et de spécificité**, initialement développés pour les analyses médicales basées sur les comparaisons des classes «vrais positifs», «faux positifs», «faux négatifs» et «vrais négatifs», pour déterminer la validité de tests diagnostiques quantitatifs. La sensibilité d'un signe pour une maladie est la probabilité que le signe soit présent si le sujet est atteint de la maladie considérée. La spécificité d'un signe pour une maladie est la probabilité que le signe soit absent si le sujet n'est pas atteint de la maladie.
- **Indices de précision et de rappel** développés pour la fouille de données ou les systèmes de recherche documentaire. Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé par le moteur de recherche pour une requête donnée. Ces indices, souvent opposés aux précédents, peuvent également se transposer à la comparaison de deux échantillons par l'étude des effectifs des classes «vrais positifs», «faux positifs», etc.

Pour notre étude nous avons choisi de travailler avec les indices de précision et de rappel. Comparer deux contours implique de se donner une marge de tolérance : on compare deux contours à x pixels près. Cette marge de tolérance est décrite par la surface obtenue par dilations successives d'une des lignes de crête. A gauche, le dilaté des lignes de référence (cf. figure ci-dessous), à droite le dilaté des lignes du watershed (obtenu dans les 2 cas par 3 dilations successives).



Clichés de l'auteur

Figure 47. Dilatation des lignes de crête de l'image de référence (à gauche) et des lignes du watershed (à droite).

On note L les lignes de crête du watershed, R les lignes de référence, $L_{D(x)}$ la surface obtenue par x dilatation de L , et $R_{D(x)}$ la surface obtenues par x dilations successives de R

Les « vrais positifs » sont les points identifiés comme limite par le watershed et dans l'image de référence.

$$TP = L \cap R_{D(x)}$$

Les « faux positifs » sont points identifiés comme limites par le watershed mais pas dans l'image de référence

$$FP = L \setminus TP = L \cap \overline{R_{D(x)}} \text{ Où « } \setminus \text{ » est la différence ensembliste et « } - \text{ » est le complémentaire}$$

Les « faux négatifs » sont les points non retenus par le watershed mais identifiés comme limites dans l'image de référence.

$$FN = \overline{L_{D(x)}} \cap R$$

Obtenir les effectifs des classes consiste à dénombrer les points noirs dans les images précédentes. Les indices de précision et de rappel sont définis comme suit :

$$\text{Précision} = \text{spécificité} = \frac{VP}{VP + FP}$$

$$\text{Rappel} = \text{sensibilité} = \frac{VP}{VP + FN}$$

Ce qui donne les résultats suivants :

Images	VP ₃	FP ₃	FN ₃	précision	rappel
RGB	62618	158989	30610	0,28256328	0,67166516
RNirPan	84249	170182	13437	0,33112710	0,86244702
RNir	79207	160982	65303	0,32976947	0,54810740

Les indices varient entre 0 et 1, 0 quand la comparaison est mauvaise et 1 quand elle est bonne. Les résultats produits ne sont pas satisfaisants, car :

- le "nettoyage expert" supprime une partie des lignes ne faisant pas partie des champs de crevette, alors que l'algorithme prend en compte toute l'image.
- les indices sont établis pour comparer des images de même granularité : il est nécessaire de développer des indices idoines permettant de prendre en considération les zones sur-segmentées ou les zones sous-segmentées, i.e. comparer un bassin de l'image de références avec la somme de n bassins de l'image calculée (et inversement).

3.3.3 Intérêt des approches multi-échelles pour maîtriser la sur-segmentation du watershed

La conclusion du §2.1 insiste sur la nécessité de mettre en œuvre des méthodes permettant de maîtriser voire de supprimer la sur-segmentation du watershed. Nous y avons parlé essentiellement des méthodes de filtrage des minima locaux a priori ou de fusion de bassins similaires a posteriori. Ici nous mettons à profit le jeu d'essai multi-spectral pour aborder le problème sous l'angle du filtrage multi-échelles : du fait, de leur forte sur-segmentation, les images de champs de crevettes se prêtent mieux que les images des cellules à ce type d'expérimentation.

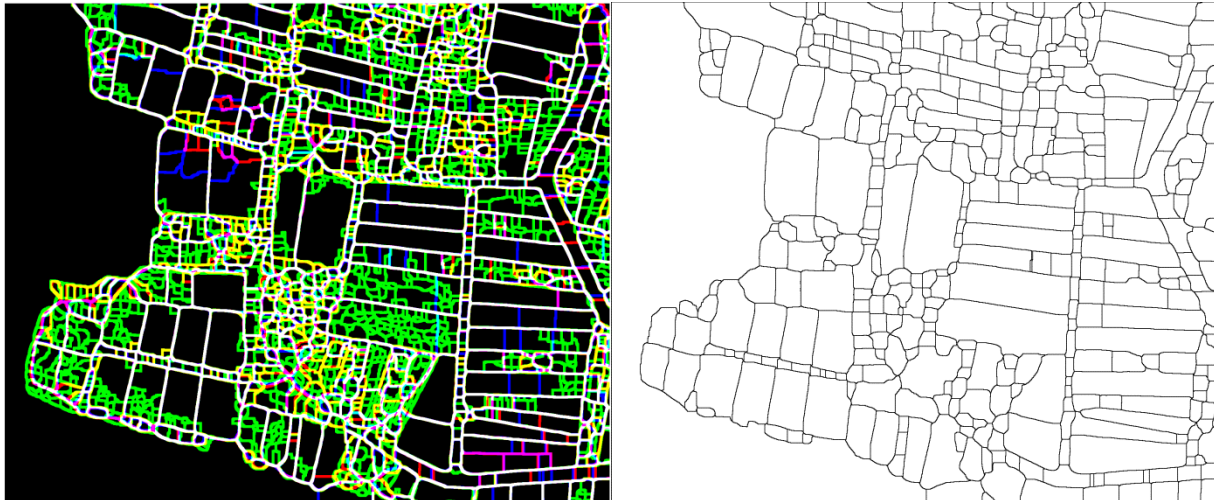
Des tests ont été menés en utilisant une pyramide de gaussiennes pour filtrer progressivement les minima locaux de l'image et identifier les lignes de crête invariantes au lissage (figure 48). Les bassins qui disparaissent entre deux niveaux consécutifs de la pyramide de lissages sont ignorés ; leur germe (minimum local) est considéré comme du bruit.



Clichés de l'auteur

Figure 48. Variation arbitraire (0, 3, 6 et 10) de la valeur du flou minimal lors du DoG, avant l'application du watershed sur l'image RedNirPan

Les lignes de crête sont sensiblement modifiées à chaque filtrage (figure ci-dessus), les germes de l'algorithme n'étant pas identiques : l'idée de ce filtrage « multi-échelle » consiste à ne conserver que les lignes qui apparaissent à tous les niveaux de filtrage. Les différents résultats sont fusionnés en ne gardant que les points de poids les plus élevés c'est-à-dire communs à plusieurs images (cf. figure 49).



Clichés de l'auteur

Figure 49. A gauche, chaque watershed est tracé dans une couleur primaire : les points communs aux quatre watershed apparaissent en blanc. A droite, la subdivision blanche correspondant aux arêtes communes aux quatre watershed notée par le symbole $\cap$.

Images	VP	FP	FN	précision	rappel
$\cap$ RNirPan (3)	51674	71056	35031	0,42103805	0,59597486
$\cap$ RNirPan (5)	69289	53441	20060	0,56456449	0,77548713

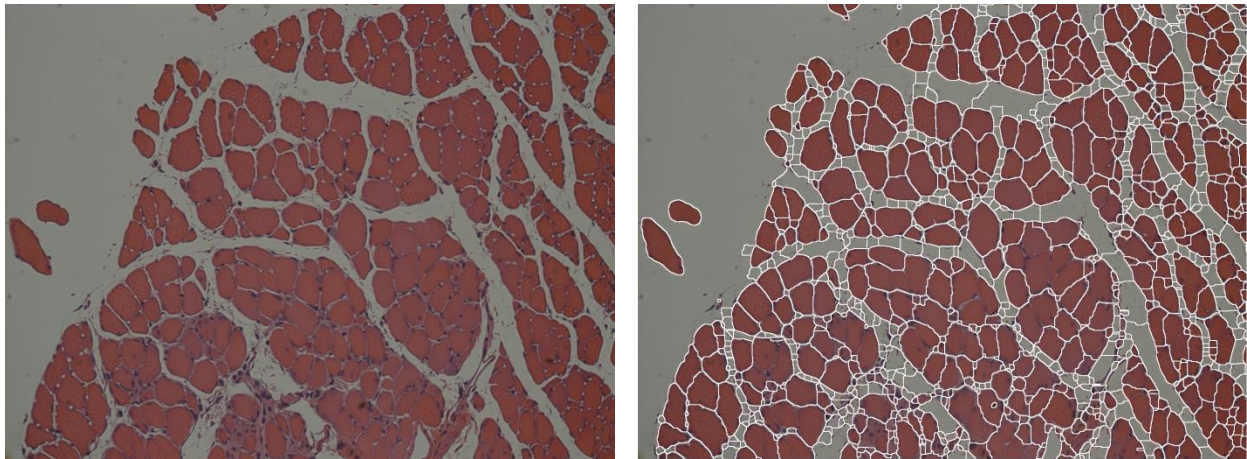
La segmentation avec l'application multi-échelles empirique améliore les résultats. Il est difficile de juger de la pertinence de l'indice car la sur-segmentation est principalement dans des zones « hors bassin » et des limites légèrement décentrées. Toutefois, le tableau semble montrer des meilleurs résultats quand on augmente le nombre de niveaux de la pyramide de flou. Une telle méthode serait donc à approfondir pour contrôler la segmentation, à condition de savoir maîtriser le niveau (nombre) et le pas (valeur) de flou à appliquer.

3.3.4 Généricité de la méthode pour la détection de cellules

La méthode de segmentation a été développée pour identifier des cellules végétales à partir d'images acquises au microscope. Mais qu'en est-il pour la détection d'autres types de cellules ?

On s'intéresse à des images de cellules de muscle de souris qui devraient être bien adaptées à une segmentation par watershed car elles sont séparées par un espace intercellulaire rappelant la paroi des cellules végétales. L'idée est de segmenter le contour des cellules pour les mesurer automatiquement.





Clichés de Melissa Bowerman (2013)

Figure 50. Image native de cellule de muscle et image de la segmentation automatique en surimpression. Les cellules sont pour la plus part individualisées. On observe des phénomènes de sur- et sous-segmentation.

La segmentation des cellules semble satisfaisante (figure 50). 88% des cellules sont correctement segmentées, si on exclut les bassins présents sur le fond de l'image. On observe des sous-segmentations sur certaines cellules, car le contraste de la limite intercellulaire est trop faible, le minima local sera commun pour les cellules.

La méthode peut segmenter des images cellulaires autres que végétales, si le contenu de l'image possède les mêmes caractéristiques photométriques, i.e. globalement deux couleurs dont une représente les objets structurants (ici les cellules).

3.4 Utilisation dans un questionnaire biologique

Nous allons nous intéresser à une étude sur les files cellulaires réalisées dans l'UMR AMAP et aux apports que pourrait apporter l'automatisation du traitement des files cellulaires dans ce contexte.

3.4.1 Contexte d'étude

La croissance d'une plante se décompose en une succession dans le temps d'évènements biologiques tels l'allongement des tiges, la mise en place des feuilles, la floraison ou encore la fructification. La phénologie des plantes représente l'étude du déroulement et de la durée de ces évènements. Elle est connue pour être, en plus des facteurs endogènes, modulée par l'environnement et le climat, qui en contexte méditerranéen impose à la végétation des conditions de croissance particulières rythmées par l'alternance de deux phases de dormance et de deux phases de végétation. Dans ce contexte d'étude nous avons choisi de travailler sur la

réactivation cambiale et dynamique de la xylogénèse en fonction de l'architecture chez *Pinus brutia*. C'est-à-dire sur l'étude des files cellulaires à partir de l'assise cambiale.

3.4.2 Méthodologie

Les prélèvements sont issus d'un micro-carottage, pris au milieu d'une unité de croissance. Ils sont ensuite préparés en suivant un protocole de coupe histologique et mis sous lame et lamelle. La numérisation de chacun des prélèvements est faite à l'aide d'un microscope Leica muni d'une platine, d'une caméra. Pour chaque prélèvement, des photos du cambium et du cerne en cours de formation sont prises, afin de déterminer les vitesses de croissance (nombre de cellules en fonction du temps). Pour chacune des phases de croissance, le nombre de cellules dans chaque zone est compté manuellement dans 10 files cellulaires à l'aide du logiciel de traitement d'images ImageJ.

Méthode de comptage des cellules de dix files radiales avec le logiciel de traitement d'images ImageJ.

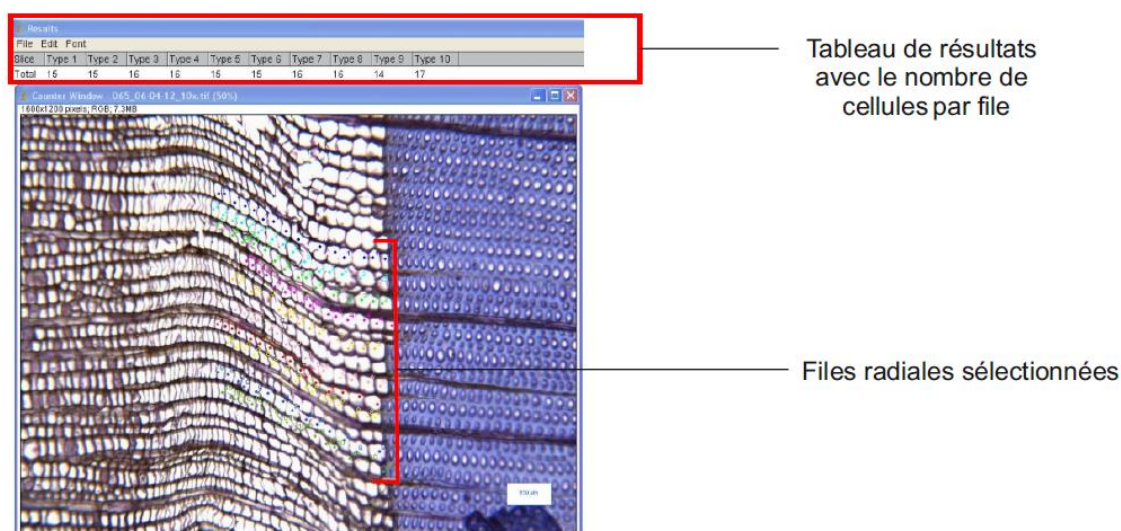


Illustration de Charlène Arnaud (UMR AMAP)

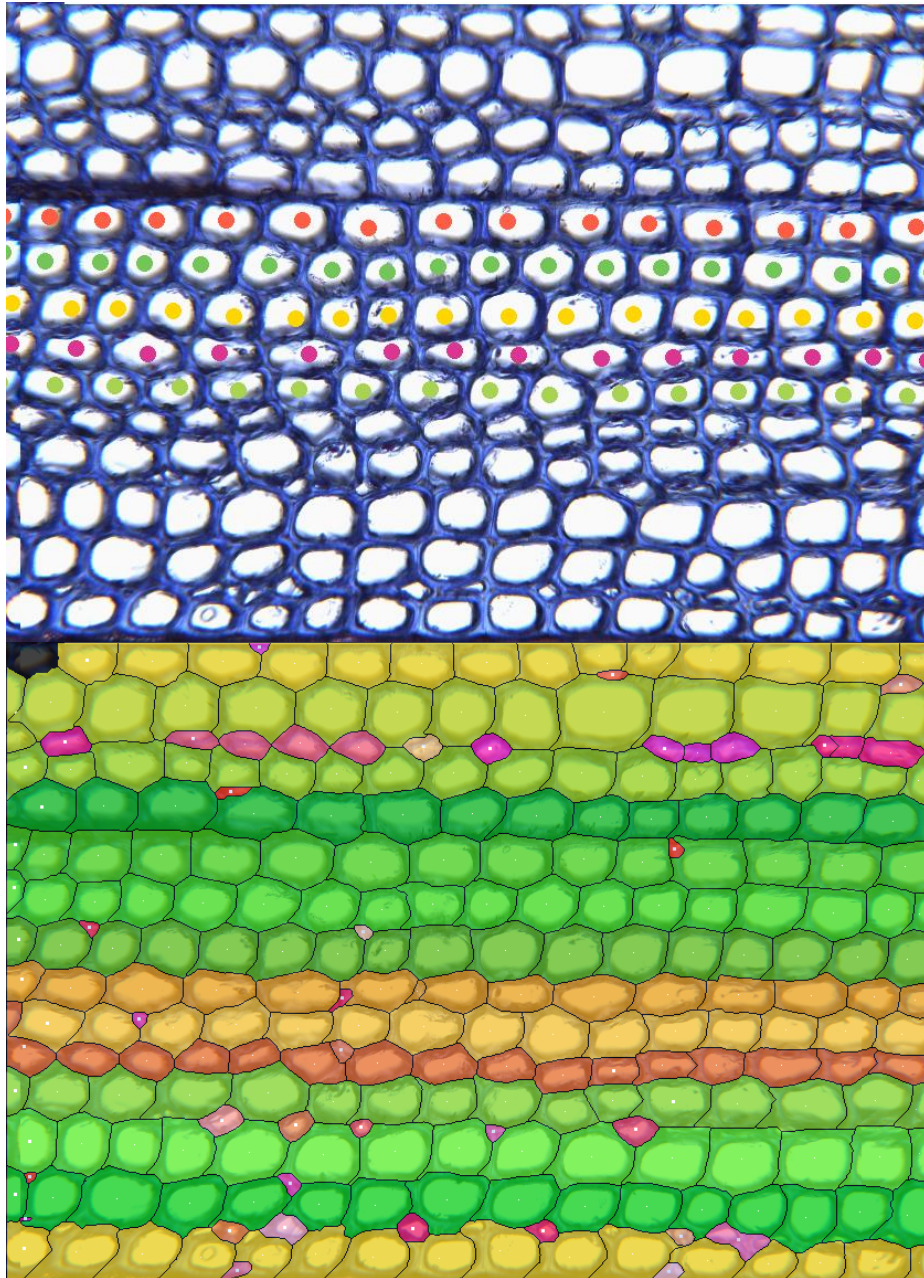
Figure 51. Utilisation d'ImageJ pour le comptage manuel des cellules de file cellulaire. Pendant la première phase de maturation des cellules, on peut voir que l'arrangement des files cellulaires n'est pas rectiligne.

Le comptage de cellules pour 10 files cellulaires sur un cerne prend environ 10 minutes hors préparation. Pour l'étude, il y a 11 prélèvements par arbre, sur une dizaine de représentants.

3.4.3 Automatisation des processus

Pour répondre aux remarques des rapporteurs, nous avons commencé à traiter des images ayant été annotées pendant l'étude. Seuls des résultats préliminaires et expérimentaux sont présentés lors de la soutenance orale. Nous avons testé l'intégralité de notre méthode d'identification

automatique des files cellulaires sur des portions de bande d'image car nous n'avons plus à disposition les images non fusionnées.



Cliché de l'auteur

Figure 52. En haut, comptage manuel des cellules de 5 files sur une coupe transversale de Pin. En bas, identification et caractérisation automatiques des files cellulaires. Le code couleur correspond de l'indice de fiabilité. On peut voir par superposition le détournage automatique des cellules.

Le traitement d'une bande image par la méthode automatique est inférieur à 10 secondes sur une machine standard (Processus double cœur Intel ® Xeon ® 2.27 GHz avec 6 Go de RAM). La totalité de l'image est traitée. Seules les cellules du bord de l'image ne sont pas prises en compte

car elles sont tronquées. Contrairement à l'approche manuelle, la méthode va, en plus du comptage des cellules par file, produire d'autres types de résultats (surface, circularité, épaisseur de la paroi, etc.) et donner des indices des fiabilités pour caractériser et/ou filtrer les files et les résultats.

3.4.4 Conclusion sur l'automatisation dans le cadre d'un questionnement biologique

La méthode que nous proposons annote une image six fois plus rapidement que l'expert et permet un traitement complet de l'image. Plus que le gain de temps, la méthode va permettre d'accéder automatiquement à d'autres types de mesure permettant l'ouverture à de nouveaux questionnements scientifiques. L'étude sur le Pin aurait pu aborder la "Dynamique de la différenciation et de la maturation cambiale chez *Pinus brutia*. Quelles contraintes endogènes ?".

4 CONTRIBUTIONS ET PERSPECTIVES

Ce travail traite de la difficulté de mise en place des processus de mesures automatiques de données biologiques à partir d'images numériques dans un cadre d'approches statistiques. Les questions qui m'ont motivé portent sur :

- l'incidence des choix méthodologiques sur la stabilité des résultats, i.e. comment les incertitudes des différentes briques d'un processus complexe peuvent ou ne peuvent pas se compenser,
- l'étude de la validation des mesures produites, i.e. quels outils ou approches peuvent être développés pour « garantir » l'exactitude des mesures réalisées,
- la généricité réelle des méthodes et modèles appliquées à la biologie végétale, i.e. à partir de quand une méthode doit être spécialisée pour produire des résultats utilisables par des botanistes ou des anatomistes,...

Autant de questions « philosophiques » qui ont conditionné l'approche et la réalisation de mon travail. L'approche et la réflexion se veulent volontairement détachées, pas vraiment théoriques mais plutôt abordées dans l'absolu, en dehors de tout contexte applicatif... L'idée est bien de se détacher du cadre d'étude pour être le plus générique possible. Bien évidemment, une fois formalisés, concepts et méthodes sont ensuite spécialisés, i.e. adaptés à l'identification des files cellulaires dans des mosaïques d'images, voire plus quand le temps l'a permis.

Répondre à ces questions nécessitent d'adopter une approche hybride consistant à plaquer sur une approche classique un regard ou un questionnement plus large que nécessaire, plus critique, peut-être plus profond...

L'objectif de ce travail était le développement de procédures de détection automatique d'organisations cellulaires à partir de mosaïques d'images microscopiques de bois. L'étude des tendances biologiques le long des files est en effet nécessaire pour comprendre la mise en place des différentes organisations et maturations de cellule. Elle ne peut être conduite qu'à partir d'une grande zone d'observation du plan ligneux. La méthode de préparation des échantillons biologiques et la mise au point d'algorithmes devaient permettre d'acquérir une description (organisation, distribution et dimension) des objets étudiés en des temps raisonnables, à des fins d'analyses statistiques. La méthode s'applique aussi bien sur le clade des Gymnospermes que sur celui des Angiospermes. Les acquisitions de coupes histologiques colorées de bois ou de plan ligneux poncé peuvent être traitées par la méthode.

Le premier apport de ce travail est la mise au point d'une chaîne de préparation des échantillons permettant l'acquisition de mosaïque d'images. Les principaux résultats obtenus sont les suivants.

1. La définition d'un protocole de préparation de rondelles de bois poncées adapté à l'étude de grandes zones, qui présente l'avantage d'être économe en temps et sans biais de découpe physique (a contrario de la coupe histologique) pour la construction d'une mosaïque.
2. Un protocole pour traiter des images produites à partir de coupes histologiques. Cette méthode permet d'obtenir un meilleur contraste entre les structures biologiques (lumen/paroi) par la coloration des échantillons.
3. La définition d'un protocole de large numérisation des échantillons, permettant, grâce à une succession d'images, de former une mosaïque couvrant la zone d'observation.

Un apport de la thèse est la définition d'un processus générique de détection automatique d'organisations, que l'on a appliqué au cas des files cellulaires. Les principaux résultats obtenus sont :

1. La mise en place d'une chaîne de traitement pour la segmentation et l'individualisation des cellules de chaque image à partir d'un algorithme de ligne de partage des eaux par inondation. L'image est tout d'abord nettoyée à l'aide de filtres standards puis un opérateur multi-échelles permet de rehausser le contraste entre les structures avant la segmentation. Les cellules sont détournées au niveau de la lamelle moyenne. L'erreur commise sur la mesure de surface des bassins est faible (en moyenne de 2,81% par rapport à une segmentation manuelle experte). Ce pourcentage est du même ordre que l'écart d'inter-variabilité expert. Elle est principalement due à la résolution de l'image qui ne permet pas de savoir, au pixel près, où se trouve la ligne de crête. Mais, à l'inverse des approches expertes, notre méthode est répétable. Les déformations optiques et de planéité génèrent souvent dans les images des flous périphériques. Notre méthode est insensible au flou de l'image, car elle se base sur des minima locaux stables au flou. Au terme de la segmentation, les cellules peuvent être mesurées sans spatialisation.
2. La définition d'un graphe d'adjacence des cellules qui décrit les relations de voisinage intercellulaires. Chaque arête du graphe relie les centres géométriques de deux bassins voisins, i.e. incidents à une même ligne de crête.
3. La détection d'organisations dans le graphe à l'aide de règles de voisinage prenant la forme de fonction à minimiser. Pour les files cellulaires, nous exploitons une règle géométrique (fonction de linéarité) et une règle morphologique (fonction de similarité). Les alignements sont construits l'un après l'autre, par agrégations successives de sommets pris dans des voisinages progressivement élargis. En moyenne 80% des files d'une image sont détectées, même avec des déformations biologiques telles que la présence des vaisseaux. Ces déformations peuvent entraîner des coupures au sein de files formant ainsi des tronçons fragmentés. Une méthode basée sur des règles topologiques d'encadrement entre files permet de concaténer les tronçons appartenant à une même file.

Pour ne pas créer de faux positifs seuls les cas simples sont automatiquement traités, les autres étant laissés à l'appréciation de l'utilisateur.

4. La mise en place d'une fonction de fusion des données permettant de suivre les files cellulaires sur l'ensemble du plan ligneux. L'agrégation des files d'images consécutives dans la mosaïque est réalisée par fusion de graphes à partir d'un noyau de fusion, c'est-à-dire à partir d'un ensemble de couples de nœuds à fusionner. Cet ensemble est obtenu à partir d'une fonction définie par un produit de termes génériques normalisés devant être spécialisés. Pour les files cellulaires, seuls les aspects géométriques – position et taille – sont nécessaires pour identifier les cellules candidates à la fusion. La fusion des graphes est facilitée par la connaissance du déplacement de la platine permettant de générer les images constituant la mosaïque.
5. La définition d'indices de fiabilité permettant de caractériser la pertinence des mesures produites au niveau des files, des cellules ou des composants cellulaires. Ces indices sont des produits de différents termes normalisés. Chaque terme est une transcription d'un facteur d'influence sur la mesure. Ces indices permettent ainsi de sélectionner les files les plus significatives ou de mettre en avant des perturbations ou configurations biologiques particulières. En pratique, la relative faiblesse potentielle du nombre de files utilisables est compensée par la quantité d'images que l'on peut traiter.

Les méthodes développées dans la thèse permettent l'acquisition et le traitement rapide d'un volume important de données. Ces données devraient permettre de conduire de nombreuses investigations : des analyses architecturales des arbres avec le suivi de file cellulaire et/ou la détection d'évènements biologiques remarquables ; des analyses de variabilité intra- et inter-arbres pourront être lancées pour mieux comprendre la croissance endogène des arbres.

Pour des échantillonnages futurs, il serait intéressant de prévoir du matériel supplémentaire pour éprouver la méthode à d'autres images issues de rondelles poncées. Nous avons à notre disposition pour le jeu d'essai principalement des essences de Gymnospermes et assez peu d'Angiospermes. Pour connaître/évaluer les limites de la méthode, il serait nécessaire, maintenant, de tester la méthode sur un large panel d'essences.

La thèse étant orientée vers la conduite ultérieure d'analyses scientifiques plutôt que vers l'application industrielle, la priorité a été mise sur la robustesse et la précision des algorithmes. Pour une utilisation en industrie, une optimisation des algorithmes pourra encore augmenter la vitesse des traitements.

A court terme, plusieurs points abordés dans cette étude gagneraient à des développements complémentaires :

- L'étude du paramétrage des filtres pour le traitement d'images issues de bois poncé. Les images des deux protocoles ayant des signatures différentes, il ne serait pas étonnant de devoir créer deux jeux de paramètres à appliquer en fonction du protocole. Une

information complémentaire devra donc être rajoutée sur l'interface du plugin pour que l'utilisateur puisse indiquer la nature de préparation de ses échantillons.

- L'approfondissement des méthodes de segmentation des objets structurants. Comme nous l'avons vu au §2.1.3, des variantes du watershed devraient permettre une meilleure initialisation des germes de départ et une fusion plus appropriée des bassins surnuméraires basées non plus sur les arêtes mais sur des informations portés par les bassins (taille, texture, etc.). De telles méthodes rendraient les détections plus robustes et devraient pouvoir couvrir une plus grande variabilité d'images.
- L'implémentation de nouvelle méthode de modélisation des files présentée en fin du §2.2.3 puis sa comparaison avec la méthode actuelle. La nouvelle méthode devrait accélérer la détection des files et réduire le nombre des étapes. De par le fonctionnement de cette dernière, les tronçons de file devraient être automatiquement raccordés dès leur création.
- Compléter et valider les indices de fiabilités présentés au §2.4.2.

A plus long terme, il s'agira d'étendre l'analyse à d'autres caractéristiques. Comme nous l'avons vu au §2.2.1, les végétaux possèdent d'autres types d'alignements dans des tissus (spiral, cercle, etc.). Théoriquement la méthode doit pouvoir traiter ces arrangements car nous l'avons pensé et implémentée pour qu'elle permette la spécialisation (générique), c'est-à-dire qu'on puisse injecter des connaissances expertes pour traiter différentes caractéristiques. Une étude devrait être menée pour compléter le typage biologique. Cette partie, de par sa difficulté apparente, n'a été que peu abordée. Il faudrait pouvoir à long terme rendre ces développements aussi fiables (à l'image des indices de fiabilité sur les files) et pouvoir définir des indices de pertinence pour l'ensemble des différents types biologiques.

5 REFERENCES BIBLIOGRAPHIQUES

Andres E. 1994. Discrete circles, rings and spheres. *Computer & Graphics* **18**: 695–706.

Baggett D, Nakaya M, McAuliffe M, Yamaguchi TP, Lockett S. 2005. Whole cell segmentation in solid tissue sections. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **67**: 137–143.

Bancroft JD, Gamble M. 2002. *Theory and practice of histological techniques*. London; New York: Churchill Livingstone.

Barghout L, Lawrence WL. 2003. Perceptual information processing system. : 543.

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

Bay H, Tuytelaars T, Gool LV. 2006. SURF: Speeded Up Robust Features. *Proceedings of the ninth European Conference on Computer Vision*.

Beucher S. 2012. Towards a unification of waterfalls, standard and P algorithms. *CMM/Mines Paristech publication*.

Beucher S, Marcotegui B. 2009. P algorithm, a dramatic enhancement of the waterfall transformation.

Bray JR, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27**: 325.

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

Brown M, Lowe DG. 2007. Automatic Panoramic Image Stitching using Invariant Features. *International Journal of Computer Vision* **74**: 59–73.

Busin L, Shi J, Vandenbroucke N, Macaire L. 2009. Color space selection for color image segmentation by spectral clustering. *2009 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 262–267.

Canny J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Clair B, Gril J, Di Renzo F, Yamamoto H, Quignard F. 2007. Characterization of a Gel in the Cell Wall To Elucidate the Paradoxical Shrinkage of Tension Wood. *Biomacromolecules* **9**: 494–498.

- Clienti C, Beucher S. 2010.** Architectures flot de données dédiées au traitement d'images par Morphologie Mathématique.
- Cormen TH, Cazin X, Kocher G-L. 2001.** *Introduction to Algorithms*. MIT Press.
- Damiand G, Resch P. 2003.** Split-and-merge algorithms defined on topological maps for 3D image segmentation. *Graphical Models* **65**: 149–167.
- Davies ER. 2005.** *Machine vision: theory, algorithms, practicalities*. Amsterdam ; Boston: Elsevier.
- Duda RO. 2001.** *Pattern classification*. New York: Wiley.
- Dunn JC. 1973.** A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* **3**: 32–57.
- Einevoll GT, Plesser HE. 2005.** Response of the difference-of-Gaussians model to circular drifting-grating patches. *Visual Neuroscience* **22**: 437–446.
- Fourcaud T, Zhang X, Stokes A, Lambers H, Körner C. 2008.** Plant Growth Modelling and Applications: The Increasing Importance of Plant Architecture in Growth Models. **101**: 1053–1063.
- Frontera-Pons J, Ovarlez J-P, Pascal F, Chanussot J. 2013.** Performance Analysis of Robust Detectors for Hyperspectral Imaging. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*. Melbourne, Australie, 1–4.
- Gindl W. 2001.** Cell-wall lignin content related to tracheid dimensions in drought-sensitive austrian pine (pinus nigra). *Iawa Journal* **22**: 113–120.
- Guo Y, Fourcaud T, Jaeger M, Zhang X, Li B. 2011.** Plant growth and architectural modelling and its applications. *Annals of Botany* **107**: 723–727.
- Heinrich L. 2007.** Reaction wood varieties caused by different experimental treatments. *TRACE, Tree Rings in Archaeology, Climatology and Ecology* **5**: 224–232.
- Hitz OM, Gärtner H, Heinrich I, Monbaron M. 2008.** Wood anatomical changes in roots of European ash (Fraxinus excelsior L.) after exposure. *Dendrochronologia* **25**: 145–152.
- Igathinathane C, Pordesimo LO, Columbus EP, Batchelor WD, Methuku SR. 2008.** Shape identification and particles size distribution from basic shape parameters using ImageJ. *Computers and Electronics in Agriculture* **63**: 168–182.
- Igathinathane C, Prakash VSS, Padma U, Babu GR, Womac AR. 2006.** Interactive computer software development for leaf area measurement. *Computers and Electronics in Agriculture* **51**: 1–16.

- Inglada J, Christophe E. 2009.** The Orfeo Toolbox remote sensing image processing software. *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*. IV–733–IV–736.
- Jeacocke MB, Lovell BC. 1994.** A multi-resolution algorithm for cytological image segmentation. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, 1994*. IEEE, 322–326.
- Jones R, Bischof L. 1996.** A graph-based segmentation of wood micrographs. In: Billard L, Fisher NI, eds. *Computing Science and Statistics*, 28. Sydney: Interface Foundation of North America, 12–20.
- Kennel P, Subsol G, Guérault M, Borianne P. 2010.** Automatic identification of cell files in light microscopic images of conifer wood. *2010 2nd International Conference on Image Processing Theory Tools and Applications (IPTA)*. IEEE, 98–103.
- Krause C, Rossi S, Thibeault-Martel M, Plourde P-Y. 2010.** Relationships of climate and cell features in stems and roots of black spruce and balsam fir. *Annals of Forest Science* **67**: 402–402.
- Lachaud S, Catesson A-M, Bonnemain J-L. 1999.** Structure and functions of the vascular cambium. *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie* **322**: 633–650.
- Ladjal S. 2006.** Estimation du flou dans les images naturelles. RFIA, .
- Lowe DG. 2004.** Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **60**: 91–110.
- Mandelbrot BB. 1983.** *The Fractal Geometry of Nature*. Henry Holt and Company.
- Marcuzzo M, Quelhas P, Campilho A, Mendonça AM, Campilho A. 2009.** Automated Arabidopsis plant root cell segmentation based on SVM classification and region merging. *Computers in Biology and Medicine* **39**: 785–793.
- McCune B, Grace JB, Urban DL. 2002.** *Analysis of ecological communities*. Gleneden Beach, OR: MjM Software Design.
- Mcinerney T, Terzopoulos D. 1999.** T-Snakes: Topology Adaptive Snakes. *MEDICAL IMAGE ANALYSIS* **4**: 840–845.
- Moëll MK, Donaldson LA. 2001.** Comparison of segmentation method for digital image analysis of confocal microscope images to measure tracheid cell dimensions. *IAWA Journal*. Vol. 22(3): 267–288.
- Najman L, Schmitt M. 1994.** Watershed of a continuous function. *Signal Processing* **38**: 99–112.

Nicolini E, CARAGLIO Y, PÉLISSIER R, LEROY C, ROGGY J. 2003. Epicormic Branches: a Growth Indicator for the Tropical Forest Tree, *Dicorynia guianensis* Amshoff (Caesalpinaceae). *Annals of Botany* **92**: 97–105.

Otsu N. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**: 62–66.

Oyallon E, Rabin J. 2013. An Analysis and Implementation of the SURF Method, and its Comparison to SIFT. *IPOLE Journal · Image Processing On Line*.

Paiva ARC, Tasdizen T. 2010. Fast semi-supervised image segmentation by novelty selection. *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.1054–1057.

Park J, Keller JM. 2001. Snakes on the Watershed. *IEEE Trans. Pattern Anal. Mach. Intell.*: 1201–1205.

Pavez R, Mora M, Gonzalez P. 2011. Objective Comparison of Contour Detection in Noisy Images. In: Martin CS, Kim S-W, eds. Lecture Notes in Computer Science. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer Berlin Heidelberg, 599–606.

Quelhas P, Nieuwland J, Dewitte W, Mendonça AM, Murray J, Campilho A. 2011. Arabidopsis Thaliana Automatic Cell File Detection and Cell Length Estimation. In: Kamel M, Campilho A, eds. *Image Analysis and Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1–11.

Rathgeber CBK, Rossi S, Bontemps J-D. 2011. Cambial activity related to tree size in a mature silver-fir plantation. *Annals of Botany* **108**: 429–438.

Rossi S, Deslauriers A, Anfodillo T, Morin H, Saracino A, Motta R, Borghetti M. 2006. Conifers in cold environments synchronize maximum growth rate of tree-ring formation with day length. *New Phytologist* **170**: 301–310.

Rossi S, Simard S, Rathgeber CBK, Deslauriers A, Zan CD. 2009. Effects of a 20-day-long dry period on cambial and apical meristem growth in *Abies balsamea* seedlings. *Trees* **23**: 85–93.

Rozenberg P, Schüte G, Ivkovich M, Bastien C, Bastien J-C. 2004. Clonal variation of indirect cambium reaction to within-growing season temperature changes in Douglas-fir. *Forestry* **77**: 257–268.

Sabri M, Macdonald RL. 2012. Vasospasm: Measurement of Diameter, Perimeter, and Wall Thickness. In: Chen J, Xu X-M, Xu ZC, Zhang JH, eds. Springer Protocols Handbooks. *Animal Models of Acute Neurological Injuries II*. Humana Press, 473–479.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez J-Y, White DJ, Hartenstein V, Eliceiri K,

- Tomancak P, Cardona A. 2012.** Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**: 676–682.
- Schneider CA, Rasband WS, Eliceiri KW. 2012.** NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.
- Scholkopf B, Smola AJ. 2001.** *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.
- Shapiro LG. 2001.** *Computer vision*. Upper Saddle River, NJ: Prentice Hall.
- Sjö dahl M, Oreb BF. 2002.** Stitching interferometric measurement data for inspection of large optical components. *Optical Engineering* **41**: 403–408.
- Tarabalka Y, Chanussot J, Benediktsson JA. 2010.** Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition* **43**: 2367–2379.
- Thibeault-Martel M, Krause C, Morin H, Rossi S. 2008.** Cambial Activity and Intra-annual Xylem Formation in Roots and Stems of *Abies balsamea* and *Picea mariana*. *Annals of Botany* **102**: 667 –674.
- Toews M, Wells W. 2009.** SIFT-Rank: Ordinal description for invariant feature correspondence. *IEEE*, 172–177.
- Travis AJ, HIRST DJ, CHESSON A. 1996.** Automatic Classification of Plant Cells According to Tissue Type using Anatomical Features Obtained by the Distance Transform. *Annals of Botany* **78**: 325 –331.
- Vapnik V. 1963.** Pattern recognition using generalized portrait method. *Automation and Remote Control* **24**: 774–780.
- Vercauteren T, Perchant A, Malandain G, Pennec X, Ayache N. 2006.** Robust mosaicing with correction of motion distortions and tissue deformations for in vivo fibered microscopy. *Medical Image Analysis* **10**: 673–692.
- Verma OP, Hanmandlu M, Susan S, Kulkarni M, Jain PK. 2011.** A Simple Single Seeded Region Growing Algorithm for Color Image Segmentation using Adaptive Thresholding. *2011 International Conference on Communication Systems and Network Technologies (CSNT)*.500–503.
- Vincent L, Soille P. 1991.** Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**: 583–598.
- Wang M, Zhou X, Li F, Huckins J, King RW, Wong STC. 2008.** Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. *Bioinformatics* **24**: 94 –101.

Wilson BF. 1984. *The Growing Tree*. Amherst: University of Massachusetts Press.

Wyant JC, Schmit J. 1998. Large Field of View, High Spatial Resolution, Surface Measurements. *International Journal of Machine Tools and Manufacture* **38**: 691–698.

Žunić J, Hirota K. 2008. Measuring Shape Circularity. In: Ruiz-Shulcloper J, Kropatsch WG, eds. Lecture Notes in Computer Science. *Progress in Pattern Recognition, Image Analysis and Applications*. Springer Berlin Heidelberg, 94–101.

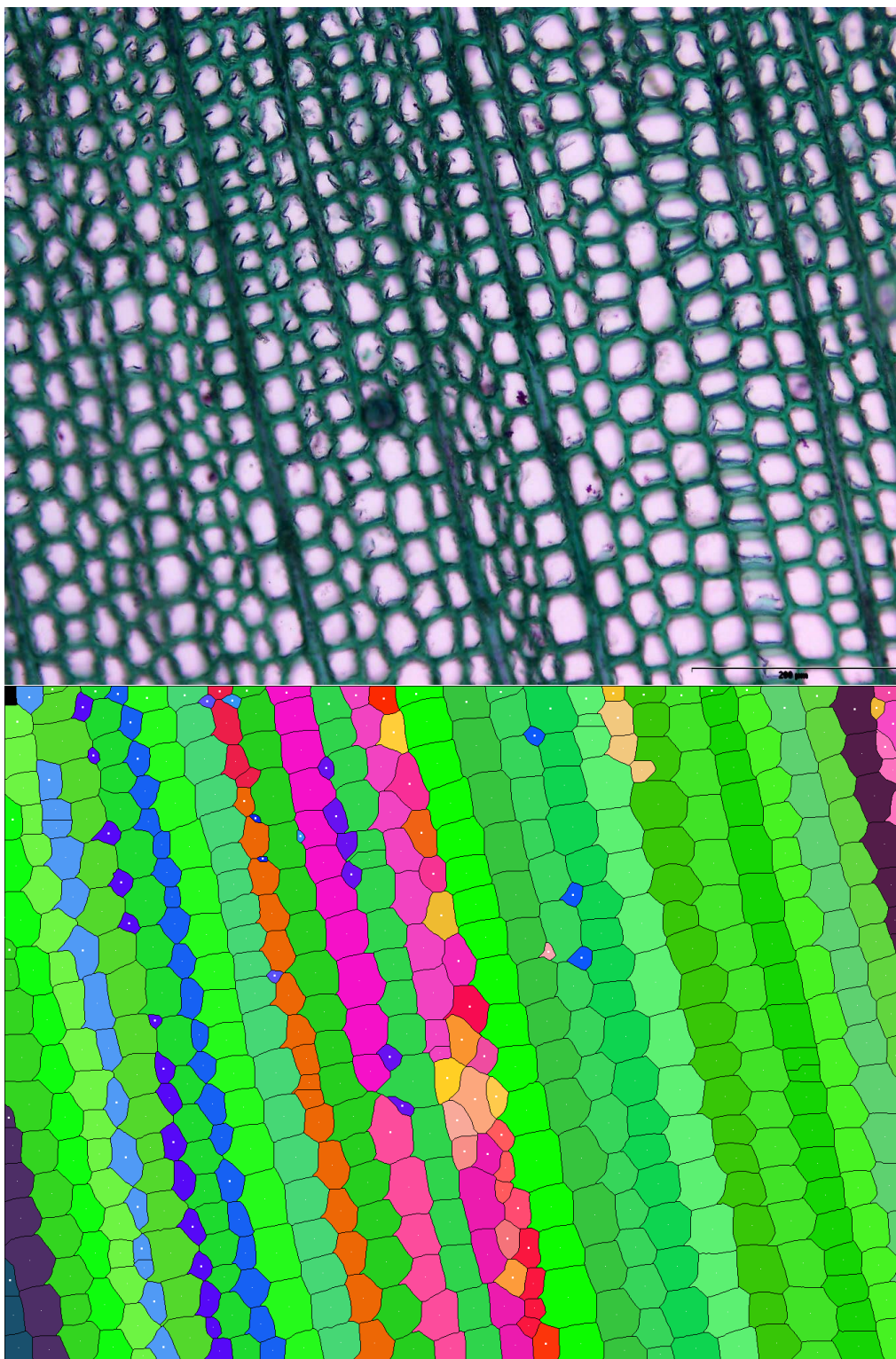
6 ANNEXES

6.1 Annexe 1 : Images et Résultats

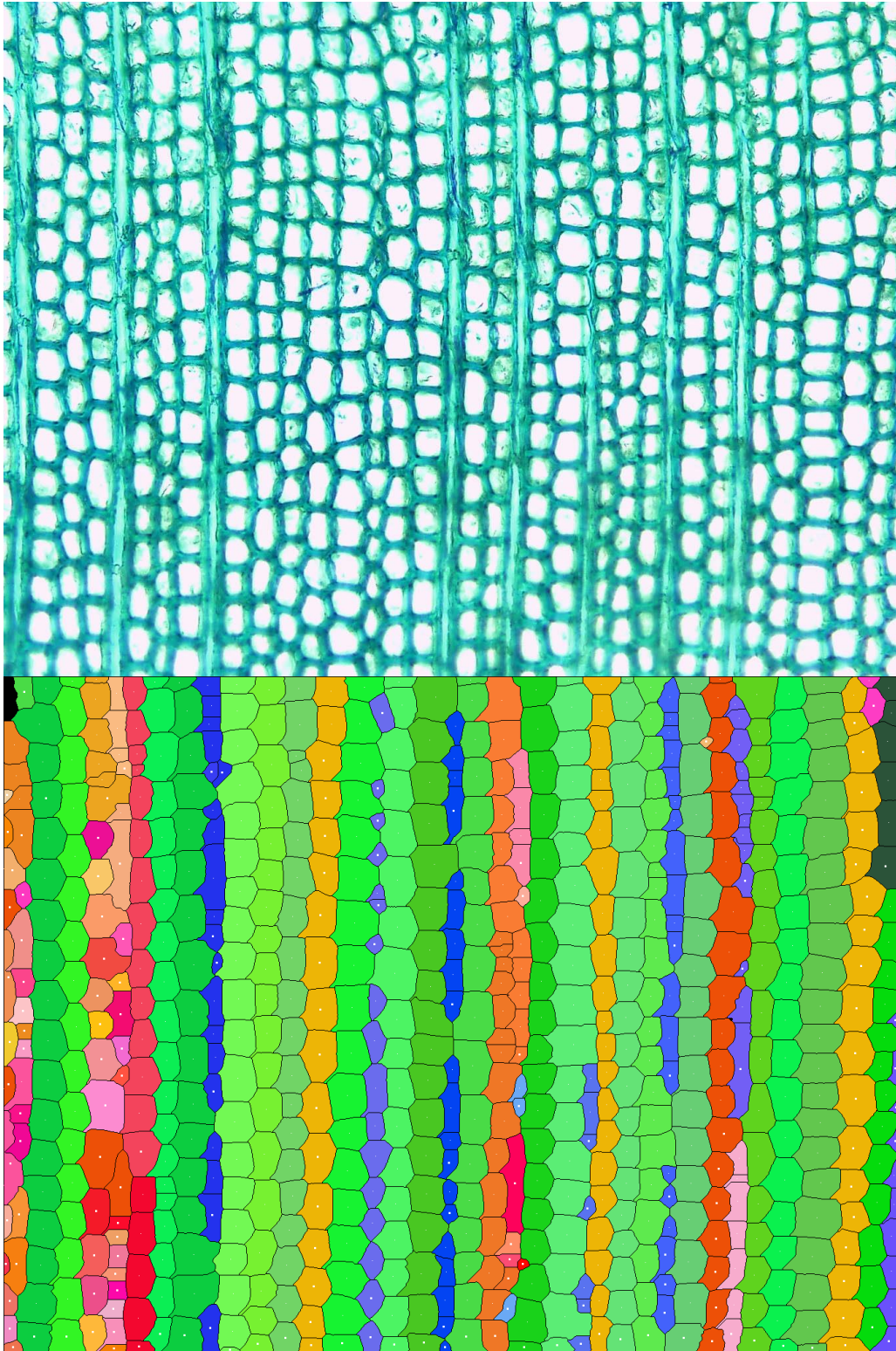
Panel d'images constituant le jeu d'essai et le résultat de la labellisation automatique de la fiabilité des files cellulaires (8 pages).

N.B. Toutes les images sont prises à un grossissement de x200.

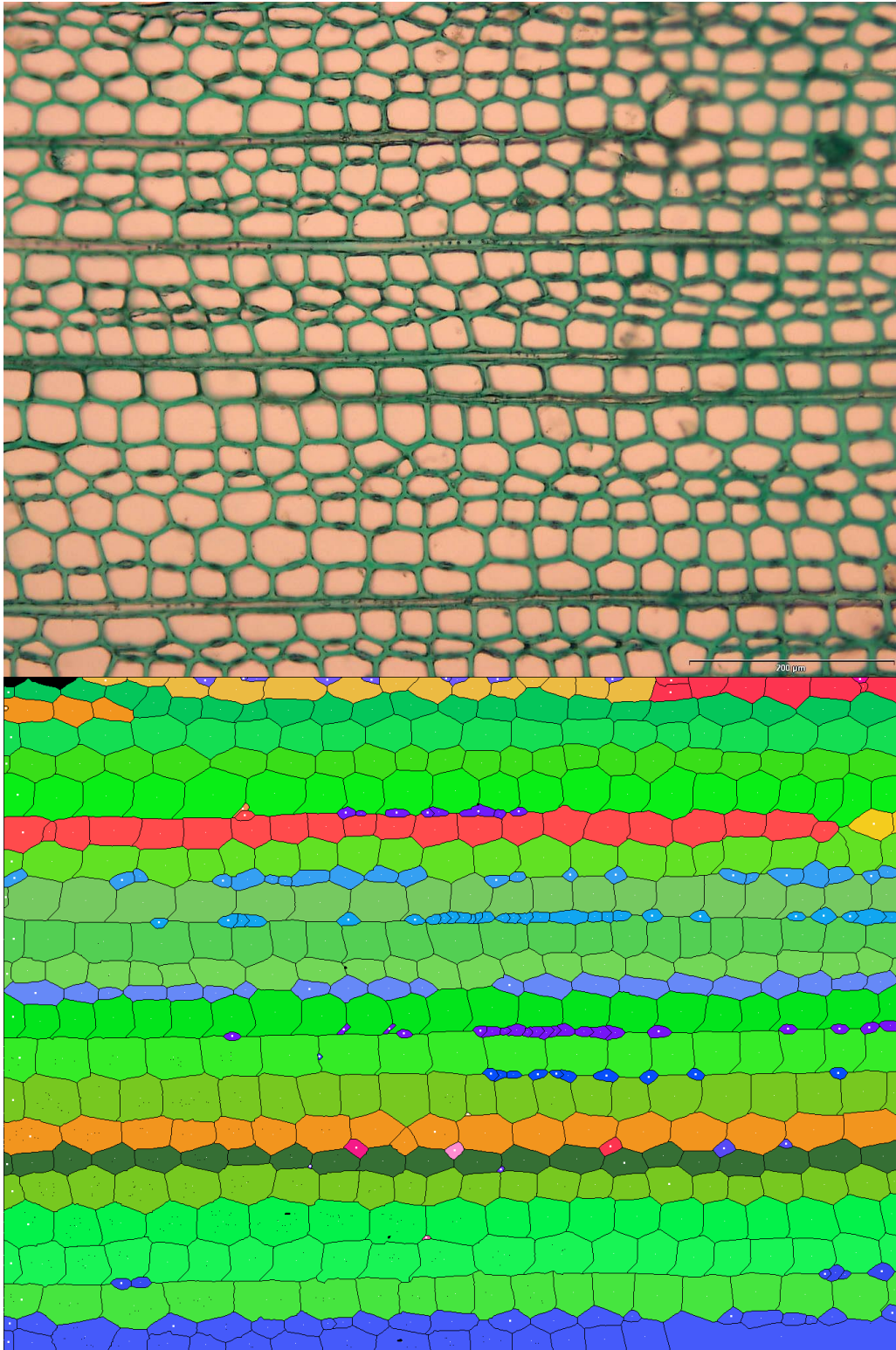
1. Coupe histologique de sapin (*Abies alba*), colorée au vert d'iode (Gymnospermes).
2. Coupe histologique de sapin (*Abies alba*), colorée au vert d'iode (Gymnospermes).
3. Coupe histologique de sapin (*Abies alba*), colorée au vert d'iode (Gymnospermes).
4. Coupe histologique d'acajou (*Pycnanthus sp.*), colorée au vert d'iode (Angiospermes).
5. Coupe histologique de sapin (*Abies alba*), non colorée (Gymnospermes).
6. Coupe histologique de pin (*Pinus nigra*), colorée à la safranine, avec un faible éclairage (Gymnospermes).
7. Coupe histologique de pin (*Pinus nigra*), colorée à la safranine, avec un fort éclairage (Gymnospermes).
8. Rondelle poncée de pin (*Pinus nigra*), non colorée (Gymnospermes).



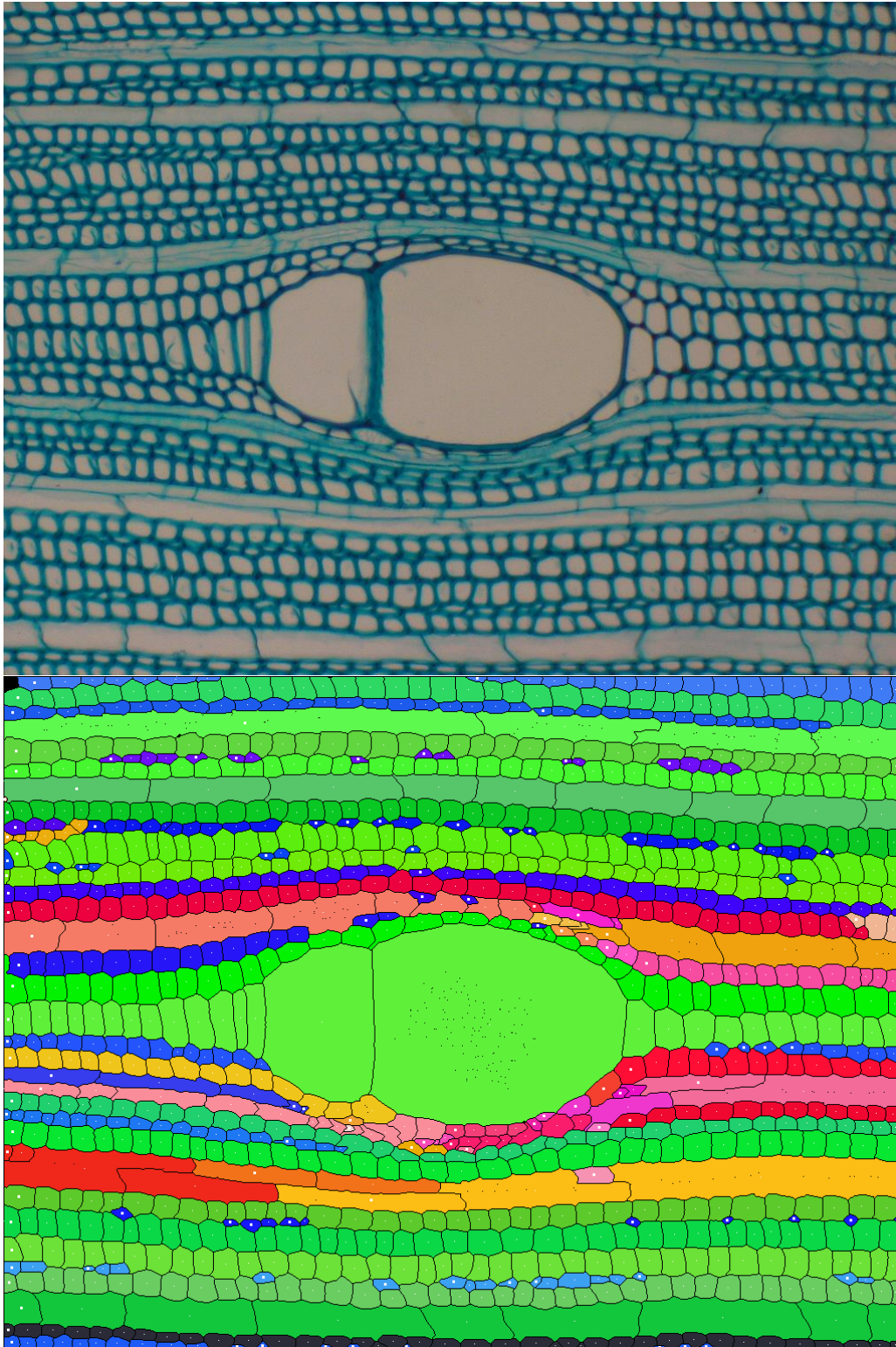
1. En haut, image native de sapin (*Abies alba*). En bas l'identification automatique des files de cellulaires.



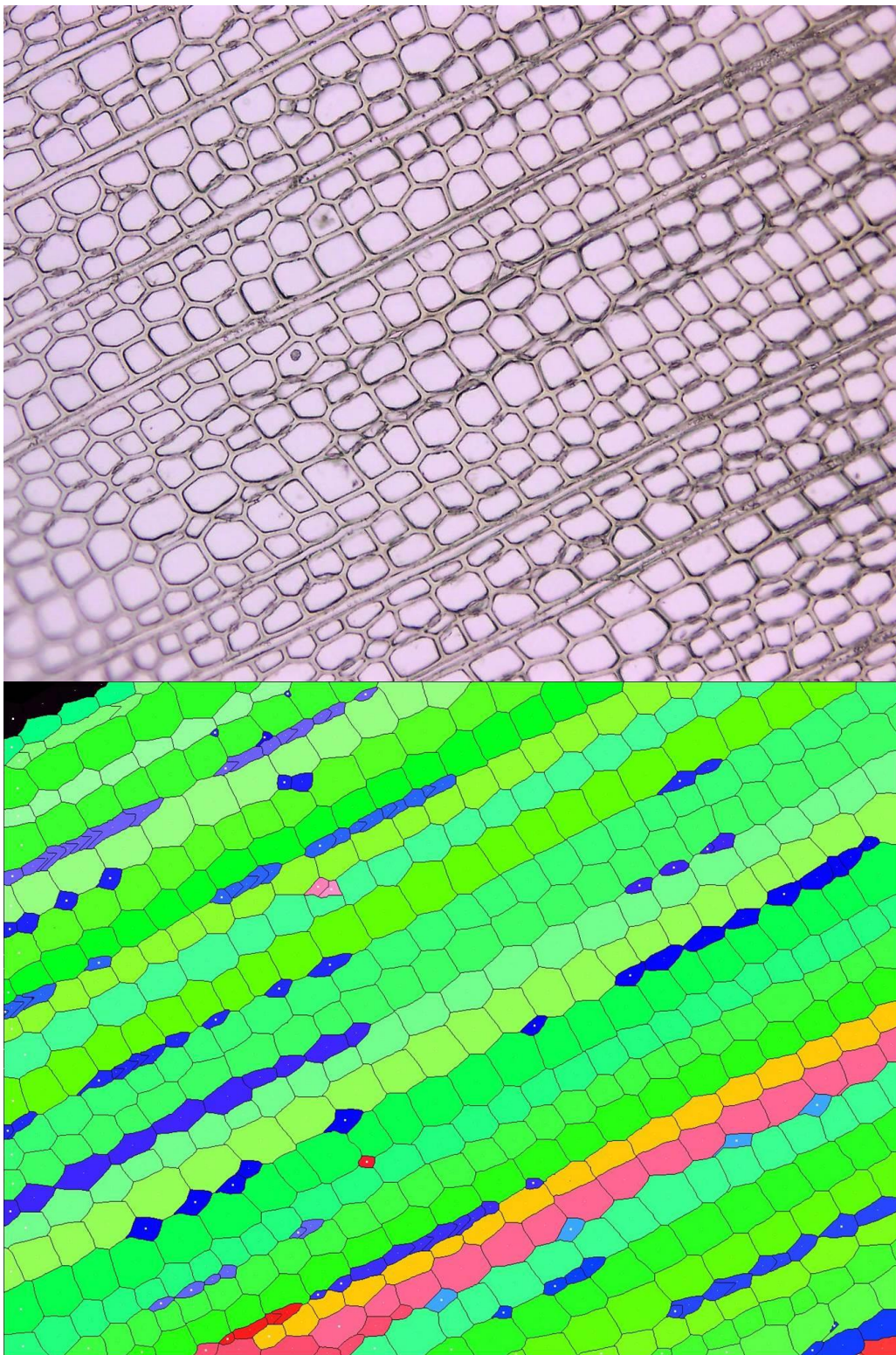
2. En haut, image native de sapin (*Abies alba*). En bas l'identification automatique des files de cellulaires.



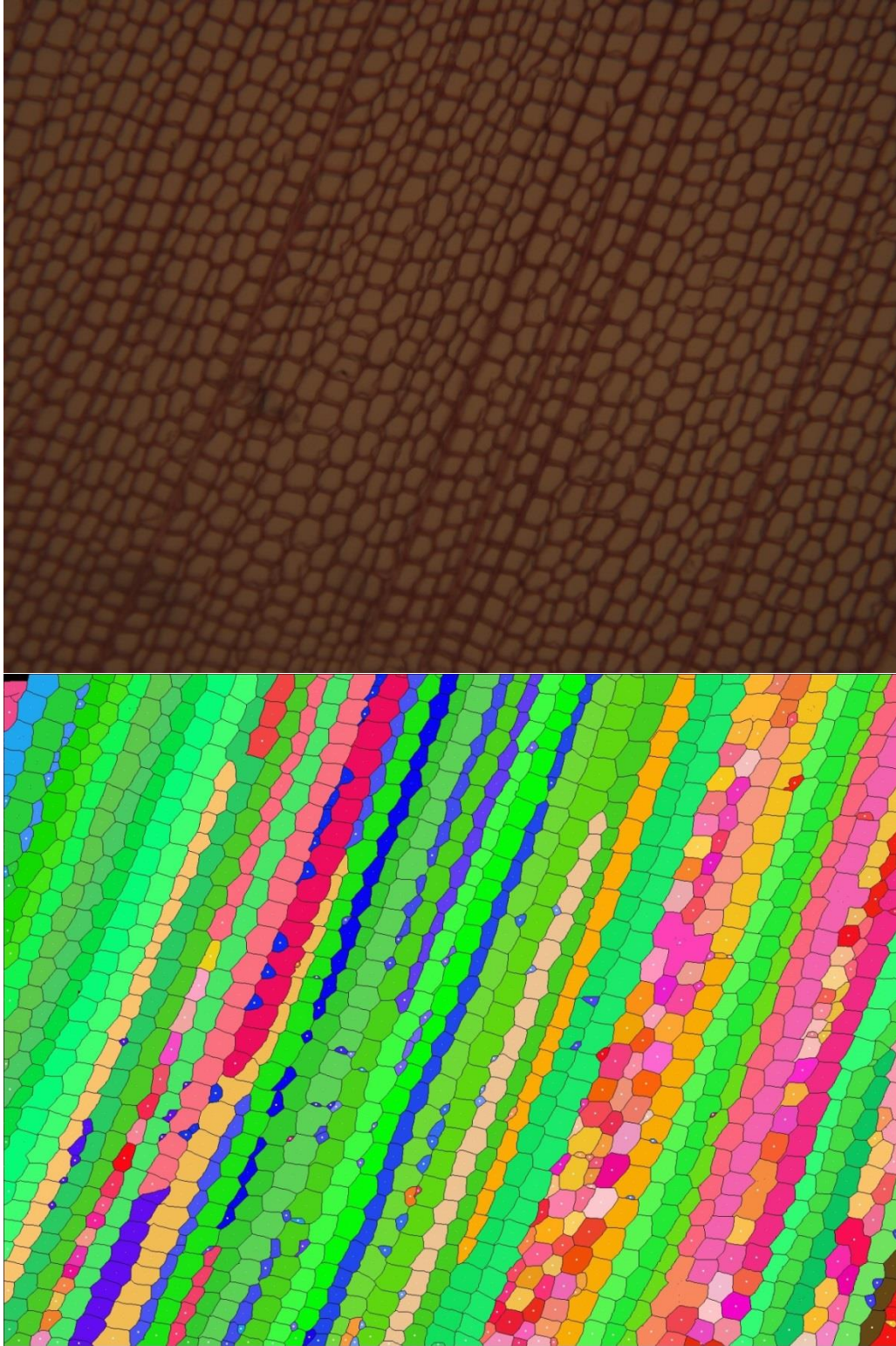
3. En haut, image native de sapin (*Abies alba*). En bas l'identification automatique des files de cellulaires.



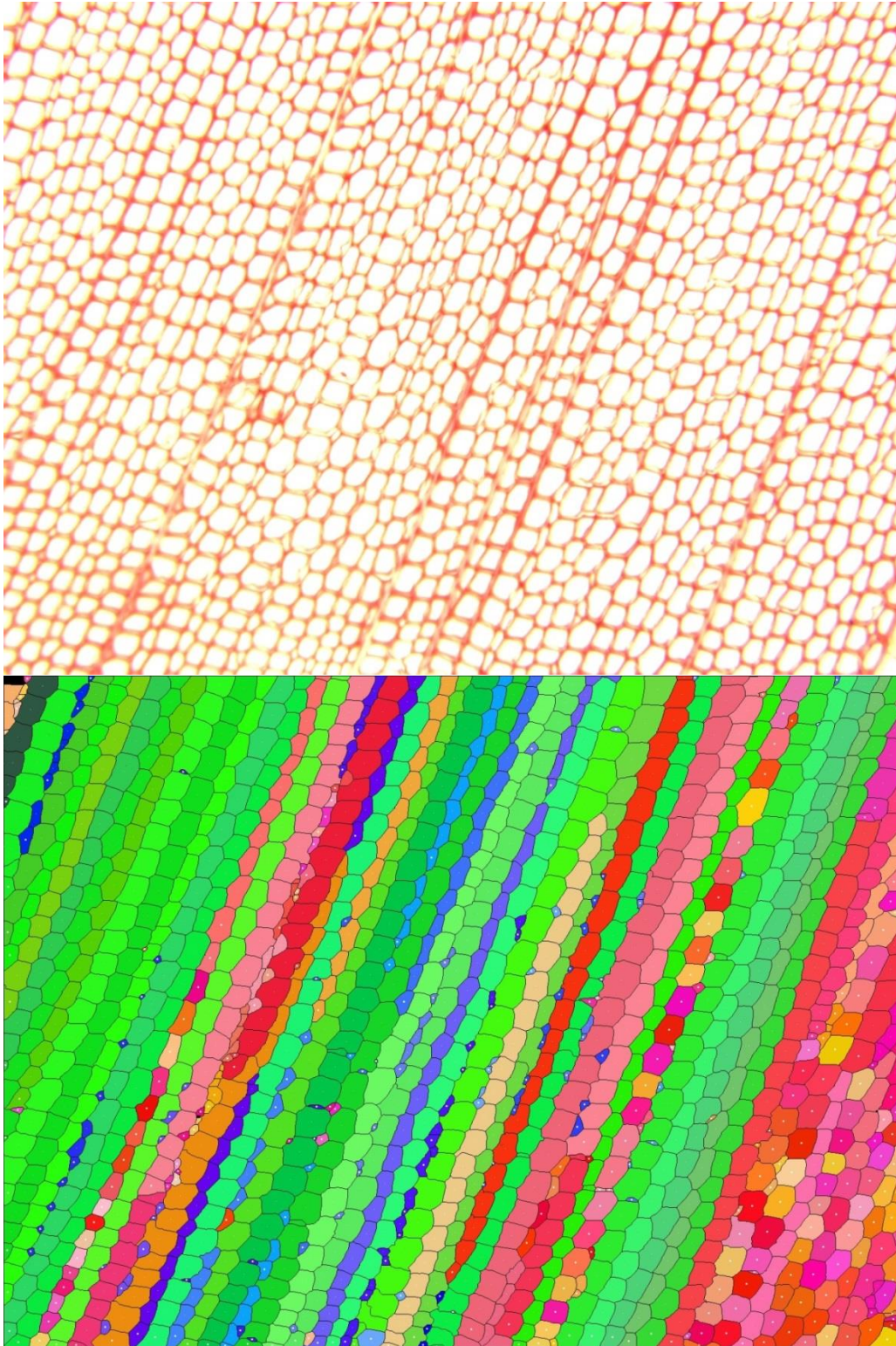
4. En haut, image native d'acajou (*Pycnanthus sp.*). En bas l'identification automatique des files de cellulaires.



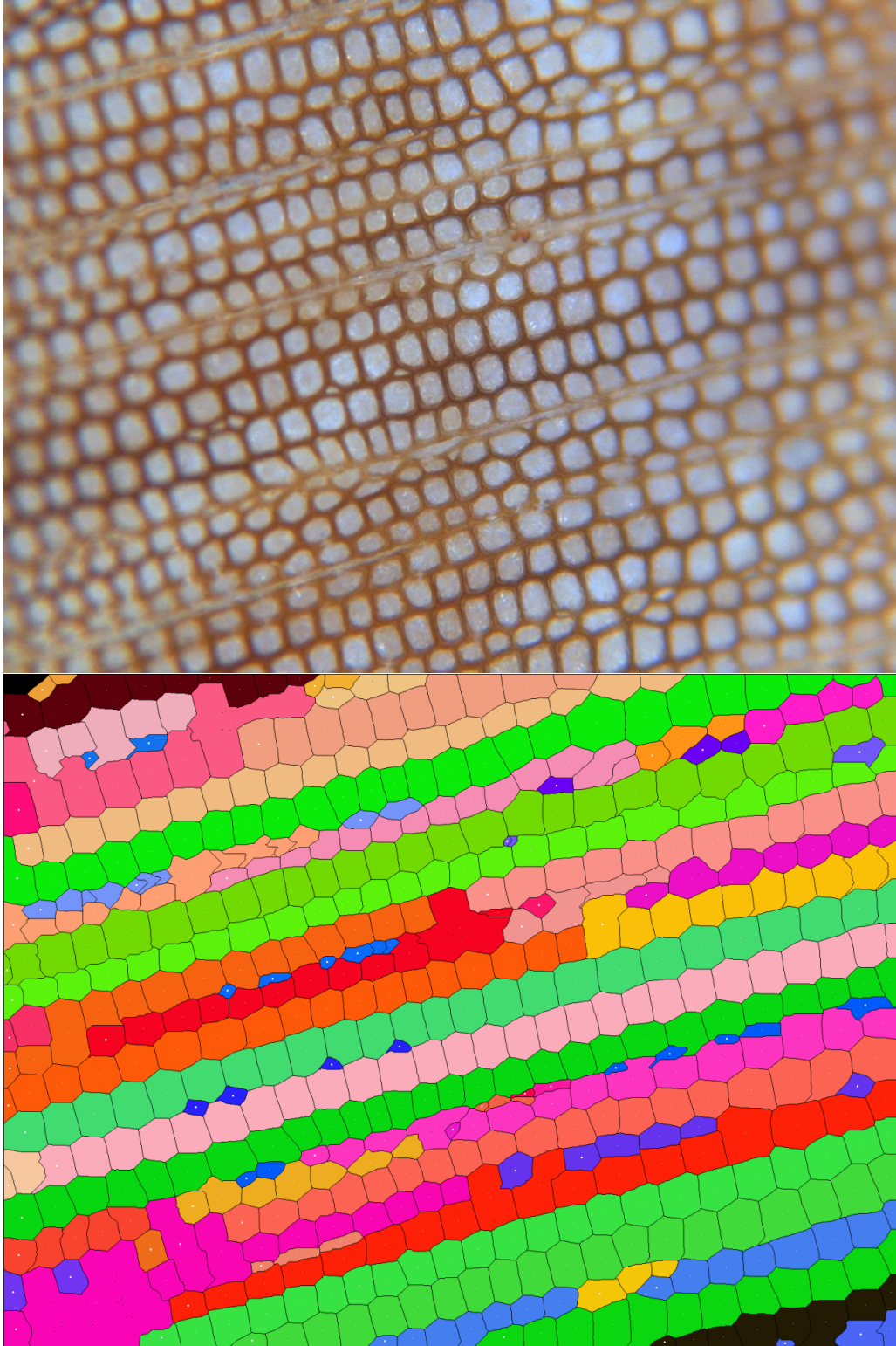
5. En haut, image native non colorée de sapin (*Abies alba*). En bas l'identification automatique des files de cellulaires.



6. En haut, image native de pin (*Pinus nigra*) avec un faible éclairage. En bas l'identification automatique des files de cellulaires.



7. En haut, image native de pin (*Pinus nigra*) avec un fort éclairage. En bas l'identification automatique des files de cellulaires.



8. En haut, image native de pin (*Pinus nigra*) issue du protocole de ponçage. En bas l'identification automatique des files de cellulaires.

6.2 Annexe 2 : Publications

Articles de références de la thèse (33 pages)

I. PMA'12. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio Y.** 2012.

II. AOB'14. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio Y.** 2014

III. FSPM'13. **Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio, Y.** 2013.

IV. SCIA'13. **Brunel G, Borianne P, Subsol G, Jaeger M.** 2013.

7 INDEX

- angiospermes** : la division des Magnoliophyta (ou Angiospermes) regroupe les plantes à fleurs, et donc les végétaux qui portent des fruits. Angiosperme signifie « graine dans un récipient » en grec par opposition aux gymnospermes (graine nue). ... 22, 23, 24, 32, 45, 52, 61, 73, 95, 97, 109
- bit** : unité de codage informatique. Une image en 8 bits sera représentée en niveau de gris et une image 24 bits sera représentée en couleur 28
- complexité algorithmique** : la théorie de la complexité est un domaine des mathématiques, et plus précisément de l'informatique théorique, qui étudie formellement la quantité de ressources (en temps et en espace) nécessaire pour la résolution de problèmes au moyen de l'exécution d'un algorithme 54
- couleur** : système de représentation de l'image. Le plus connu est le système RGB car c'est celui utilisé par la majorité des capteurs CCD. Dans ce système, la couleur est définie par la combinaison additive de trois couleurs primaires plus ou moins intenses. le Rouge, le Vert et le Bleu. Il existe plusieurs système de couleurs : le plus proche du système rétinien est le système HLS (Teinte, Luminance, Saturation)..... 28, 29, 33, 34, 36, 37, 39, 46, 49, 63, 78, 80, 83, 88, 92
- fibres** : faisceaux de cellules résistantes, disposées dans le sens axial (de la tige), assurant la rigidité et la résistance mécanique du bois 17, 25, 26, 52, 53, 67, 68, 69, 70
- graphe** : un graphe est un ensemble de points, dont certaines paires sont directement reliées par un (ou plusieurs) lien(s). Ces liens peuvent être orientés, c'est-à-dire qu'un lien entre deux points u et v relie soit u vers v, soit v vers u : dans ce cas, le graphe est dit orienté. Sinon, les liens sont symétriques, et le graphe est non orienté ... 18, 20, 42, 43, 44, 48, 49, 50, 52, 53, 63, 67, 96
- gymnospermes** : les Gymnospermes sont des plantes faisant partie d'un sous-embranchement paraphylétique des Spermaphytes (plantes à graines) qui inclut les plantes dont l'ovule est à nu (non enclos dans un ovaire à la différence des Angiospermes) et est porté par des pièces foliaires groupées sur un rameau fertile (cône). Le nom Gymnospermes provient du grec gymnospermis signifiant « semence nue » 18, 22, 23, 25, 32, 73, 95, 97, 109
- image native** : L'image native est l'image produite par un système d'acquisition n'ayant pas subi de traitement (filtrage) 35, 36, 80, 81, 110, 111, 112, 113, 114, 115, 116, 117
- intensité d'un signal** : puissance ou force du signal. l'intensité d'un niveau de gris est une valeur entière variant entre 0 et 255 qui représente la luminosité d'une teinte grise 33, 34, 36, 37, 38, 40, 52, 60, 62
- mosaïque d'image** : Une mosaïque d'image est la fusion de plusieurs images .. 4, 17, 19, 20, 26, 28, 49, 52, 53, 54, 63, 72, 95, 96, 97
- phloème secondaire** : Le phloème est le tissu conducteur de la sève élaborée qui est une solution riche en glucides tels que le saccharose, le sorbitol et le mannitol chez les plantes vasculaires. Le phloème a aussi un rôle de réserve avec les parenchymes et un rôle de soutien avec les fibres libériennes et les sclérites. Dans les troncs des arbres, le phloème peut faire partie de l'écorce, pour cette raison son nom provient d'un dérivé du mot grec phloos qui signifie écorce. 24
- pixels** : partie élémentaire de l'image numérique (abv de Picture Element). Il porte une information qui dépend de la nature et du codage de l'image. 28, 33, 34, 36, 40, 41, 49, 57, 58, 64, 65, 66, 72, 79, 80, 85
- prétraitement d'image** : opération visant à modifier les valeurs des pixels d'une image 35
- RGB** : le système RGB (Rouge Vert Bleu en français) est le plus simple des systèmes de codage informatique des couleurs. Il est composé de 3 canaux permettant d'afficher la couleur sur les écrans 36
- segmentation** : la segmentation d'image est une opération de traitement d'images qui a pour but de rassembler des pixels entre eux suivant des critères prédéfinis. Les pixels sont ainsi regroupés en régions, qui constituent un pavage ou une partition de l'image. Il peut s'agir par exemple de séparer les objets du fond. Si le nombre de classes est égal à deux, elle est appelée aussi binarisation. 18, 19, 20, 33, 34, 35, 37, 38, 39, 40, 46, 48, 62, 64, 70, 80, 81, 82, 83, 84, 86, 88, 90, 96, 98, 101, 102, 103, 104, 105
- stabilité algorithmique** : en analyse numérique, une branche des mathématiques, la stabilité algorithmique (ou numérique) est une propriété globale d'un algorithme numérique, une qualité nécessaire pour espérer obtenir des résultats ayant du sens 38
- stitching** : l'assemblage de photos est un procédé consistant à combiner plusieurs images numériques se recouvrant, dans le but de produire un panorama ou une image de haute définition 20, 49
- traitement d'image** : opération visant à extraire une information d'une image . 4, 11, 17, 19, 33, 35, 36, 37, 40, 46, 47, 49, 60, 72, 75, 79, 80, 81, 90, 91, 92, 93, 96, 97, 98, 102
- xylème secondaire** : le xylème (du grec xylon, « bois ») ou tissu xylémique, est un constituant des tissus végétaux formé de l'association de vaisseaux, de cellules mortes ou vivantes de soutien et de cellules associées. 24, 25

Automatic Characterization of the Cell Organization in Light Microscopic Images of Wood: Application to the Identification of the Cell File

Guilhem Brunel^{1,2}, Philippe Borianne², Gérard Subsol³, Marc Jaeger² & Yves Caraglio²

¹UMII; ²CIRAD – UMR AMAP
Montpellier, France
Guilhem.brunel@cirad.fr

³CNRS – LIRMM
Montpellier, France

Abstract — Automated analysis of wood anatomical sections is of great interest in understanding the growth and development of plants. In this paper, we propose a novel method to characterize the cell organization in light microscopic wood section images. It aims to identify automatically the cell file in a context of mass treatment. The originality of the proposed method is our cell classification process. Unlike many supervised methods, our method is self conditioned, based on a decision tree which thresholds are automatically evaluated according to specific biological characteristics of each image. In order to evaluate the performances of the proposed system and allow the certification of the cell line detection, we introduced indices of quality characterizing the accuracy of results and parameters of these results. Those are related to topological and geometrical characters of the cell file at both global and local scales. Moreover, we propose an index of certainty for selective results exploitation in further statistical studies. The proposed method was implemented as a plugin for ImageJ. Tests hold on various wood section well contrasted images show good results in terms of cell file detection and process speed.

Keywords: *Image processing, pattern recognition, wood microscopic images, cell segmentation, file identification.*

I. INTRODUCTION

The development of the tree results from the primary growth and secondary growth. The primary growth concerns the extension and branching of the axes. Whereas the secondary growth is on their thickening. Among others, issues of carbon sequestration and wood energy leads to consider both structural and functional aspects using accumulating approaches. However, these approaches, concerning secondary growth and its relationships between secondary and primary growth, are based on fragmentary studies, mainly raising from data acquisition costs. Secondary growth results from the cambial activity, which is among other things to the origin of wood (xylem) and consequently the different cellular elements that constitute it. The understanding of mechanisms of growth of the cambium will go through the study of rhythmicity of cell patterns, their disruption or modification through space and time. Environment and its fluctuations influence the differentiation of wood's elements (from the divisions of cambial cells) which constrain us to follow the production of specific cell organizations. To simplify, two organizations are considered: firstly the growth rings which represents cell's production at a given time [1], and secondly

the cell lines which represents the activity of an initial cell during time [2]. For example, [3] underlie the importance of secondary growth organizations by studying cellular patterns, their spatial rhythmicity, their variability along different growth rings. For its part, the study of cell lines is presented as a promising enable to understand the development, differentiation and temporal rhythmicity of cells [4]. Automating the study of cell lines should permit to closer links with the functional aspects and ecology of species: Wood is a continuous recording of changes in the development of the tree (the lesser known) and its environment (the most work, eg dendrochronology, tree-rings).

Progress in terms of the image acquisition devices and their analysis allow us to consider the access to quality's information on various extensive areas in the plant. Cross section microscopic images analysis shows interest in both cell feature characterization and detection but also for higher structure character estimation as shown for instance by Wu et al on root sections [5].

Automatic identification of wood cellular structures as cell files is a new challenge in structural biology of plants [6], which requires a multidisciplinary expertise.

In bio-imaging, cell segmentation has been intensively studied by researchers in image processing with various approaches [7][8][9]. For example, [10] combine four classical algorithms to segment white blood cells in bone marrow image: watershed [11], snake [12], multi-resolution analysis [13] and dynamic programming. This kind of approach based on mixing many algorithms, shows the difficulty of segmenting cell images with high biological variability. The algorithms should be adapted to the application and to the specific cells characters. In particular, for wood cross section, we have to consider both cell local geometry and arrangement.

In such wood cellular organizations, some authors propose to introduce some geometric models to describe the topological neighborhood of cells. For example, [14] or [15] rely on an oriented graph to extract cell files in images of the gymnosperms.

From a technical point of view, we did not found any software solution allowing automatic identification of cell files. Some specific commercial tools as WinCell [23], exist for the analysis of wood cells, but they do not allow recognition and characterization of the cellular organization. Moreover, it is not possible to add new functionalities. More

generic commercial software such as Visilog [24], offer a rich development environment in image processing but they are quite expensive and not sufficiently specialized for our application. Open source platforms as ImageJ [20][25] are an attractive option: they allow to enrich the source code with specific functionalities while facilitating the exchange and dissemination of methods. Therefore we chose this solution to implement our algorithm.

The proposed method of automatic identification and modeling of cell files extends the research presented by [15]. In this work, cell file identification in light microscopy images is based on a preliminary identification of the cells. It is a supervised classification requiring the definition, then the use of a large training database. As the variability of wood anatomy is huge, significant results can be obtained only for a limited number of wood species, sharing close and simple anatomical features (gymnosperms in this case). We introduce here a specific model of the cell lines recognition avoiding supervised training and allowing results quality quantification. Applying this model, as shown in the following sections, we can build a generic tool suitable for both gymnosperm and angiosperm wood cross section images.

II. MATERIAL AND METHOD

A. Preparation and digitization

We aim to process histological stained sections of angiosperm and gymnosperm wood, and more specifically of Mahogany, *Fraxinus Angustifolia*, Pine Black, Pine Carabinsis and Fir tree.

Wood cross sections of 20-25 μm of thickness are produced with a vibratome. They are then immersed in an aqueous chlorinated solution to clean the cellular content. A colouring of the cell wall is then performed to increase the contrast with the lumen. Various types of colouring agents can be used as the methylene blue, the toluidine blue or the safranine. The sections are then digitized by a Olympus DP71 LCD camera mounted on a Olympus BX51 microscope. The magnification is chosen around $\times 100$.

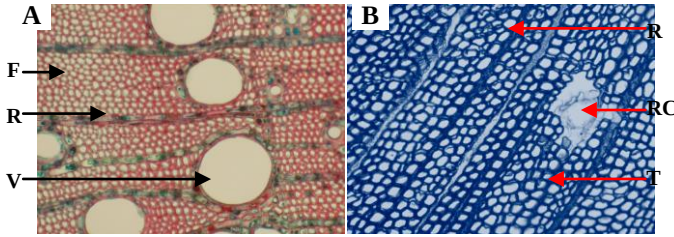


Figure 1. Images of histological stained wood sections. A: *Fraxinus* (angiosperm) colored with safranin. B: *Pinus Caribensis* (gymnosperm) colored with methylene blue. Notice the different anatomical structures: Vessel (V), Ray (R), Fiber (F), Tracheid (T) and Resin Canal (RC).

Different anatomical structures as vessels, fibers, tracheids, resin canals, or rays are visible around the cell files (see Fig. 1). With such a magnification value, we can count at least 20 consecutive cells and then distinguish the cell files and the smallest interesting structure covers a surface of at least 10 by 10 pixels.

The resulting images are in color (coded on 24 bits) with a resolution close to 2000 by 2000 pixels.

B. Cell file processing

Identification of cell files is based on the search of alignment of cells which share similar geometric (i.e. size, shape) and densitometric (i.e. color) properties. The concept of alignment implies a specific cellular organization which is emphasized by the neighborhood relations between cells.

Our approach is divided into three steps (see Fig. 2): first cell identification is performed in order to individualize the cells in the image, then the cellular organization step detects and individualizes the alignments of anatomical structures, and finally, the classification step classifies anatomically and types qualitatively the cell files. In the following sections, we detail these three steps.

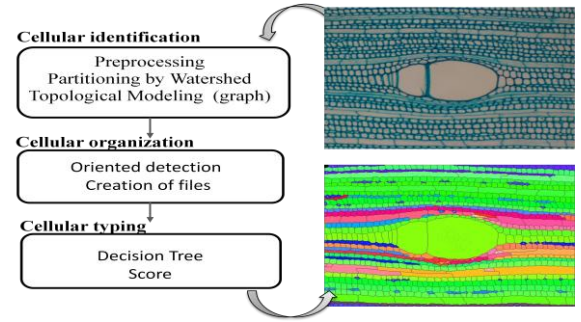


Figure 2. Overview of the cell file identification algorithm.

1) Cell identification step

Microscope images present a "pepper-and-salt" noise inferred by the thermal effect of the lamp. We attenuate this impulsive noise by applying a median filter of radius 3 pixels. Images show alternations of clear and dark areas representing respectively the cell lumens and the cell walls. To increase the contrast between the two areas, we apply the Difference of Gaussian (DoG) [16] method. This is a bandpass filter. It thresholds the frequencies corresponding to the lumens. To obtain the lumens (see Fig. 3), the DoG filter subtracts a Gaussian light blurred image (blurring parameter close to the size of the cell wall) to the highly blurred image (blurring parameter set to 1/10 of the image width). Nota that the blurring parameters were set experimentally and are different from the classical 1.6 ratio used for DoG smoothing in edge detection.

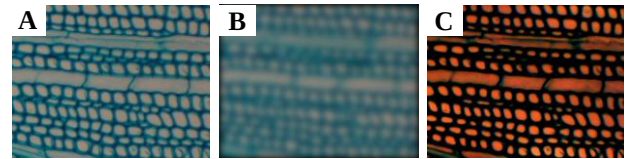


Figure 3. Difference of Gaussian filtering on a cross-section of Mahogany stained with toluidine blue. A: the image obtained by applying a small Gaussian blurring ($\sigma = 3$ pixels). B: the image resulting from a strong Gaussian blurring ($\sigma = (\text{image width})/10$). C: subtraction of images A and B which increases the contrast wall / lumen.

We obtain an image where the walls have zero intensity and the lumens have a higher intensity.

This colour image is then converted in grey levels. Average colour channel value is used since no specific channel seems significant. Moreover this conversion simplifies specific process definition potentially resulting from different colouring protocols.

The greyscale image cell extraction is obtained by the classical Watershed algorithm [11]. The idea is to consider a grey level image as a topographic relief, and to calculate the watershed lines by “flooding” the relief. The resulting watershed lines define a partition of the image. The ridge lines constitute the intercellular boundaries and correspond to the middle lamella. The lines of the ridges of the Watershed (fig. 4-c) give the boundary between two adjacent watersheds. The dual graph is the adjacency graph of the basins, it connects two by two the geometric centers of the basins incident to the same edge. The basins are superimposable to image cells, we speak of the adjacency graph of the cells (Fig. 4-d).

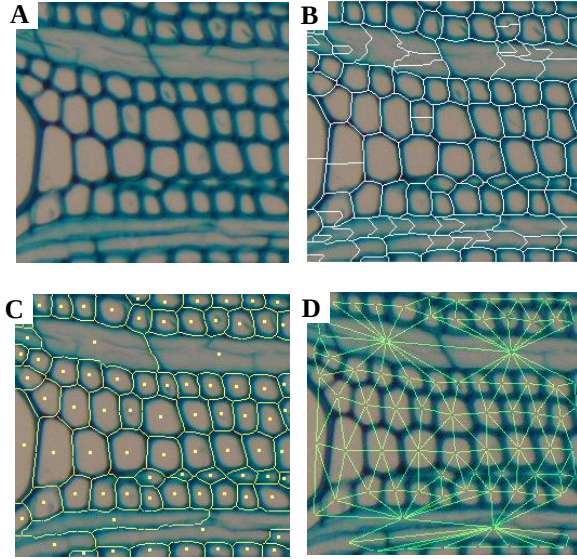


Figure 4. Cell individualization of a Mahogany cross section. A: native image. B: watershed segmentation result with the crest lines crossing the lumen. C: cleaned watershed: the crest lines are replaced by curvilinear edges shown in yellow; they correspond to the middle lamella; the points extra yellow match the geometric center of the watershed therefore biological cells. D: dual watershed graph. This adjacency graph, in green connects cell centers with their neighbour; notice that the degree of vertices is greater than four showing the staggered pattern of the cells.

2) Cellular Organization

Unlike approaches [15] and [14], our recognition of lines based exclusively on a constraint path of the adjacency graph of the cells. More specifically, the construction of cell lines is based on the one hand on geometric criteria, in particular the general orientation of adjacencies cell, and, secondly on topological criteria, in particular on configurations of cellular adjacency.

The edges of the adjacency graph represent connections between neighboring cells and their orientations give an indication of the preferred directions of the cell arrangements. On our studies, cells are arranged in staggered rows, and cell lines are rectilinear and two in two parallels. Therefore, the orientations of edges adjacency follow three directions: a main one, corresponding to the cell alignments and two secondaries, corresponding to the organization in staggered rows (fig. 4-d). To compute the main direction, we use R. Jones (1996) method, studying the distribution of angles formed by each edge of the adjacency graph with the horizontal. The distribution is then trimodal : the major mode shows the orientation most present in the adjacency graph, that is to say the general orientation of the lines (Fig. 5).

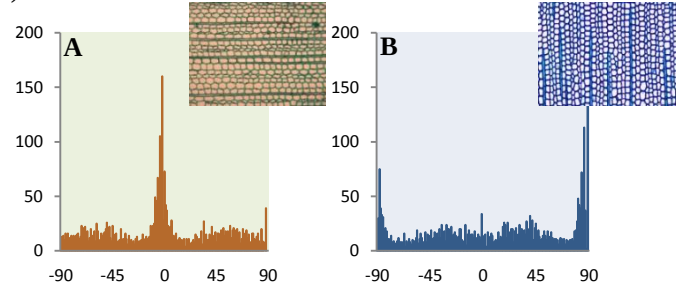


Figure 5. Direction distribution histograms formed by edges of the adjacency graph with the horizontal. A: The main mode centered on 0° corresponds to horizontal lines; B: The histogram shows one mode splitted on -90 and 90° values corresponding to vertical lines.

Considering now the lines. The construction is progressive and based on four principles: alignment search, feedback check, overlap management and chaining. The alignment principle allows for each vertex V of the graph to find the next vertex NV in the cell lines. Each vertex neighbor of V is a potential candidate. We will retain one that maintains the best spatial alignment (in a range of $\pm 35^\circ$), and whose underlying cell shows the highest geometrical similarities with the previous one. The line is gradually built from the initial edge, by successive additions of consecutive edges preserving the geometric continuity. At each step, the drift of the line is minimized under this geometrical constraint. We chose the criterion of Bray-Curtis [21] (see Fig. 6) for similarity cost evaluation since well adapted to surface characterizations. When this criterion tends to zero, the compared cells can be considered as similar.

$$BC_{nm} = \frac{|n - m|}{(n + m)}$$

Figure 6. Formula of Bray Curtis. Where « n » is the surface of current cell and « m » is the surface of neighbor cell. When the criterion tends to 0, the cells show a similar surface.

Each cell bellowing to a cell line path is given a score value : the sum of the angle deviation with its predecessor and the criterion of Bray-Curtis.

The path of the line stops when there are not vertices anymore candidate.

A cell line direction is independent from graph traversal orientation. This property is used to validate or invalidate any line segment or part of it. The process described above is applied again starting from the end edge, by inverting the direction of progress in the graph. We note this path as "backward", opposed to the initial path "forward". This define the principle of Forward / Backward : a line is validated when both forward and backward paths are strictly superposables. In the case where the paths are not identical, the initial forward path is gradually reduced until reaching a stable forward/backward sub-path. Specifically, whenever a difference appears in backward path, the corresponding segment is deleted from the original line: the edges and their scores are released, initialized to an infinite value. The backward tip is updated from the last reached position. In the strictly superposable case, the score of each cell of the line is decreased by half insuring high stability of the line towards overlapping (see below).

Indeed, a cell belongs to a unique file. This property leads to the following rule: each vertex belongs to one cell line and only one. In some cases, during its construction, a line may use one or several cells, mobilized by a line already constructed. In this case, each of these vertex is attributed to the line for which its score is minimal: it is the principle of overlapping. Two scenarii may occur:

- The existing line keeps its vertex; the line under construction will have to find another path or stop.
- The existing line loses its summit to the detriment of the line under construction. Here, all the summits of the existing line situated after the lost summit are released: their respective score takes an infinite value, indicating they are again available.

Due to the presence of intrusions, tears or a lack of adjacency (Fig. 7-a), it is not unusual to detect a given line in splitted part.

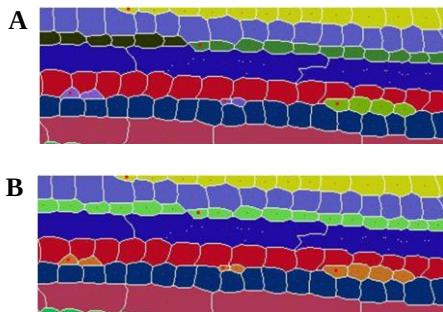


Figure 7. A: Result of treatment chain without connection. B: result whit connection (Bottom). There is one color by file.

The principle of merging allows to concatenate several part of line, using simple topological rules, to establish whole lines. The idea is based on two observations : (i) the lines cross the image throughout, (ii) the lines do not

intersect (Fig. 8-a). As a result, if the parts T2 and T2' are adjacent to the line F1 and F3, it is likely that T2 and T2' are two parts of the same line F2. At present only the parts in two same lines are concatenate that is to say as belonging to the same line. In other scenarii the process is stopped.

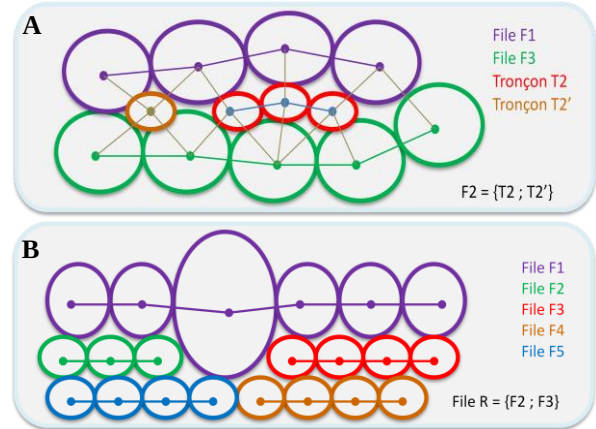


Figure 8. A: a two neighbours connection scheme. Sections T2 and T2' are squeezed between lines F1 and F3; therefore T2 and T2' are probably derived from a common file F2. B: a multiple neighbours scheme; in this case, rules of adjacency are used to drop the case in a two neighbours one.

3) Classification step

The cell typing is the ultimate step: it allows to classify different cells (fibers, tracheids, vessels, rays, ...). It is realized from geometric and densitometric characterization of the watersheds associated with the vertices of the adjacency graph of cells. Unlike the work of [15] and [14], supervised classification was discarded because of the difficulty of building learning games sufficiently complete and discriminating, due to the existing biological variability.

A decision tree discriminates the anatomical structures. It was established with wood anatomists from the two following hypotheses:

- 1) The perimeter of cells, noted parameter T, allows to differentiate the "big" structures from the "small" cells.
- 2) The circularity of cells, noted parameter C, allows to differentiate rays (globally lengthened) from the vessels (globally circular).

The decision tree given in Fig. 9 is applied to basins with a clear area corresponding to the lumen. It requests two threshold values which are automatically estimated for every image. For each parameter, both values are calculated from the grey image, from two disjoint groups using a two-means clustering [22]. This classical method divides n observations into two clusters, so minimizes the intra-classe variance and maximizes the interclasse variance. The threshold is then given by the median value between the upper bound of the lowest group and the lower bound of the stronger group.

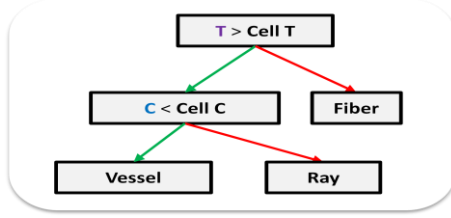


Figure 9. Decision Tree. T represents the threshold of the perimeter and C the threshold of circularity. The thresholds are automatically rated by a 2-means clustering applied to all values from the image. Only cells with a lumen are processed by the decision tree.

After the alignments identification and the biological typing, we can deliver a set of figures characterizing the shape, size, nature of biological structures (walls, lights, cells, lines ...). These characteristics, of interest for anatomists, are defined by self conditioned procedures. As part of a mass treatment it is interesting to characterize the accuracy of these assessments. Both characteristic quality and automated define parameters are computed defining indexes of certainty. An index is assigned for each calculated parameter. The certainty index concerns the geometric parameters of structural elements and the steps of the process of identifying lines.

At the scale of structural elements (fiber, vessels, radius ...), evaluation of this index depends on the parameter itself. For example, the certainty of the surface or the shape of the lumen is directly connected to the local degree of sharpness of the image (Fig. 14): the boundary between the wall and the lumen being obtained by distribution of the pixels of the cell two intensity classes (light intensities and dark intensities), it depends irremediably on local dynamics of the image.

At scale structure elements (lines), evaluation of the index depends on the construction process and the nature of the result. To "certify" the cell lines, three criteria are used :

- The overall score of the line, characterizing topological relationships of the line with its neighbors. The identification of the topological position allows to assign a global additional score to the line: 0 for a line crossing the image throughout, 1 for a line composed of simple parts, 2 for non identified segment, 3 for single cell, 4 for defects.
- The average of the scores of its cells obtained during the construction of the line (not to be confused with the index of certainty parameters). Each vertex added to the line will be assigned a score defined by the sum of the deviation of the angular aperture and the similarity coefficient. This method allows to restrict the path of the line without guaranteeing possible overlapping or intersection with a nearby line.
- The length, characterizing the representativeness in the image: the more a line is long, the more it is sure.

An index of certainty is assigned to the line according a linear combinaison of these three scores. The indices of

certainties (parameters and lines) are used to filter or classify the numerical results generated for each lines identified.

The indices of certainty of the lines are totally independent from indices certainties of the parameters associated with the cells.

The visual representation of the global score of the lines, and in particular the use of a color code, allows a fast and efficient processing check, distinguishing the whole lines form the cellular insertions. The colour code is green for the lines automatically identified, is blue for lines automatically rebuilt and is red for complex configurations (Fig. 10).

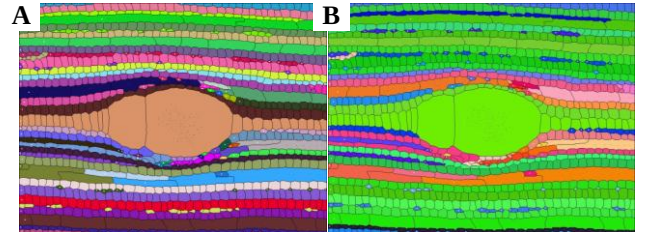


Figure 10. Automatic identification of cell lines from a cross section of mahogany. A: lines random colouring is used. B: colour qualifies the cell files; the lines automatically identified in green, the lines automatically rebuilt in blue and the complex configurations in red.

The method has been implemented in Java and integrated as a plugin in the ImageJ free platform. We use additional free libraries such as the Java Universal Network / Graph Framework [26] for efficient data structure management. We are currently posting the application on the web [27].

III. RESULTS AND DISCUSSION

The tests were led from a sample of images representative of biological variability, possessing a dozen colored sections of different species of angiosperms (list) and gymnosperms (list) (Fig. 11).

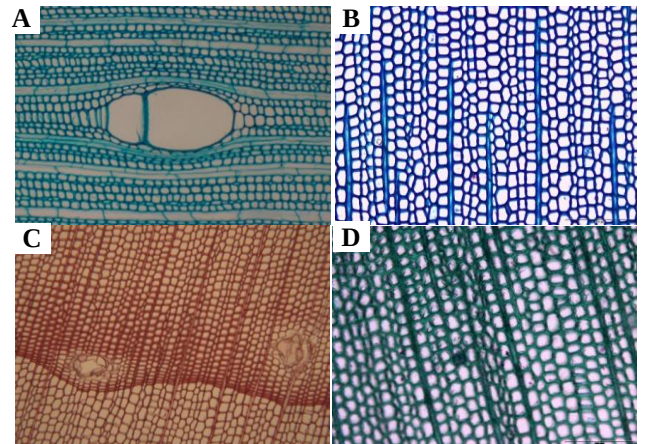


Figure 11. Extract from the test set consists of cross sections of angiosperms and gymnosperms. A : mahogany colored with toluidine blue for which the continuity of queues is preserved in spite of the vessel elements. B : fir colored with methylene blue. C : black pine colored with safranin having resin canals. D : fir colored with toluidine blue showing cells blocked.

A. Biological aspects

Moreover, the colour encoding can quickly locate the complex biological configurations requiring the expertise of an anatomist to be identified (in red, Fig. 12).

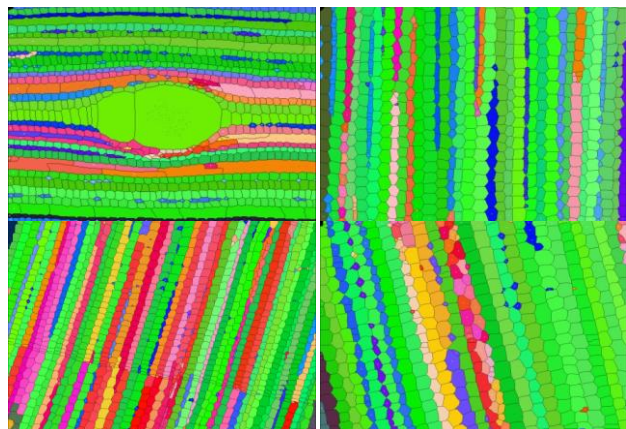


Figure 12. Automatic identification of cell lines of the extract on the game of trying to figure 11. The colour code is given in figure 10.

As shown in the table below, our method works well on images presenting a structured organization with a marked wall / lumen contrast, in regard to data acquisition, management of lines and the magnification used. For configurations showing low colour contrast or complex cellular organizations the lines are badly detected.

Table 1. Summary of some significant results: the size of images processed, the number of cells extracted, the CPU time obtained on a machine with an Intel Xeon at 2.3GHz and a Total Quality Index defined as the ratio of the number of lines automatically identified and reconstructed on the total number of lines.

Species	Size (pixels)	Cells numbers	Times (sec)	Total Quality (%)
Mahogany	1024x768	1359	14.3	83
Fir	1360x1024	800	12.4	92
Black Pine 1	1600x1200	1873	23.2	73
Caribbean Pine	1360x1024	828	11.5	91
Black Pine 2	1600x1200	1458	16.1	93

What about in terms of quantity, that is to say in terms of numerical measures?

The overall rating of the line remains the most important indicator. It enables us in particular to detect all single files: all lines described as obvious by anatomists have an overall zero score.

On the test set, the lines are automatically detected at 60% in the images. This figure is not significant since it depends on the anatomical configurations. The unidentified lines can be automatically excluded from statistical processing. The amount lost can be easily offset by the increase of the images treatable.

It is important to evaluate the accuracy of measurement. The Figure 13 presents the comparison of about sixty normalized areas got from manual and automated method.

The coefficient of determination tends of 1 showing the areas are well correlated. The slope of regression line is weakly superior to 1 indicating a weak over-valuation of the automated method. The weak gap of 0.0127 confirms the middle error of 5% on right areas. The automated method seems over-valuation the measurements (at least the expert under-valuation the measures!). Only certainty: the automated method is repeatable.

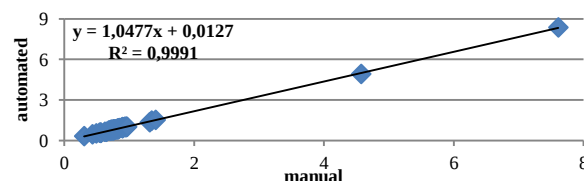


Figure 13. Basins surfaces study on Mahogany. In abscissa, the normalized areas got from the fully manual method. In ordinate, those got from the automated method. The coefficient of determination is close to 1, showing a very well adjustment of areas.

The major limit to the automatic identification of lines comes from the image content, that is to say the photometric characteristics of the image, and biological configurations.

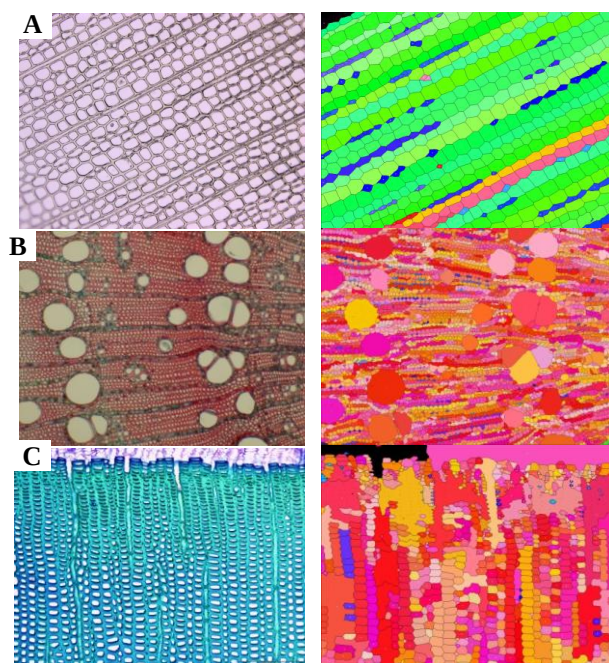


Figure 14. A right cross sections of three native species. On the left, the automatic identification of cell lines. A : not colored pine treated in transmission; Detection is good despite the low constrate Wall / Lumen. B : ash colored with safranin; Detection of cells is good, but the method only produces sections due to the complexity of biological organization. C : pine colored with methylene blue; Identification is biased by the presence of local inversion of contrast at the junction wood summer / winter.

The images with a dynamic intensity locally reversed, for example the junction wood winter / summer wood (Fig. 14-c), remain difficult to treat. Our method is based on the image contrast, so watersheds corresponding to the cells are badly detected, and thus the detection of files is incorrect

and incomplete. The images of washer sanded wood (unstained) are treated in reflection: light does not pass through the sample. Thus the lumens appear in colour almost close to the walls one, with not sufficient contrast to ensure proper recognition of watersheds and proper identification of lines. The wood sections not colored are treated by transmission, crossed by light. They are less contrasted than the stained sections. But the difference wall / lumen is pronounced enough to allow the algorithm to correctly identify the lines (Fig. 14-a). For complex cellular organizations (Fig. 14-b), where cell alignments are not obvious, new rules for the path and for the reconstruction should be established with the biologists.

The files are created using only geometric and topological rules on the basins. It is important to make sure of the robustness of the detection of the basins, especially in terms of insensitivity to blur and image orientation. A first study was conducted on the absorption of the blur found in images related to microscope views side effect or sample flatness.

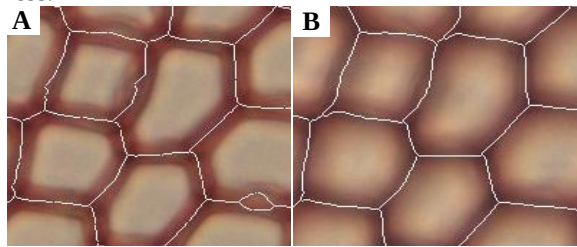


Figure 15. Compared watershed result on Pine cross section. A : crests lines of a sharp image. B : crests lines of a blurred image. Outline of the basins is nearly identical between a sharp and a blur image. On the sharp image, a small additional basin of a detachment of the wall issued from cutting. This basin is skipped during the supernumerary removal phase.

Acquisitions of the same zone of the image with different focus were performed to evaluate the stability of our computational method. In particular, we made a statistical study on the variability of cropping cells. The intercellular lines obtained by the algorithm of Watershed are generally invariant to blurring of the image. Indeed, they correspond to the curves reverse slopes of intensity: the smoothing of the image produced by optical blur attenuates the intensity without changing the look of intensity variations. The crest lines remain unchanged (Fig. 15).

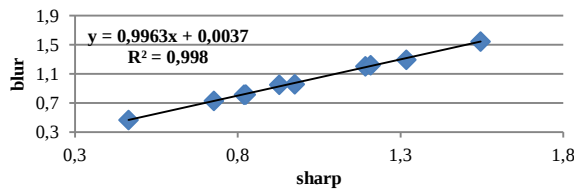


Figure 16. Basins surfaces study on Mahogany. In abscissa, the normalized areas got from sharp images. In ordinate, those got from blur images. The coefficient of determination is close to 1, showing a very well adjustment of areas.

Figure 16 shows the comparison of a sixty normalized area obtained from a sharp and a blurred image. The coefficient

of determination tends to 1 showing areas are well correlated. The slope of the regression line is slightly less than 1 indicating low undervaluation on blurred images. The difference of 0.0037 confirms the average error of 2.81% on sharp image. The method gives results almost identical regardless of the sharpness of the image

The method described is working correctly for images that contain the one hand high contrast between the walls and lumens (without local inversion of color) and other hand a visible cellular organization. Under these conditions all the obvious lines by qualified experts are correctly identified, with a significant time save (a typical manual expert identification requires ten hours on our samples). Moreover, the index of certainty allows the selective exploitation of results for statistical studies.

IV. CONCLUSIONS

Automated analysis of wood anatomical sections is of great interest in understanding the growth and development of plants. In a frame of over studies on wood structure, we propose a original method for automatic detection of cell lines, operational in mass treatment. It applies to digital images of coloured transverse cut wood. Identification of lines based on research alignment of cells presenting similar geometrical and densitometric characteristics. The notion of alignment implies to know how to set up the relations of neighborhood between cells. Identification of lines is made from an adjacency graph for the path constrained graph with an orientation and similarity criteria. The originality of the method compared to a supervised method is that it automatically creates rules evaluated for each image and not a set of rules to be applied on all images. The lines described as obvious by anatomists are correctly identified by our method, with a drastic drop of time cost. To "certify" the cell lines, we have introduced the indices characterizing the quality of results and parameters of these results. The method described works correctly for images with high contrast between the walls and lumens and a clear cellular organization. Under these conditions all the obvious lines by qualified experts are correctly identified. Moreover, the index of certainty allows the selective exploitation of results for statistical studies.

Three future work axes are being considered:

1. The enlargement of the major zones of study in order to follow the lines of several rings (Fig. 17). For this, it is necessary to adapt the processing of image mosaics. Two problems are directly related: first, the treatment of different photometric behavior in the same image, in particular the transition wood winter / summer wood. Secondly, the joining of the results at the edges of the images pavement.
2. The study can be extended to sanded wood, for which the contrast wall / lumen is less contrasted.
3. The enrichment of the cell typing method that is not currently able to differentiate all kind of vessels. Work

on the walls texture can be considered to resolve the limitation.

Cell file identification has to deal with other open difficulties like the passage tree ring or the junction between the different images. The construction of cell lines will be reviewed to match between images that may or may not be recovered.

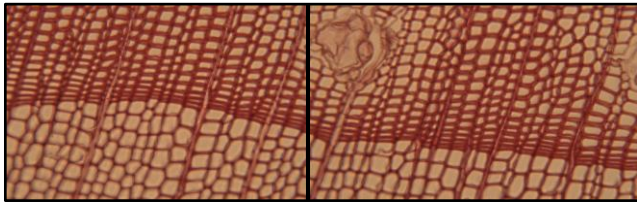


Figure 17. Abutting mosaic of a cross section of Black Pine, colored with safranin, 100x.

As a summary, this work is a first contribution to develop methods for automatic image processing that will mainly to identify and characterize, in mosaics depicting field's broad observation, cellular organizations, and the cells that compose them. The implementation aims to obtain a rapid cell typing, automatic and reliable to process statistically significantly large sets of data.

ACKNOWLEDGMENT

The authors gratefully acknowledge Christine HEINZ for her involvement in this work, her advices and availability. A big thank to Michael GUEROULT without whom the laboratory work would not have been possible.

This work is jointly funded by a doctoral fellowship of the Labex NUMEV and the SIBAGHE Graduate School of the University Montpellier 2 and by the Scientific Council of the University Montpellier 2.

REFERENCES

- [1] L. Heinrich, "Reaction wood varieties caused by different experimental treatments," *TRACE, Tree Rings in Archaeology, Climatology and Ecology*, vol. 5, p. 224–232, 2007.
- [2] P. Rozenberg, G. Schüte, M. Ivkovich, C. Bastien, and J.-C. Bastien, "Clonal variation of indirect cambium reaction to within-growing season temperature changes in Douglas-fir," *Forestry*, vol. 77, no. 4, pp. 257–268, Jan. 2004..
- [3] E. Nicolini, Y. Caraglio, R. Pélissier, C. Leroy, and J. Roggy, "Epicormic Branches: a Growth Indicator for the Tropical Forest Tree, *Dicorynia guianensis* Amshoff (Caesalpinaceae)," *Annals of Botany*, vol. 92, no. 1, pp. 97–105, Jul. 2003..
- [4] W. Gindl, "Cell-wall lignin content related to tracheid dimensions in drought-sensitive austrian pine (*pinus nigra*)," *Iawa Journal*, vol. 22, no. 2, pp. 113–120, 2001..
- [5] Haiwen Wu, Marc Jaeger, Mao Wang, Baoguo Li and Baogui Zhang. 3D-Reconstruction and Visualization of Xylem Vessels of Wheat Nodal Root. In: Li, B. and Jaeger, M. and Guo, Y. (Eds). 2010. Proceedings of Plant growth Modeling, and their Applications (PMA09), Beijing, China, November 9-13, 2009, IEEE CPS, pp. 384-390
- [6] P. Quelhas, J. Nieuwland, W. Dewitte, A. M. Mendonça, J. Murray, and A. Campilho, "Arabidopsis Thaliana Automatic Cell File Detection and Cell Length Estimation," in *Image Analysis and Recognition*, vol. 6754, M. Kamel and A. Campilho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–11.
- [7] M. K. Moëll and L. A. Donaldson, "Comparison of segmentation method for digital image analysis of confocal microscope images to measure tracheid cell dimensions," *IAWA Journal*. Vol. 22(3), pp. 267–288, 2001.
- [8] D. Baggett, M. Nakaya, M. McAuliffe, T. P. Yamaguchi, and S. Lockett, "Whole cell segmentation in solid tissue sections," *Cytometry A*, vol. 67, no. 2, pp. 137–143, Oct. 2005.
- [9] T. Fourcaud, X. Zhang, A. Stokes, H. Lambers, and C. Körner, "Plant Growth Modelling and Applications: The Increasing Importance of Plant Architecture in Growth Models," *Ann Bot (Lond)*, vol. 101, no. 8, pp. 1053–1063, May 2008.
- [10] J. Park and J. M. Keller, "Snakes on the Watershed," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1201–1205, 2001.
- [11] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, Jun. 1991.
- [12] Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int J Comput Vis*, vol. 1, no. 4, pp. 321–331, 1988.
- [13] M. B. Jeacocke and B. C. Lovell, "A multi-resolution algorithm for cytological image segmentation," in *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, 1994, 1994, pp. 322–326.
- [14] R. Jones and L. Bischof, "A graph-based segmentation of wood micrographs," in *Computing Science and Statistics*, 28, Sydney, 1996, pp. 12–20.
- [15] P. Kennel, G. Subsol, M. Guérout, and P. Borianne, "Automatic identification of cell files in light microscopic images of conifer wood," in *2010 2nd International Conference on Image Processing Theory Tools and Applications (IPTA)*, 2010, pp. 98–103.
- [16] G. T. Einevoll and H. E. Plesser, "Response of the difference-of-Gaussians model to circular drifting-grating patches," *Vis. Neurosci.*, vol. 22, no. 4, pp. 437–446, Aug. 2005.
- [17] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, 1st ed. Chapman and Hall/CRC, 1984.
- [18] O. M. Hitz, H. Gärtner, I. Heinrich, and M. Monbaron, "Wood anatomical changes in roots of European ash (*Fraxinus excelsior* L.) after exposure," *Dendrochronologia*, vol. 25, no. 3, pp. 145–152, Mar. 2008.
- [19] A. J. Travis, D. J. Hirst, and A. Chesson, "Automatic Classification of Plant Cells According to Tissue Type using Anatomical Features Obtained by the Distance Transform," *Annals of Botany*, vol. 78, no. 3, pp. 325–331, 1996.
- [20] B. Clair, J. Gril, F. Di Renzo, H. Yamamoto, and F. Quignard, "Characterization of a Gel in the Cell Wall To Elucidate the Paradoxical Shrinkage of Tension Wood," *Biomacromolecules*, vol. 9, no. 2, pp. 494–498, 2007.
- [21] J. R. Bray and J. T. Curtis, "An Ordination of the Upland Forest Communities of Southern Wisconsin," *Ecological Monographs*, vol. 27, p. 325, Oct. 1957.
- [22] E. Forgy, "Cluster analysis of multivariate data: efficiency vs interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.
- [23] <http://www.regentstruments.com/>
- [24] <http://www.noesisvision.com/fr/index.html>
- [25] <http://rsbweb.nih.gov/ij/index.html>
- [26] <http://jung.sourceforge.net/doc/index.html>
- [27] http://umramap.cirad.fr/amap2/logiciels_amap/index.php?page=ficeler

TECHNICAL ARTICLE: PART OF A SPECIAL ISSUE ON FUNCTIONAL–STRUCTURAL
PLANT MODELLING

Automatic identification and characterization of radial files in light microscopy images of wood

Guilhem Brunel^{1,2,*}, Philippe Borianne², Gérard Subsol³, Marc Jaeger² and Yves Caraglio²

¹CIRAD – UMR A51 AMAP, France, ²University of Montpellier 2, France and ³LIRMM – CNRS,
University of Montpellier 2, France

* For correspondence. E-mail guilhem.brunel@cirad.fr

Received: 31 October 2013 Returned for revision: 9 January 2014 Accepted: 31 March 2014

• **Background and Aims** Analysis of anatomical sections of wood provides important information for understanding the secondary growth and development of plants. This study reports on a new method for the automatic detection and characterization of cell files in wood images obtained by light microscopy. To facilitate interpretation of the results, reliability coefficients have been determined, which characterize the files, their cells and their respective measurements.

• **Methods** Histological sections and blocks of the gymnosperms *Pinus canariensis*, *P. nigra* and *Abies alba* were used, together with histological sections of the angiosperm mahogany (*Swietenia* spp.). Samples were scanned microscopically and mosaic images were built up. After initial processing to reduce noise and enhance contrast, cells were identified using a ‘watershed’ algorithm and then cell files were built up by the successive aggregation of cells taken from progressively enlarged neighbouring regions. Cell characteristics such as thickness and size were calculated, and a method was developed to determine the reliability of the measurements relative to manual methods.

• **Key Results** Image analysis using this method can be performed in less than 20 s, which compares with a time of approx. 40 min to produce the same results manually. The results are accompanied by a reliability indicator that can highlight specific configurations of cells and also potentially erroneous data.

• **Conclusions** The method provides a fast, economical and reliable tool for the identification of cell files. The reliability indicator characterizing the files permits quick filtering of data for statistical analysis while also highlighting particular biological configurations present in the wood sections.

Key words: Functional–structural plant modelling, image processing, microscopic wood images, radial cell file, reliability coefficients, cellular structure, light microscopy, *Pinus canariensis*, *P. nigra*, *Abies alba*, *Swietenia*.

INTRODUCTION

Tree development is the result of both primary growth, involving the elongation and branching of the axes, and secondary growth, consisting of their thickening over time. Many approaches have been developed to study the structural and functional aspects of plants, particularly when used to address issues of carbon sequestration and wood energy. These approaches, however, whether examining secondary growth and its inter- or intra-specific variations, or the relationships between primary and secondary growth, are often based on only fragmentary studies due to the fact that data acquisition is often costly (Rossi *et al.*, 2009).

The description of primary growth can be more or less extensive thanks to retrospective approaches (Barthélémy and Caraglio, 2007; Krause *et al.*, 2010) while dendrochronological studies are generally performed only on portions of the ligneous plane (coring). Here, the more detailed the observation level, the shorter the part of the ligneous plane described by the data. Secondary growth originates from cambial activity, the source of the wood (xylem), and thus from the different cells that make up the xylem (ligneous plane).

The biological typing (Lachaud *et al.*, 1999) of cells can be approached by examining their shape, size and spatial distribution,

independent of their organization. The findings can then be used to determine the anatomical changes that characterize xylem production and differentiation phases (Thibeault-Martel *et al.*, 2008; Guédon *et al.*, 2013) without truly explaining wood production or growth mechanisms.

Indeed, the mechanisms of cambial growth can be understood only by studying cell pattern rhythmicity and cell disruption or modification in space and time. A spatio-evolutionary perspective must therefore be introduced to discuss these issues. Environmental fluctuations and their effects on the differentiation of wood elements (originating from the divisions of cambial cells) require production to be monitored in specific cellular organizations. Here, for the sake of simplicity, we consider two organizations: first, the growth ring that reflects cell production over a given time (Heinrich, 2007) and, secondly, the cell file that reflects the activity of the initial cell over time (Rozenberg *et al.*, 2004).

Nicolini *et al.* (2003) highlighted the importance of secondary growth organization by studying cell motifs, spatial rhythmicity and the variability of successive growth rings to characterize plant space occupation strategies. This approach does not, however, address the mechanisms of cell differentiation or the construction of cell motifs and structures, which are time-dependent mechanisms. The precise, dynamic monitoring of cambial

production at the initial cells (Rossi *et al.*, 2006; Rathgeber *et al.*, 2011) is technically limited to a small portion of the ligneous plane. Thus, the study of cell files, which result from cambial production and its local and overall fluctuations, is a promising approach to understanding the establishment, differentiation and temporal rhythmicity of cells (Gindl, 2001).

Biological questions concerning interactions between the development and growth of trees under environmental constraints (Moëll and Donaldson, 2001) may be addressed by: (1) determining the contribution made to conduction and storage by the different cell types in the ligneous plane; (2) determining the variability of cell characteristics (vascularization elements, fibres, ray cells, vertical parenchyma); and (3) breaking down this variability by identifying that which can be attributed to ontogeny and that which can be attributed to an environmental response.

To interpret and compare internal observations with observations of external morphology, various parts of the plant's architecture must be studied simultaneously (Barthélémy and Caraglio, 2007). In particular, observations should distinguish between the constitutive anatomical features of the plant (i.e. trunk, branch) over time, and generate a sufficiently large sample for the use of statistical tools.

However, these biological questions and their associated characterizations are highly constrained due to cumbersome protocols and the difficulties inherent in acquiring large datasets. Indeed, a precise understanding of growth mechanisms and their modelling has mostly been achieved by searching for invariants in large observation samples, thus restoring intra- and inter-species variability.

Progress made in the preparation of complete cross-sections of axes for ligneous plane observation from a macroscopic to a microscopic level without using histological sections (polishing), combined with modern image acquisition equipment (stage for programmed movements, vibratome, etc.), today provide access to quality information on broad areas of the ligneous plane. This technology, however, is hampered by its limited ability to manage and process data. Acquiring a slice 10 cm in diameter at $\times 200$ magnification requires a mosaic of about 1000 images. On average, each image contains 500 cells and the manual evaluation of each image requires about 40 min using an image editor (ImageJ, Photoshop). Cellular arrangements are identified by studying several successive rings in several growth units, and here the number of images required is too large for manual processing. Usually, counts are made on three files per section (Rossi *et al.*, 2006) to save time and reduce costs.

Automating the study of cell files would make it easier to establish tighter links to the functional and ecological aspects of the species by having a better appreciation of the range of phenomena variability. Wood provides a continuous record of all the developmental changes undergone by a tree (ontogeny, the least known) and a record over time of its environment (dendrochronology and dendroclimatology, the most studied). One particular concern with automated methods is the need to assess and quantify the reliability of their results. This means establishing reliability indicators for the results produced.

The automated identification of cell structures is one of the new challenges to be met in studies of the structural biology of plants (Quelhas *et al.*, 2011), and requires a multidisciplinary approach. In matters of (bio)-imaging, cell segmentation is a problem that has been widely discussed in the literature

(Baggett *et al.*, 2005; Fourcaud *et al.*, 2008). For example, Park and Keller (2001) combined four conventional image analysis approaches to segment cells: watershed algorithm (Vincent and Soille, 1991), snakes (McInerney and Terzopoulos, 1999), multi-resolution analysis (Jeacocke and Lovell, 1994) and dynamic programming. In the specific context of wood, cell segmentation is generally combined with the determination of cell type, i.e. its biological characterization. For example, Wang *et al.* (2008) and Marcuzzo *et al.* (2009) sorted cells using the support vector machines (SVM) classification. This is a supervised incremental training method where the major challenge is to create the training data set that is sufficiently representative to recognize the intrinsic variability of individuals.

In studies of cellular organization, certain authors have used geometric models based on topology. This also provides a description of the cell's surrounding environment. For example, the studies of Jones and Bishof (1996) and Kennel *et al.* (2010) were based on a graph that was orientated so as to extract cell files from images of gymnosperms. More specifically, cells are extracted by applying the watershed algorithm to the converted greyscale image. The adjacency graph of the cells is then built from the basin diagram. Cell typing obtained by the supervised classification method (CART; Breiman *et al.*, 1984) can thus be used to extract tracheid alignments from the graph. This method is sensitive to the reinforcement training used to classify biological types.

From a technical perspective, the searches we conducted failed to find any software solutions suitable for the automatic identification of cell files. Specialized commercially available tools such as WinCell (Hitz *et al.*, 2008) are intended for use in the analysis of wood cells, but do not recognize or characterize cellular organizations. Users have only limited or no possibility of improving their functions. By contrast, commercial platforms such as Visilog (Travis *et al.*, 1996) offer a rich environment, but are not sufficiently specialized, because although it is possible to create macro functions (by assembling and configuring basic ones), it is impossible to add new basic functions.

Open source platforms such as ImageJ (Clair *et al.*, 2007) are an interesting option as their source codes may be enriched by special functions, and specialized macros may be developed. At the same time, they facilitate method sharing and dissemination. It is thus possible to develop new functions for automated processing. We thus preferred this solution for our technical development.

In this paper we describe a cell file detection method that is based solely on cell geometry (size and shape), not on biological type, thus avoiding the bias and uncertainty inherent to the creation of reinforcement learning representing biological variability. Reliability indices are employed to determine the accuracy of the results produced for both the geometrical parameters characterizing the cell and its components (lumen and wall), and the identification of cellular alignments. Results are presented from application of the process to real images of wood slices, obtained by a plugin implementation in ImageJ (Schneider *et al.* 2012), a Java-based image analysis freeware.

MATERIALS AND METHODS

Study species

For the purposes of our study we examined histological sections and blocks of gymnosperms, specifically *Pinus canariensis*,

TABLE 1. *Sanded wood discs protocol.*

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Abrasive type	Sic Foil 320	Sic Foil 500	Sic Foil 800	Sic Foil 1200	Sic paper 4000	MD MOL
Disc rotation (r.p.m.)	150	150	150	150	150	150
Rotating sample holder	150	150	150	150	150	150
Direction of rotation	»	»	»	»	»	»
Vertical force (N)	120	120	120	120	120	120
Duration of step (min)	2	2	2	2	2	5
Liquid lubricant	Water	Water	Water	Water	Water	DP blue
Abrasive product						DP spray P3

P. nigra and *Abies alba*. Results were extrapolated to angiosperms by studying histological sections of *Khaya ivorensis*.

Any detailed study of cell typology requires an examination on multiple planes. In our study, we therefore considered only three biological cell types: tracheids, which provide support and are the main cells present in the radial cut; vessels; and rays. For each cell we calculated its circularity, height, wall thickness, and surface area of the cell and its lumen.

Preparation protocol and microscopy

Two preparation methods were evaluated to verify the general applicability of the processing and analytical techniques used.

The conventional preparation method for the study of cell files consists of creating a mosaic of histological sections from wood cores. In practice, this approach does not guarantee the juxtaposition of samples or tissue integrity. It is also relatively time-intensive as it takes approx. 1 h to prepare a mosaic of ten stained histological sections. Thus, this protocol is better suited to the study of relatively smaller areas of observation. The study of larger areas relies instead on the observation and measurements of discs of sanded wood. We thus developed a protocol that combined the sanding and polishing of wood discs to facilitate automated measurements on the entire ligneous plane.

Thin histological cuts. Producing a single cross-section of 20–25 μm thickness using a vibratome takes approx. 5–15 min depending on the case. The sections are then stained to increase the wall/lumen contrast, with the dye binding to the wall. The cleaning and staining steps take 30 min. The full protocol therefore requires approx. 1 h per cut. This method is able to produce histological section mosaics on several rings, depending on core dimensions.

Sanded wood discs. Samples were taken from pine felled in April 2013. Wood discs (about 2.5 cm thick) were cut from the centre of each growth unit. At this stage, it is important, although not always easy, to obtain samples in which the two sides are as parallel as possible. Samples were dried outdoors in the shade, sheltered from the rain, from May to July. A belt sander (Festool BS 105) was then passed over the samples using successively finer grit paper (60, 120 and 220).

The following equipment was used for the pre-polishing and polishing steps: a semi-automatic polisher fitted with a 30-cm-diameter magnetic plate (Struers); abrasive discs (Sic foil or paper depending on the grit used); wool polishing cloth (MD MOL); alcohol-based lubricant (DP blue); and diamond spray (spray DP). The discs were placed in the sample holder

and fixed with double-sided tape. Disc processing is outlined in Table 1.

After the last step, the samples were removed from the holder and rinsed with deionized water. To avoid damage, care was taken not to touch the fragile polished side of the samples. To protect from dust, the samples were wrapped in a laboratory fibre paper. Images were acquired the next day.

Scanning. Scanning, which is performed within 24 h of sample preparation, consists of producing a mosaic of images representing the observation area. The time between sample preparation and scanning must be as short as possible such that the natural drying of the wood does not produce microdeformations affecting the uniformity or ‘flatness’ of the surface, thus compromising the sharpness of the images produced.

In our study, the transverse planes were scanned by an Olympus DP71 LCD camera mounted on an Olympus BX51 microscope equipped with a movable stage, thus allowing, by gradual movements of the slide, the creation of a mosaic of images.

A magnification of $\times 200$ was used to obtain significant cell surfaces while maintaining a large enough study area.

The position of the mosaic image is determined by the x and y movements of the stage. Here, the maximum viewing area was $10 \times 10 \text{ cm}^2$, corresponding to a mosaic of about 1500 images. Each image was defined by 1600×1200 pixels, which were encoded on 3 bytes in the RGB system. The images may be blurred locally or peripherally due to optical distortions or variations in sample thickness.

Image analysis

The image analysis process followed a three-step sequence, starting with cell identification, then cell geometrical characterization and finally file detection. Cell typing and a determination of the reliability of the results obtained are detailed below.

Each image processing step was itself divided into several tasks, as described below.

Cell segmentation. This first step aimed to identify the different cells in the original image. It was divided into three tasks: image noise reduction preprocessing, contrast enhancement by filtering and colour to greyscale conversion, and cell identification from basins generated by the watershed algorithm.

Image noise reduction. At the magnifications used, the images acquired showed ‘salt and pepper’-type noise, caused by heat from the lamp. This noise was attenuated using a median filter applied on a 5×5 -pixel neighbourhood.

Contrast enhancement. The micrographs showed alternating light areas, corresponding to the lumen, and dark areas, corresponding to the wall. To enhance the contrast between different areas in the image and highlight the walls and lumens, we used a Difference of Gaussian (DoG) operator (Haddad and Akansu, 1991). This filter acts as a band-pass, only letting through the frequency range corresponding to the lumen and filtering out the noise present at higher frequencies. The first image was built from a slightly blurred Gaussian filter application with smoothing factor being selected close to average wall size. The second image was built from a heavily blurred Gaussian filter application with smoothing factor tenfold that used for the first image. The heavily blurred image was then subtracted from the slightly blurred image. These operations were performed on colour images, with negative values being set to zero. This process boosts the intensity of the lumen while substantially reducing wall intensities.

From colour to greyscale images. The choice of colour system is an important aspect of colour image processing. Several studies have shown that the ‘best system’ depends on image contents (Busin et al., 2009). For our study, we preferred the RGB system in which colours correspond to the wavelengths that stimulate the three cones of the human eye. This system can define all the hues but not all the saturations.

However, if the watershed algorithm is to be applied, a colour to greyscale image conversion is required, given that our intention was to avoid basing automated processing on human perception. Instead to use the relative luminance formula, which give less weight to blue and more weight to red and green components, we preferred to apply a simple average of the three colour channels.

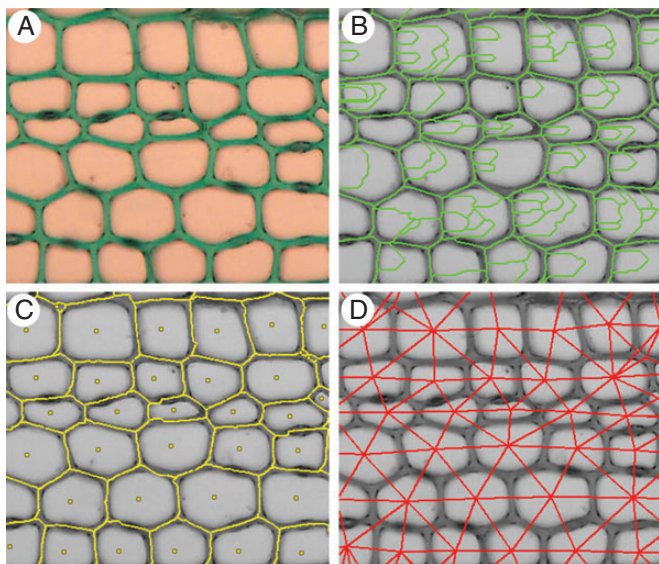


FIG. 1. Cell individualization for a Mahogany cross section. (A) Native image. (B) Watershed segmentation results with crest lines (in green) crossing the lumen. (C) cleaned watershed image with crest lines replaced by curvilinear edges (in yellow). The edges correspond to the middle lamella while the geometric centre of the watershed (extra points in yellow) correspond to the biological cells. (D) Dual watershed image, in which the red lines connect each cell centre with its neighbours (adjacency graph). Note that the degree of connections is greater than 4, showing the staggered pattern of the cells.

Cell extraction and identification. Cell extraction from the gray-scale image was based on the watershed algorithm. This powerful algorithm (Vincent and Soille, 1991) considers the image as a landscape, with altitude is given by grey values. Here, the lower points in the topographic relief, corresponding to dark regions, act as catchment basins (as if a drop of water were to fall onto the topography) and are separated by a watershed, represented by the lighter pixels.

These catchment basins are thus obtained from a partition of the image. The crest lines constitute the intercellular border, corresponding to the middle lamella (Fig. 1C). The watershed algorithm is known to produce an exaggerated partitioning, resulting in lines that *abnormally* intersect cells. This phenomenon is due to the fact that each local minimum in the image produces a potential catchment basin. Typically, two methods are used to reduce or even eliminate the exaggerated partitioning: *a priori* filtering of the local minima values (Gilmore and Kelley, 1995) used to initialize the watershed, or *a posteriori* merging of similar adjacent basins in terms of both mean and minimal intensity (Beucher, 2012). Neither method guarantees perfect cleaning of the watershed lines. Here, we propose a method based on the biological observation that the lumen of the cell is empty. Hence, if a watershed line crossing a lumen shows a characteristic intensity profile whose overall maximum is greater than a threshold experimentally set from average lumen values, then it will be deleted. Conversely, a watershed line near the middle lamella contains only values below the low-intensity threshold (Brunel et al., 2012).

After cleaning the watershed. An adjacency graph of the basins was used to describe the neighbouring relationships between the different basins. This is a simple graph conventionally defined by a set of vertices and a set of edges. More precisely, each edge of the graph connects the geometric centres of two neighbouring basins, i.e. incidents at the same crest line.

Geometric properties at the cell scale. Each cell was then individualized. Cells are composed of two anatomical structures: the wall and the lumen. On the original colour image, these structures are automatically divided into two classes by an unsupervised two-means classification algorithm (Forgy, 1965), also equivalent to Otsu’s thresholding operator. The bright pixels class corresponds to the lumen and the dark pixels class to the wall. Surface areas are calculated simply by counting the pixels in each class.

Cell geometrical properties at larger scales: the global cell arrangement orientation. When considering a basin in the adjacency graph, it can be seen that the cells are arranged in a staggered manner, and the edges of the graph are orientated in three directions. At the full graph scale, the most frequent orientation corresponds to the cellular alignments. This is defined as the main direction. We will use the method originally described by Jones and Bishof (1996) and reused by Brunel et al. (2012), studying the distribution (via a histogram) of the angle that each edge of the adjacency graph forms with the horizontal. In theory, due to the arrangement of the cells, the distribution should be trimodal and show at least one marked amplitude mode. The mode of the greatest amplitude indicates the most represented orientation in the adjacency graph, i.e. the main

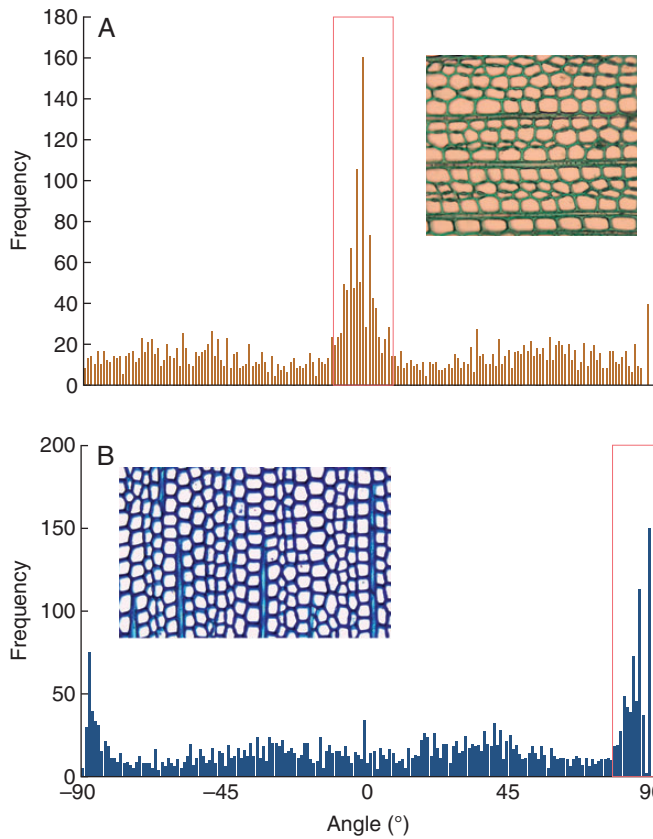


FIG. 2. Histograms showing the angle distribution on the full adjacency graph; each angle is defined by an adjacency graph edge direction on the x axis. (A) The main mode centred on 0° corresponds to horizontal files. (B) The main mode is split on the $[-90^\circ, +90^\circ]$ interval borders.

orientation of the files (Fig. 2). This mode is retrieved by searching the histogram maximum amplitude.

File detection

The process used to identify cell files in an image is based on the following assumptions: (1) files are pair-wise alignments of similar cells (in terms of size and shape) and (2) cell alignments are independent of the orientation of the image.

File identification is based on a double scale constructive approach, applied to the adjacency graph. The first step builds cell alignments, i.e. candidate cell files under spatial and geometric constraints. The second step manages the case of alignment recoveries. Finally, the resulting isolated alignments are linked to build cell file fusions.

(1) *Building alignments by applying spatial and geometric constraints.* This task extracts the longer rectilinear straight paths of ‘geometrically similar’ vertices from the adjacency graph. In other words, it detects a vertex sequence along which the surface of the underlying basin varies gradually. These alignments are constructed step by step, by successive aggregations of vertices. More specifically, it means finding which vertex w of the graph must be added at the end of the v line under construction to complete it.

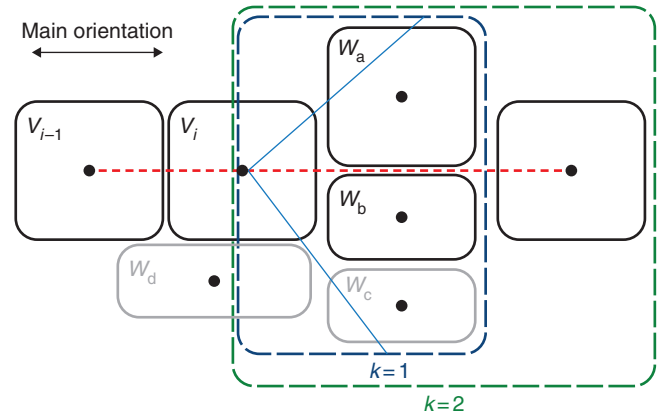


FIG. 3. The angulation (direction) constraint. The red dashed line corresponds to the file axis. The vertices (v_{i-1}, v_i) define the current alignment. The blue and green dashed line boxes show the study neighbourhood at different ranks. Cells w_a and w_b are the candidates located at a geodesic distance of 1. Vertex w_d is not considered because it is above the angular threshold (blue line) and vertex w_c is not considered because vertex w_b is closer to the file axis.

The angulation (direction) constraint. Let us define the vertex neighborhood $N_k(v_i)$ of rank k as the set of vertices located at a geodesic distance k from vertex v_i . Let us then define the angulation $A(v_i, w)$ as the angle defined from the complement to 180° of (v_{i-1}, v_i, w) where v_i and v_{i-1} are the last two vertices of the built alignment and w is the candidate vertex. The angulation can be used to choose the best candidate w that minimizes the angular offset from the reference direction, i.e. the current cell alignment. In no cases may this angle be greater than a threshold which is given by a decreasing linear function (to 35° at 10°) that itself is dependent upon the geodesic distance. For the first two file cells the angulation value is defining according to the main orientation. The vertex solution w is selected from among the w_k vertices of the $N_k(v_i)$ set and is the vertex most geometrically similar and nearest to the file axis. In fact, for a given geodesic distance k , at most two vertex candidates are considered, i.e. the two closest vertices (w_a and w_b) lying on either side of the axis formed by the last two geometric centres of the cells of the file under construction (Fig. 3).

The geometrical similarity constraint. The candidate most geometrically similar to vertex v_i is retained. The geometrical comparison is performed based on their mutual underlying basins. Let GS_{pq} be the geometric similarity index between two vertices p and q that is given by the normalized difference between the surface areas S_p and S_q

$$GS_{pq} = \frac{|S_p - S_q|}{(S_p + S_q)} \quad (1)$$

The more the index tends toward 0, the more similar the cell surface areas. This index is particularly well suited to softwoods in which the shape and surface area of the tracheid varies very little. For angiosperms, the index is a good indicator of rough size ruptures while permitting continuous and progressive surface area variations.

In practice, if the smallest GS_{pq} value is lower than 0.5, the candidate is assigned to the current alignment. If the geometric similarity index of the vertices is greater than 0.5, the candidate is

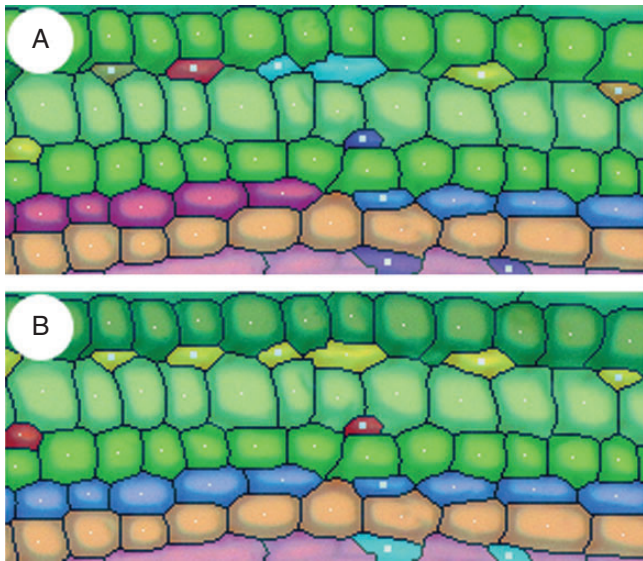


FIG. 4. Linking isolated sections. (A) Result of linkage processing without connection. (B) Result with connection. One colour is used per file.

rejected, the geodesic distance k is incremented and the process is repeated. The neighbourhood explored is progressively enlarged to a geodesic width of 5, at which point the construction of the alignment stops. This neighbourhood width was an experimental limit set arbitrarily by the wood anatomists involved in the project.

Before validating the alignment as a cell file section, a feedback control test is applied as the construction process should be reversible.

This feedback uses the main orientation of the file as a reference to reduce file drift. When a vertex w is chosen, a check is performed that the reciprocal is true, i.e. that we obtain current vertex v when falling back from vertex w using the main orientation of the files as a directional reference. If reciprocity is not confirmed, the construction of the file stops. If it is confirmed, the vertices located on the shortest path linking v to w are added to the cell alignment by the Dijkstra algorithm (Cormen et al., 2001). The criteria employed are described in the next section. Finally, vertex w becomes the new extremity of the alignment and the search process is repeated until the alignment construction stops.

Once stopped, the detected alignment builds a candidate cell file. The two next steps (recovery management and fusion) then concern the cell file level.

(2) *Managing recoveries.* During construction, some vertices may be assigned to multiple files due to the presence of biological organizations (vessels and channels) that disrupt the staggered arrangement. The assignation of these vertices to the most optimal file is solved by the shortest path search algorithm in the adjacency graph. The Euclidean distance between the geometrical centres of the watershed basin is used as a criterion. This distance sums the angle and shape considered in file building. The greater the angular deviation, the greater the Euclidean distance. The greater the size difference between successive cells, the greater the Euclidean distance. The shortest path in terms of Euclidean distance corresponds to the best alignment. As above, we use the Dijkstra algorithm here, and the

corresponding vertices are then removed from the other files, which will be fractionated.

(3) *Linking isolated cell file sections.* Due to the presence of intrusions, cell tears or limitations of the watershed algorithm, the file identification process may cause isolated file sections to appear (Fig. 4). Here, merging can be used to concentrate several parts of the files using simple topological rules to establish whole files, based on file adjacency (Brunel et al., 2012). Let us assume that two files are adjacent when two of their cells are adjacent.

The merging process is then based on two further assumptions: (1) each file crosses the image from one side to the other and (2) the files do not intersect. A file section does not fulfil the first assumption, i.e. one or both tip(s) do(es) not match with the underlying image border.

From those assumptions, we deduce a property resulting from file adjacency.

Given two file sections T2 and T2' adjacent to files F1 and F3, it is highly likely that T2 and T2' are two parts of a single F2 file. In other words, T2 and T2' are located between F1 and F2.

Note that, under the exposed assumptions, restricted cases of file merging concern only sections between two lines crossing the image from side to side. In other cases, it was decided to leave the situation at the expert's discretion and evaluation.

This method creates files or sections of files with high confidence (Fig. 5, right). Note that isolated cells or artefacts (due to segmentation errors or to poor image resolution preventing an acceptable segmentation) cannot be part of any file and no attempt is made to connect them to a file. Complex cases are left to the discretion of the user.

Typing

Cell typing is the ultimate step as used to classify the different cells present (fibrous, tracheids, vessels, rays, etc.). It is based on the geometric and densitometric characterization of the catchment basins associated with the cells' adjacency graph vertices. Unlike Kennel et al. (2010), we decided to discard supervised classification due to the difficulties encountered in providing reinforcement training that is both sufficiently complete and representative of biological variability.

A decision tree (Fig. 6) was established with wood anatomists and is based on the following observations that are sequentially tested:

- (1) The perimeter of the cells, denoted P , can be used to differentiate 'large' structures from generally 'small' cells.
- (2) The circularity coefficient (Zunic and Hirota, 2008) is given by a weighted ratio between the surface area and the squared perimeter -4π (area/perimeter²). Circularity ranges between 0.0 and 1.0: the circularity of a circle is 1, and is far less than for a starfish footprint. Values may not be valid for very small particles due to the definition of perimeter and area in a square grid.

The thresholds used in the decision tree are re-evaluated for each image by automated classification of their numerical values. The classifier used is an implementation of two-means clustering. This conventional data mining method consists of

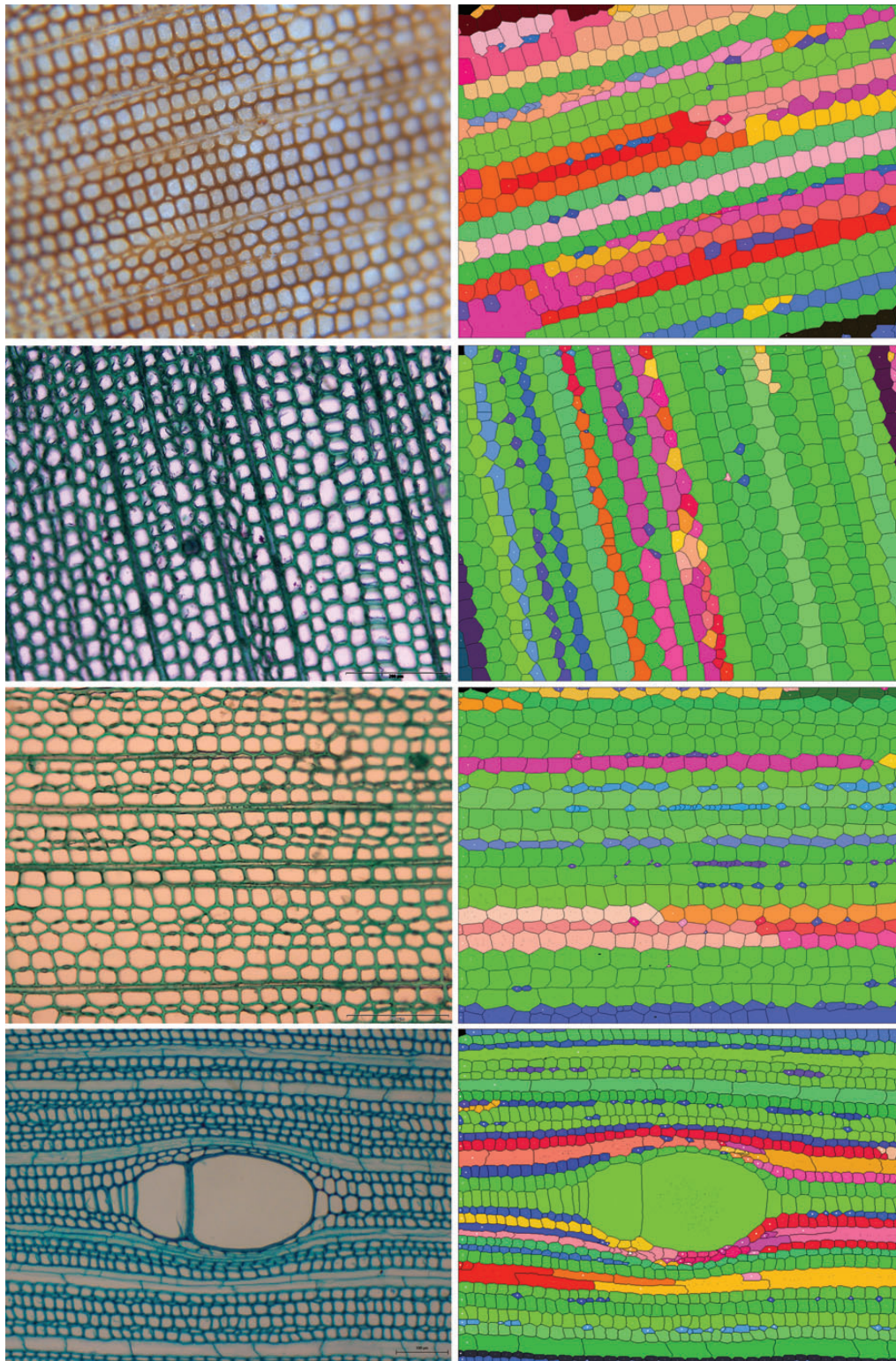


FIG. 5. Left. The native image; the images show different the preparations (sanded wood and stained histology sections), species (*Pinus*, fir, *Pinus*, mahogany) and clades (gymnosperm and angiosperm) that were processed. Right. Automatic identification of cell files; random colouring is used to enhance the visibility of the identified files (these colours should not to be confused with the reliability index colour).

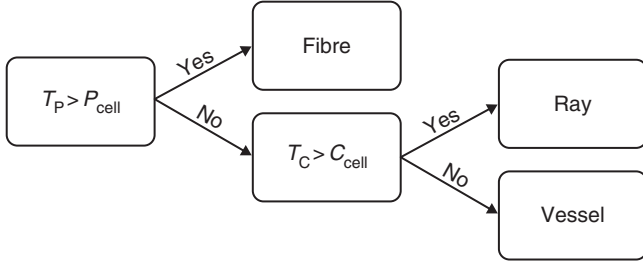


FIG. 6. Decision Tree. T_P corresponds to the threshold of the perimeter and T_C the threshold of circularity. The thresholds are automatically rated by a two-means clustering applied to all values from the image. Only cells with a lumen are processed by the decision tree.

allocating n observations in two clusters to minimize the intra- and maximize inter-class standard deviation. The threshold is therefore assigned by the median value between the upper bound of the weakest group and the lower bound of the strongest group.

Beyond the identification of alignments and biological typing, the anatomist requires a set of numerical results characterizing the shape, size and nature of more or less complex biological structures (walls, lumen, cells, files, etc.). These characteristics are defined by parameters that are automatically evaluated. In the context of mass processing, it is advantageous to be able to qualify the accuracy of these evaluations. A certainty index is thus assigned to each one calculated, whether it focuses on the cells, their components or their alignments.

Reliability

Intuitively, the computed measures will be more or less close to the true value depending on the cumulative errors stemming from image quality, biological configuration and algorithmic approximations. It is therefore important to assign each result a reliability estimation.

A file f must be sufficiently long to be significant; its length L_f must be greater than threshold L . L is defined from the length L_f distribution analysis. For each image the threshold is re-evaluated to reflect the biological characteristics inherent to each of them as a file must be composed of similar cells in terms of shape, size and appearance. For the sake of simplicity, these three criteria are reduced to the single height of cells defined from all points in the normal direction of the file. This simplification stems from the maturation mechanism of tracheids, for which cell variations only concern cell extension, not thickening. Clearly, this mechanism cannot be applied to broadleaved trees, but in practice this single criterion appears to be sufficient to characterize the continuous variation observed in the cell files of broadleaved trees. As a result, the heights H and H_{j+1} of consecutive cells in file f are compared. The reliability coefficient R_f of file f is described by a product of standardized terms ranging from 0 to 1.

$$R_f = \left(1 - \max\left(\frac{L - L_f}{L}, 0\right)\right) \prod_{j=0}^{n-1} \left(1 - \frac{|H_j - H_{j+1}|}{H_j + H_{j+1}}\right) \quad (2)$$

The closer the coefficient to 1, the more reliable file f .

By analogy, we can determine reliability coefficients for the geometrical parameters that characterize file cells. The expression for this coefficient will depend on the nature of each parameter and more specifically on the factors influencing that parameter. The surface area of the basin is independent of local image blur because it is obtained by watershed segmentation, while the surface area of the lumen is highly sensitive to local image blur because it is obtained by an intensity classification. The reliability coefficient for the lumen is therefore directly related to blur level. Ladjal (2006) defined a blur indicator that characterizes the spread of maximum intensity amplitudes, i.e. the speed at which the signal passes from the lowest to the highest intensity. This is defined by the following expression where $p(x)$ shows the pixel intensity of a basin and $p'(x)$ the variance.

$$\mathfrak{Z}(p) = \frac{\max_{x \in \mathbb{R}} p(x) - \min_{x \in \mathbb{R}} p(x)}{\max_{x \in \mathbb{R}} |p'(x)|} = \frac{\text{amplitude}(I)}{\max(|\nabla I|)} \quad (3)$$

When the coefficient is high, blur is strong. The reliability coefficient for lumen surface area (I_{ls}) is defined below where threshold T_s is defined by studying lumen variation using a Gaussian blur. Surface areas above the threshold are not significant.

$$I_{ls} = \begin{cases} 0 & \text{if } \mathfrak{Z}(p) > T_s \\ 1 - \frac{\mathfrak{Z}(p)}{T_s} & \text{else} \end{cases} \quad (4)$$

Note that the accuracy of the surface areas and circularities are sensitive to the size of the objects measured; for example, the circularity of a discrete circle (Andrès, 1994) is meaningful only above 10 pixel rays.

Parameter computation

Our method produces several layers of results corresponding to different observation levels: (1) radial files are classified according to their length and their cellular self-similarity, (2) cells are characterized by geometric size, diameter, shape, etc., and (3) cell components (wall, lumen) are characterized by geometric parameters (size, thickness, diameter, etc.).

The method has been implemented in Java and integrated as a plug-in for ImageJ freeware. As an indication, the plug-in allows images consisting of 1600×1200 pixels to be processed in less than 20 s on a computer with an Intel Q720 processor clocked at 1.6 GHz.

RESULTS AND DISCUSSION

The study described herein aimed to compare the results produced by our method with those in an expert database containing 12 images whose cells had been manually cropped by wood anatomists. Our results were obtained without any specific user settings (Fig. 5): all parameters were automatically re-evaluated by the application based on the intensity histogram generated for each image. The different tests conducted allowed us to evaluate the limitations of our method, namely with regard to different biological contexts, different preparation protocols, and diverse and varied acquisition conditions. The tests were essentially

TABLE 2. Outline of some significant results: size of processed images, number of extracted cells, CPU time on a PC with an Intel Xeon at 2-3 GHz, and total Quality Index defined as the ratio of number of files automatically identified and reconstructed over total number of lines; for the purposes of the expert analysis, the cells were manually cropped using ImageJ software

Species	Size (pixels)	Expert		Plugin		% of files detected
		Cell numbers	Times (s)	Cells numbers	Times (s)	
Mahogany	1024 × 768	1302	2520	1359	14.3	86
Fir	1360 × 1024	787	1300	800	12.4	88
Black Pine 1	1600 × 1200	1794	2750	1873	23.2	92
Caribbean Pine	1360 × 1024	819	1500	828	11.5	91
Black Pine 2	1600 × 1200	1411	2450	1458	16.1	83

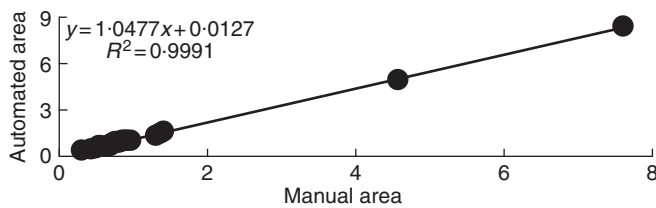


FIG. 7. Study of basin surfaces on mahogany. The x-axis shows the normalized areas obtained by the fully manual method. The y-axis shows those obtained by the automated method. The coefficient of determination is close to 1, showing a good fit.

motivated by simple questions: (1) is our method suitable for both angiosperms and gymnosperms? and (2) is our method sensitive to preparation protocols, to optical blur, to luminosity levels, etc.?

Qualitative performance

At this stage in our studies, the performance of our method was subject to a qualitative assessment by comparing – in a fairly global manner – the results produced with those of an expert appraisal. Of course, it is always possible to compare ‘processing’ time, number of cells detected or files identified (Table 2). For example, Table 2 shows that the automatic method detects a consistent number of cells about 80–100-fold more rapidly than the expert, and this regardless of preparation protocol, anatomical configuration or magnification used, and that, overall, our method works well (88 % of files detected on average) on images presenting a structured organization with a marked wall/lumen contrast.

Geometric precision

We have seen that files are created only from geometric and topological rules applied to the basin adjacency graph. It is therefore important to ensure the robustness of the basin detection, i.e. the method used must be precise in cell cropping and insensitive to image blurring.

Figure 7 compares about 60 normalized areas obtained by the manual and automated methods. The areas were divided by the median area to refine the regression. The coefficient of determination tends toward 1, showing that the areas are closely correlated. The slope of the regression is weakly greater than 1, indicating slight over-estimation by the automated method.

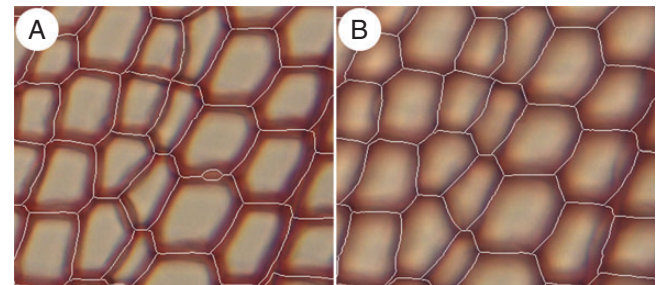


FIG. 8. Comparison of watershed results on a *Pinus* cross-section. (A) Crest lines from a sharp image. (B) Crest lines from a blurred image. Basin outlines are virtually identical between the sharp and blurred images. The sharp image has a small additional basin corresponding to wall detachment caused by cutting. This basin was deleted in the supernumerary removal phase.

The low shift of 0.0127 confirms the middle error of 5 % on right areas. The automated method appeared to over-estimate the values (or the expert appraisal under-estimated the values). The only certainty is that the automated method is repeatable.

The major limitation to the automatic file identification method stems from image content, namely the photometric characteristics of the image and the structural biological layouts.

On file identification, only geometric accuracy is studied. The geometric shift is deliberately not taken into account as it is considered to be negligible with respect to cell surface area. However, the study could feasibly be extended to geometric shift, namely to assess the geometric overlap between expert annotations and segmentation algorithms.

A study was conducted on the effect of optical blur, which is often present in images and is caused by peripheral deformations due to the microscope lens or sample flatness. The algorithmic stability of our method was evaluated by acquisitions of the same image area with different adjustments. A statistical evaluation of cell cropping variability was then performed. We did highlight the fact that the intercellular lines obtained by the watershed algorithm are generally unchanged by image blurring. They correspond, in fact, to gradient slope inversions: the image smoothing produced by the optical blur reduces the intensity dynamic of the image without removing the changes in intensity variations. Overall, the crest lines remain unchanged (Fig. 8).

Figure 9 compares 60 areas obtained from a sharp and a blurred image. The coefficient of determination tends toward 1 showing that the areas are closely correlated. The slope of the regression is slightly less than 1, indicating slight under-estimation for blurred

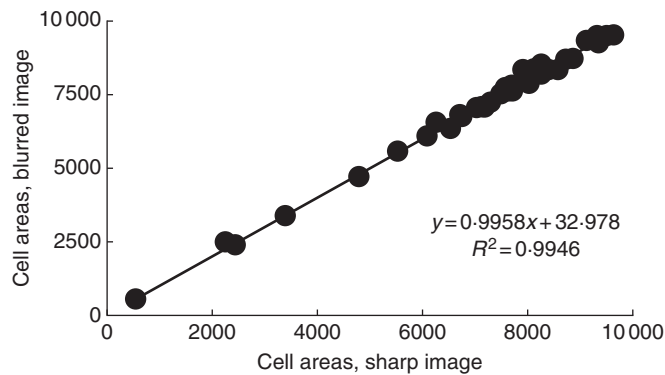


FIG. 9. Basin surface areas study on *Pinus*, with cell areas expressed as pixels. The x-axis shows the areas estimated from sharp images. The y-axis shows those estimated from simulated blurred images. The slope of the regression is slightly less than 1, indicating minor under-estimation of basin areas in the blurred images.

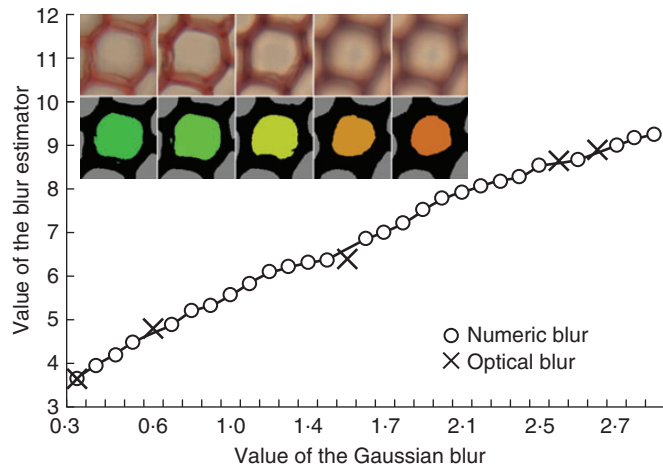


FIG. 10. The behaviour estimator shows the blur tendency and the red crosses correspond to the five optical microscopy blurs. The five lumen areas are shown with their corresponding optical blurred snapshots. The colour indicates the reliability coefficient, from green for true lumen area to orange for reduced lumen area.

images. The method therefore yielded almost identical results for both sharp and blurred images.

Reliability and blur

Image blurring is a source of potentially major bias when evaluating lumen areas and wall widths. As indicated before, the cell segmentation is based on a two-means clustering. This classification is sensitive to the intensity dynamics of the cellular points and is therefore modified if intensity dynamics are reduced by optical blur (Fig. 10). We observed that the surface area of the lumen for a given cell was reduced when blur increased, i.e. when intensity dynamics decreased.

The local blur indicator introduced earlier is used to estimate the error on the surface area measurements of cell constituents. The figure on the left shows the consistent behaviour of the blur

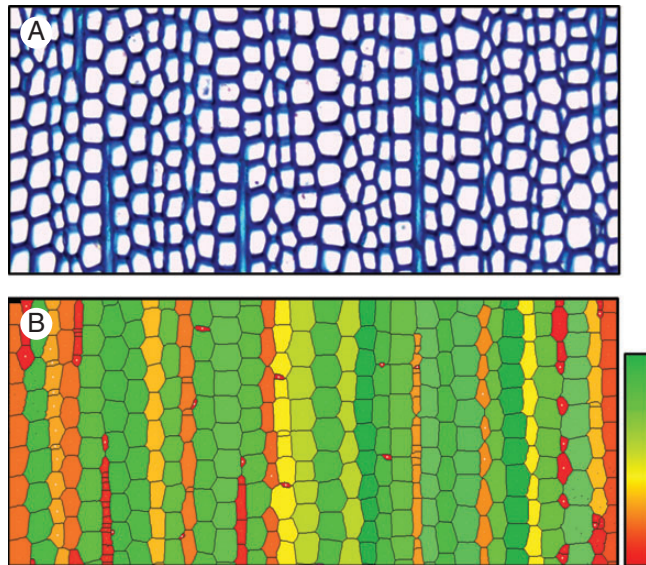


FIG. 11. (A) Cross-section of native *Abies*. (B) The automated identification of radial files. The colour value indicates the reliability coefficient: poorly reliable files appear in red while very reliable files appear in pale green.

estimator that was calculated from cells considered in images that were digitally disrupted by an incremental Gaussian blur.

Therefore, the local blur indicator can be used to filter the results and retain only the sharpest cells. It is used, in particular, to create mosaic images (Brunel et al., 2013), for retention of the most reliable basins. It could also possibly provide a correction factor for measures made on areas that are more or less subjected to optical blur.

Reliability of files

The overall reliability coefficient can be used for file characterization of files, in terms of both significant expression and geometric behaviour, i.e. a ‘good’ file is a sufficiently long line in which cell size varies gradually. This coefficient acts directly to filter the files (Fig. 11). Files with a good coefficient are shown in green while those with a poor coefficient are in red. The red–orange hues indicate configurations with cuts or sharp discontinuities. This estimator is well suited to softwoods with very organized and continuous cell arrangements. In angiosperms, it may be used for a rapid visualization of vessels and adjacent files. These files are useful for analysing the endogenous geometric constraints caused by vessels on files.

This indicator is ultimately a very integrated aspect of the files identified. It is more comprehensive than the performance indices mentioned in the first part of the results, and is also more complex to explain as it goes beyond file identification reliability to actually reflect file quality. For example, some alignments, shown in yellow in the image below, appear to be ‘visually correct’ while their coefficient is rather low. They are penalized by excessively elevated surface area variations. For example, the third orange file from the left looks to be good, but small basins can be seen within the file arising from over-segmentation caused by wall detachment. Files can be filtered for statistical analysis thanks to their cell coefficients.

General conclusions

The method described herein works correctly for images that show both high contrast between walls and lumens (without local color inversion) and a visible cellular organization. Under these conditions, all the obvious lines determined by qualified experts are correctly identified with significant time savings (a typical manually expert identification required 10 h on our samples). Moreover, the certainty index allows for selective exploitation of the results for statistical studies.

CONCLUSIONS

The automated analysis of anatomical wood sections will contribute greatly to understanding secondary plant growth and development. Herein, we propose a fully automatic method for the recognition of operational cell files in mass processing.

We have also established reliability coefficients to characterize the files, their components and their measurement. These reliability coefficients are mainly used to filter results for statistical studies conducted by botanists and aim to draw their attention to specific biological configurations.

We have shown that the method is equally applicable to digital images of stained histological wood sections and images of sanded ligneous planes.

File identification is based on detection of cell alignments showing gradual geometric variations. The cells in the image are automatically cropped by watershed lines calculated from a set of judiciously filtered local minima, thus reducing the risk of exaggerated segmentation. Cell files are defined by the longest straight vertices that are 'geometrically similar' to the adjacency graph created from the watershed catchment basins. These alignments are constructed one after the other by successive aggregations of vertices in progressively enlarged neighbourhoods. Unlike traditional approaches, cell typing here is not based on the creation of reinforcement training, but on a decision tree established with wood anatomists.

In addition to file detection, the cells are characterized by a set of values (wall thickness, lumen surface, cell circularity, etc.) that are necessary for the statistical analysis of files and their variability. To ensure that the measures are relevant, reliability indices have been created, based on existing methods and biological reality (file length and continuity). These indices are used to select the most important files or highlight particular biological disturbances or configurations.

The percentage of unusable files is offset by the quantity of images that can be processed. Mass processing allows us to study spinal files up to the cambium over several wood growth units.

These initial results are already very interesting and confirm that the method has great potential and could be suitable for statistical studies and the detection of invariants or trends in large observation samples. Comparative studies undertaken with images annotated by experts have shown that the proposed method is time-efficient and generally insensitive to the optical blur potentially present in micrographs. A blur indicator has been introduced to study the stability of the method and in the short term could be used to correct the measurements obtained. Finally, a reliability indicator to characterize files, in terms of both significant expression and geometric behaviour, is given either to draw

the attention of experts to specific biological configurations or to remove potentially erroneous data from statistical studies.

ACKNOWLEDGEMENTS

We thank the NUMEV Laboratory of Excellence (University of Montpellier 2) and the SIBAGHE Graduate School of the University of Montpellier 2 for their support. We extend special thanks to Christine Heinz and Christophe Godin for their availability and the relevant and constructive observations they provided during different interviews and meetings. Additionally, a big thank you to François Pailler and Merlin Ramel for their patience and help in preparing the different samples and optimizing technical protocols. Special thanks also to Michaël Guérout who took the time to provide assistance and share his skills and knowledge in the use of optical microscopy.

LITERATURE CITED

- Andr  s E. 1994. Discrete circles, rings and spheres. *Computers and Graphics* **18**: 695–706.
- Baggett D, Nakaya M, McAuliffe M, Yamaguchi TP, Lockett S. 2005. Whole cell segmentation in solid tissue sections. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **67**: 137–143.
- Barth  l  my D, Caraglio Y. 2007. Plant architecture: a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Annals of Botany* **99**: 375–407.
- Beucher S. 2012. Towards a unification of waterfalls, standard and P algorithms. CMM/Mines Paristech publication.
- Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and regression trees*. London: Chapman and Hall/CRC.
- Brunel G, Borianne P, Subs  l G, Jaeger M, Caraglio Y. 2012. Automatic characterization of the cell organization in light microscopic images of wood: Application to the identification of the cell file. In: Guo Y, Kang MZ, Dumont Y, eds. *Plant growth modeling, simulation, visualization and applications. The Fourth International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications*, Shanghai, China, 58–65.
- Brunel G, Borianne P, Subs  l G, Jaeger M. 2013. Simple-graphs fusion in image mosaic: application to automated cell files identification in wood slices. In: K  m  r  inen J-K, Koskela M, eds. *Lecture notes in computer science. Image analysis*. Berlin: Springer, 34–43.
- Busin L, Shi J, Vandenbroucke N, Macaire L. 2009. Color space selection for color image segmentation by spectral clustering. *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 262–267. <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5478603&queryText=3DColor+space+selection+for+color+image+segmentation+by+spectral+clustering>.
- Clair B, Gril J, Di Renzo F, Yamamoto H, Quignard F. 2007. Characterization of a gel in the cell wall to elucidate the paradoxical shrinkage of tension wood. *Biomacromolecules* **9**: 494–498.
- Cormen TH, Cazin X, Kocher G-L. 2001. *AD. Introduction to algorithms*. Cambridge, MA: MIT Press.
- Forgy E. 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics* **21**: 768–769.
- Fourcaud T, Zhang X, Stokes A, Lambers H, K  rner C. 2008. Plant growth modelling and applications: the increasing importance of plant architecture in growth models. *Annals of Botany* **101**: 1053–1063.
- Gilmore P, Kelley C.T. 1995. An implicit filtering algorithm for optimization of functions with many local minima. *SIAM Journal* **5**: 269–285.
- Gindl W. 2001. Cell-wall lignin content related to tracheid dimensions in drought-sensitive Austrian pine (*Pinus nigra*). *Iawa Journal* **22**: 113–120.
- Gu  don Y, Taugourdeau O, Caraglio Y, Sabatier S. 2013. Characterizing the balance between ontogeny and environmental constraints in forest tree development using growth phase duration distributions. In: Nygren P, Lintunen A, Godin C, Nikinmaa E, Siev  nen R, eds. *Proceedings of the 7th International Conference on Functional Structural Plant Models*, Saariselk  , Finland, 237–239.

- Haddad RA, Akansu AN. 1991.** A class of fast Gaussian binomial filters for speech and image processing. *IEEE Transactions on Signal Processing* **39**: 723–727.
- Heinrich L. 2007.** Reaction wood varieties caused by different experimental treatments. TRACE, Tree Rings in Archaeology. *Climatology and Ecology* **5**: 224–232.
- Hitz OM, Gärtner H, Heinrich I, Monbaron M. 2008.** Wood anatomical changes in roots of European ash (*Fraxinus excelsior* L.) after exposure. *Dendrochronologia* **25**: 145–152.
- Jeacocke MB, Lovell BC. 1994.** A multi-resolution algorithm for cytological image segmentation. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, 1994.* IEEE, 322–326.
- Jones R, Bischof L. 1996.** A graph-based segmentation of wood micrographs. Computing science and statistics 28. *Proceedings in Computing Science and Statistics* **28**: 12–20.
- Kennel P, Subsol G, Guérout M, Borianne P. 2010.** Automatic identification of cell files in light microscopic images of conifer wood. In: Djemal K, Deriche M, eds. *2nd International Conference on Image Processing Theory, Tools and Applications.* IEEE, 98–103.
- Krause C, Rossi S, Thibeault-Martel M, Plourde P-Y. 2010.** Relationships of climate and cell features in stems and roots of black spruce and balsam fir. *Annals of Forest Science* **67**: 402–402.
- Lachaud S, Catesson A-M, Bonnemain J-L. 1999.** Structure and functions of the vascular cambium. *Comptes Rendus de l'Académie des Sciences – Series III. Sciences de la Vie* **322**: 633–650.
- Ladjal S. 2006.** Estimation du flou dans les images naturelles. In: *Quinzième congrès francophone Reconnaissance des formes et intelligence artificielle (RFIA).* François-Rabelais University Press (in French).
- Marcuzzo M, Quelhas P, Campilho A, Mendonça AM, Campilho A. 2009.** Automated Arabidopsis plant root cell segmentation based on SVM classification and region merging. *Computers in Biology and Medicine* **39**: 785–793.
- McInerney T, Terzopoulos D. 1999.** T-Snakes: topology adaptive snakes. *Medical Image Analysis* **4**: 840–845.
- Moëll MK, Donaldson LA. 2001.** Comparison of segmentation method for digital image analysis of confocal microscope images to measure tracheid cell dimensions. *IAWA Journal* **22**: 267–288.
- Nicolini E, Caraglio Y, Pélissier R, Leroy C, Roggy J. 2003.** Epicormic branches: a growth indicator for the tropical forest tree, *Dicorynia guianensis* Amshoff (Caesalpinaceae). *Annals of Botany* **92**: 97–105.
- Park J, Keller JM. 2001.** Snakes on the watershed. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**: 1201–1205.
- Quelhas P, Nieuwland J, Dewitte W, Mendonça AM, Murray J, Campilho A. 2011.** *Arabidopsis thaliana* automatic cell file detection and cell length estimation. In: Kamel M, Campilho A, eds. *Image analysis and recognition.* Berlin: Springer, 1–11.
- Rathgeber CBK, Rossi S, Bontemps J-D. 2011.** Cambial activity related to tree size in a mature silver-fir plantation. *Annals of Botany* **108**: 429–438.
- Rossi S, Deslauriers A, Anfodillo T, et al. 2006.** Conifers in cold environments synchronize maximum growth rate of tree-ring formation with day length. *New Phytologist* **170**: 301–310.
- Rossi S, Simard S, Rathgeber CBK, Deslauriers A, Zan CD. 2009.** Effects of a 20-day-long dry period on cambial and apical meristem growth in *Abies balsamea* seedlings. *Trees* **23**: 85–93.
- Rozenberg P, Schüte G, Ivkovich M, Bastien C, Bastien J-C. 2004.** Clonal variation of indirect cambium reaction to within-growing season temperature changes in Douglas-fir. *Forestry* **77**: 257–268.
- Schneider CA, Rasbaand WS, Eliceiri KW. 2012.** NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.
- Thibeault-Martel M, Krause C, Morin H, Rossi S. 2008.** Cambial activity and intra-annual xylem formation in roots and stems of *Abies balsamea* and *Picea mariana*. *Annals of Botany* **102**: 667–674.
- Travis AJ, Hirst DJ, Chesson A. 1996.** Automatic classification of plant cells according to tissue type using anatomical features obtained by the distance transform. *Annals of Botany* **78**: 325–331.
- Vincent L, Soille P. 1991.** Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**: 583–598.
- Wang M, Zhou X, Li F, Huckins J, King RW, Wong STC. 2008.** Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. *Bioinformatics* **24**: 94–101.
- Zunic J, Hirota K. 2008.** Measuring shape circularity. *Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes. Computer Science* **5197**: 94–101.

Defining reliability coefficients in an automated radial file identification and characterization method in microscopic images of gymnosperms

Guilhem Brunel^{1,2}, Philippe Borianne², Gérard Subsol³, Marc Jaeger² and Yves Caraglio²

¹CIRAD - UMR AMAP, France. ²Université Montpellier 2, France. ³CNRS – LIRMM, France.
guilhem.brunel@cirad.fr

Highlights: The analysis of wood anatomical sections is of great interest for understanding the growth and development of plants. We propose a novel method for automatically identifying and characterizing radial files in wood microscopic images of gymnosperms. A key point is to be able to assign *a priori* reliability coefficient to the results, particularly for statistical processing in large-scale analyses. We describe in this paper the principle used to establish reliability coefficients to evaluate the radial file identification process and the geometrical measurements of cells and their components.

Keywords: image processing, wood microscopic images, radial file identification, *a priori* reliability.

INTRODUCTION

Wood structure reflects the physiological and molecular regulation of cambium activity (Cato et al. 2006) and also registers environmental conditions (Barlow et al. 2005). Understanding cambium growth mechanisms calls for a study of cell pattern regularity, and of cell disruption or modification in space and time (Liang et al. 1997). Two types of organizations are considered: the growth ring representing cell production at a given time, and the radial file representing the activity of an initial cell over time (Rozenberg et al. 2004); radial files help in understanding the development, differentiation and temporal changes of cells.

Morphological fluctuations of successive cells along a radial file and the influence of external factors have been investigated for a long time (Ford et al. 1978), but those investigations were greatly restricted by a limited number of radial file comparisons and a restrictive number of successive cell descriptions. Those studies, concerning secondary growth and its relationships with primary growth, were based on fragmentary studies, due to high acquisition and processing costs. Large-scale studies for understanding wood growth and disturbance are available nowadays thanks to automated radial file identification and cell characterization (Brunel et al. 2012, Kennel et al. 2011). In fact, wood continuously records changes in the development of the tree. Such new methodologies and approaches offer real new opportunities for more in-depth investigations of tree biology and climate change (Fonti et al. 2010). The size and shape fluctuations observed along radial files are usually considered to be the result of external constraints (Frankenstein et al. 2005). This is a restrictive and simplified viewpoint: these fluctuations are probably due to a combination of external and internal factors. We aim to promote an objective quantification of these measurements, on the basis of statistics based on numerous image datasets.

This paper focuses on the reliability of such data from statistical studies, i.e. produced by automated radial file identification (Brunel et al. 2012). For large-scale data analyses, it is necessary to ensure *a posteriori* that the measurements are valid in order to find invariants or validate hypotheses. This is a classic line taken in statistics (Bruton et al. 2000) and experimental approaches.

MATERIALS

We processed histological sections of several gymnosperm species: *Pinus caramanica*, *Pinus nigra* and *Abies alba*. Wood cross-sections with a thickness of 20 µm were produced with a vibratome. The sections were stained to increase the contrast between the lumen and cell walls. They were then digitized with an Olympus DP71 LCD camera on an Olympus BX51 microscope. The square color images produced had a resolution of 4 million pixels.

OVERVIEW OF THE METHOD

Radial files are alignments of substantially similar cells in terms of color dynamics, shapes and sizes. Intuitively, the notion of cell alignment implies the existence of cellular organization based on the neighborhood relationships between cells. Those properties are used to identify radial files.

Our approach, described in (Brunel et al. 2012) was divided into three steps: cell identification which specified single cells, cell organization which detected radial files, and cell classification which gave the biological and qualitative typing of cells and radial files.

The incremental construction of radial files was based on two assumptions: (i) such a file is independent from image orientation; (ii) two consecutive cells of a file are very similar. From a methodological point of view, cell individualization was derived from a watershed algorithm (Vincent et al. 1991), and the cell pattern was described by an adjacency graph created from the watershed crest lines. Radial files were then built by finding a reversible path in the graph under spatial constraints (maintaining a specific direction) and similarity constraints (two consecutive cells should show close geometric characteristics). Our method produced several layers of results, respectively corresponding to different observation levels: (i) radial files were classified according to their length, their fractionation and their cell self-similarity, (ii) cells were characterized by geometric parameters (such as size, diameter, shape, etc.) and topological parameters (number of neighbors), (iii) cell component elements (wall, lumen) were characterized by geometric parameters (size, thickness, diameter, etc.).

DEFINITION OF A PRIORI RELIABILITY COEFFICIENTS

Quality assessment of these results depends on many factors. It concerns the data set itself (image noise, biological configuration complexity, etc.), the process (algorithm approximations, computation costs, stability, etc.), and the geometrical scale (related to the observation level). Each result should be qualified using different indicators, notably to keep the significant outcome while processing high volume data. It is thus relevant to assign an estimated reliability coefficient to each result. We illustrate cases on file and cell scales.

Radial file construction reliability. We propose defining this radial file reliability coefficient R_f from three criteria: length, fractionation and similarity. The first criterion derives from the file length, L_f , compared to the threshold value T . T is set by 2means and defines the minimum significant length. Fractionation is evaluated from the file construction algorithmic cost. It involves the total number of cells L_{fm} stacked to build the file, the length L_f of the file, and the number of sections N_f . The similarity criterion is built from the product of the $n-1$ consecutive cell surface variations. The final reliability coefficient R_f of the file f is then defined by a product of normalized terms ranging between 0 and 1. When this coefficient tends to 1, the reliability of the radial file increases.

$$R_f = \left(1 - \max\left(\frac{T - L_f}{T}, 0\right)\right) \left(1 - \frac{L_{fm} - L_f}{N_f * L_{fm}}\right) \prod_{j=0}^{n-1} \left(1 - \frac{|Sf_j - Sf_{j+1}|}{Sf_j + Sf_{j+1}}\right)$$

Lumen area estimation reliability. The reliability coefficient of the lumen surface is defined on a similar principle, by a product of two normalized terms, based on the cell size and an image local blur estimation. In fact, the watershed algorithm is known as a robust method for extracting cell boundaries, since it is based on intensity discontinuities which are less sensitive to blurring, whereas the lumen area shows high sensitivity to local blur. The cell area definition involves a two-class clustering split, corresponding to the lower and the higher intensity classes. The local blur estimator is derived from (Ladjal 2006), based on the relationship between local dynamics and intensity differences.

EXPERIMENTS, CONCLUSION AND WORK IN PROGRESS

Experiments were performed on ten colored sections of different species of gymnosperms, representative of biological variability. The color quantifies the reliability results, both on a global (Fig. 1) and local (Fig. 2) scale. On a global scale, we can see in Fig. 1 that the reliability estimator properly qualifies the cell files: green colors are assigned to straight and regular cell lines whereas red colors correspond to incomplete cells on the side of the image or incrustated cells. Such a color map alerts the expert to potential unsure classifications and emphasizes the ambiguous biological status of a file or cells. On a local scale, for lumen area computation, the estimator also retrieves obvious cases. The reliability factor also ensures areas showing lower intensity variations (areas in green in Fig. 2, right).

Several points were studied. (i) in angiosperms, the vessel morphology is different and calls for a different evaluation of cell similarity *which defines the third term of the file reliability coefficient*, by taking into account the larger diameter instead of its surface. (ii) other blur estimators were studied, such as estimators based on multi-scale filtering in order to assess the independence of the reliability coefficient with

respect to the blur estimation method. (iii) how the local reliability coefficient could be used to correct the measured values according to the local blur estimator relationship (Fig 2. Left).

To conclude, we introduced some reliability coefficients for image processing related to cell organization identification and illustrated the cases of radial file detection and of lumen area computation. The reliability coefficients were essentially used to filter the input data of statistical studies conducted by botanists or to draw attention to special biological configurations.

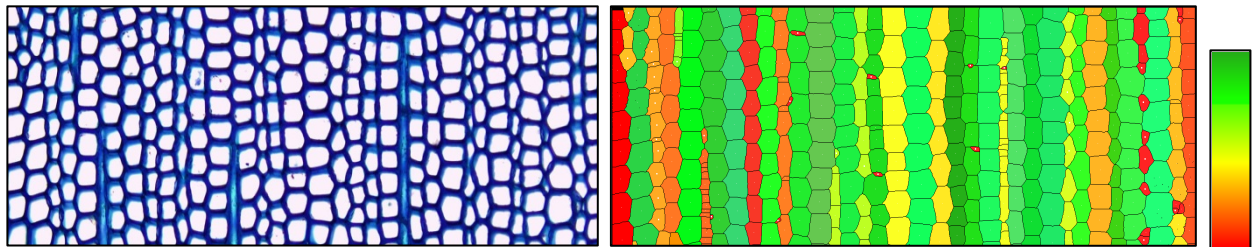


Fig. 1. Left: cross-section of *Abies alba*. Right: automated identification of radial files. Color value qualifies the reliability coefficient: weakly reliable files appear in red tones and highly reliable files are in green tones.

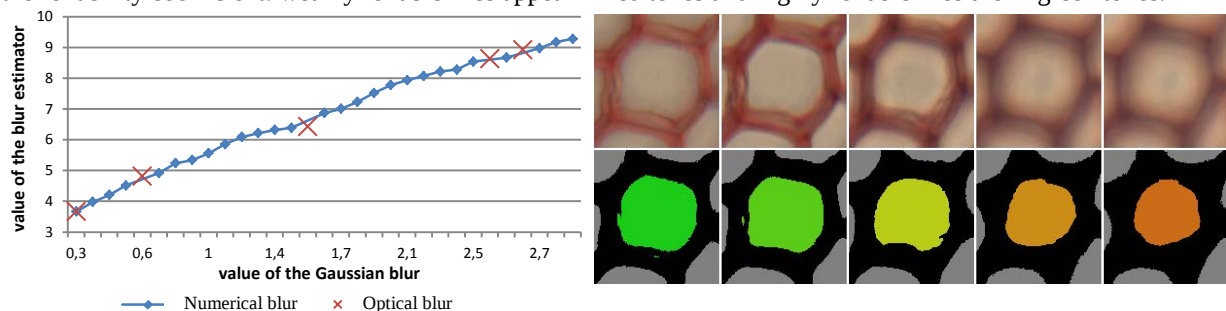


Fig. 2. Left: relationship between a Gaussian blur applied to a cell image and the local blur estimator. The estimator emphasizes the blur tendency; red crosses stand for five microscopic optical blurs, matched from their area variance. Right: the five corresponding lumen areas from the optical blurs: the color qualifies the reliability coefficient, from green for the real lumen area to orange for the reduced lumen area.

The authors acknowledge NUMEV Labex and SIBAGHE Graduate School of the University of Montpellier 2 for their support.

LITERATURE CITED

- Barlow PW, Powers SJ. 2005. Predicting the environmental thresholds for cambial and secondary vascular tissue development in stems of hybrid aspen. *Annals of Forest Science* **65**: 565-573.
- Brunel G, Borianne P, Subsol G, Jaeger M, Caraglio Y. 2012. Automatic characterization of the cell organization in light microscopic images of wood: application to the identification of the cell files, *Plant Growth Modeling, Simulation, Visualization and Applications*, IEEE press, ISBN 978-1-4673-0070-4, 58-65.
- Bruton A, Conway JH, Holgate ST. 2000. Reliability: What is it, and how is it measured?, *Physiotherap* **86**:94-100
- Cato S, McMillan L, Donaldson L, Richardson T, Echt C, Gardner R. 2006. Wood formation from the base to the crown in *Pinus radiata*: gradients of tracheid wall thickness, wood density, radial growth rate and gene expression. *Plant Molecular Biology* **60**:565–581.
- Fonti P, von Arx G, Garcia-Gonzalez I, Eilmann B, Gärtner H, Eckstein D. 2010. Studying global change through investigation of the plastic responses of xylem anatomy in tree rings. *New Phytologist* **185**: 42–53.
- Ford ED, Robards A, Piney MD. 1978. Influence of environmental factors on cell production and differentiation in the early wood of *Picea sitchensis*. *Annals of Botany* **42** (3): 683-692.
- Frankenstein C, Eckstein D, Schmitt U. 2005. The onset of cambium activity – A matter of agreement? *Dendrochronologia* **23**: 57-62.
- Kennel P, Subsol G, Guérout M, Borianne P. 2010. Automatic identification of cell files in light microscopic images of conifer wood. *2nd International Conference on Image Processing Theory Tools and Applications* 98–103.
- Ladjal S. 2006. Blur estimation in Natural Images, *15^e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle* 112-124.
- Liang C, Filion L, Cournoyer L. 1997. Wood structure of biotically and climatically induced light rings in eastern larch (*Larix laricina*). *Canadian Journal of Forest Research* **27**: 1538.1547.
- Rozenberg P, Schüte G, Ivkovitch M, Bastien C, Bastien JC. 2004. Clonal variation of indirect cambium reaction to within-growing season temperature changes in Douglas-fir. *Forestry* **77**:257-268.
- Vincent L, Soille P. 1991. Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations, *IEE Transactions on Pattern Analysis and Machine Intelligence* **13-6**:583–598.

Simple-Graphs Fusion in Image Mosaic: Application to Automated Cell Files Identification in Wood Slices

Guilhem Brunel^{1,2,*}, Philippe Borianne¹, Gérard Subsol³, and Marc Jaeger¹

¹ CIRAD - UMR AMAP, France

² Université Montpellier 2, France

³ CNRS – LIRMM, France

guilhem.brunel@cirad.fr

Abstract. Results aggregation by disjoint graph merging is potentially a good alternative to image stitching. During the processing of image mosaics, it allows to be free of radiometric and geometric corrections inherent in image fusion. We have studied and developed a generic merging method of disjoint graphs for tracking cell alignments in image mosaics of wood.

Keywords: graphs theory, graphs fusion, image processing, pattern recognition, cell segmentation, cell organization.

1 Introduction

Graph theory is widely used in image processing [1] especially for region encoding or image segmentation. The image is then considered as a graph whose vertex set is made of pixels, and edge set is given by an adjacency relation between them. Numerous segmentation methods based on region merging are proposed in literature, the most known being the Split and Merge algorithm. Several authors combine it with the watershed algorithm by using graph fusion to segment grey [2] or color [3-4] images. In all these approaches, the fusion operates on the vertices of a single graph.

Our specific segmentation method of cell alignments, so called cell files, in wood slices produces an adjacency graph. Nevertheless, identifying cell files on very large areas given by images mosaic implies either to process the image resulting from combining all the images of the mosaic or to merge the results produced by processing each image.

In the first case, stitching is the method which is the most often used. Image stitching combines multiple images with overlapping fields of view in order to produce a high-resolution image. Most of the algorithms require nearly exact overlaps between images and identical exposures to produce seamless results [5-6]. Moreover, they often request local deformation correction [7-9] due to optical acquisition properties, or local offset rectification [10] due to the automated views shooting. However, the resulting high resolution images may still be difficult to be processed, especially from the final resolution (up to several billion pixels) and the possible radiative variation due to the microscope acquisition properties.

* Corresponding author.

So, we preferred to use the second class of methods. Our framework is based on the understanding of the cambium growth mechanisms which requires the studies of cell patterns regularity, of their disruption or of their modification through space and time [11-12]. We are interesting by radial files (cell alignments); they help understanding the development, the differentiation and the temporal changes of cells [13]. The automation of the cell files identification and the characterization of their components [14-15] allow today to process a large quantity of images. However, we aim to study a large area of observation, following cells of some microns diameter on a given wood slice of several centimeters diameter. But these recognition processes being constrained to single image detections; studies of large area are thus impracticable. Identifying radial files on large areas described by images mosaic involves therefore merging the results produced by the processing of each image.

Our dataset is defined by color images mosaic of wood histological slices. Transversal slices of 20 μm thickness are prepared with a vibratome, colored, and then fit in between glass strips before being digitalized by a camera coupled with the microscope. The mosaic images (Fig. 1.) are successively generated by progressive displacement of the microscope stage taking the strip. The position of the image in the mosaic is given by the stage displacement in abscissa X and ordinate Y. The observation area is limited to $10 \times 10 \text{ cm}^2$, requiring nearly 1500 images mosaic. Each image is encoded in RGB system, with a resolution of 1600 by 1200 pixels. The images may show local or peripheral blur due to the optical deformations or specimen thickness variations.

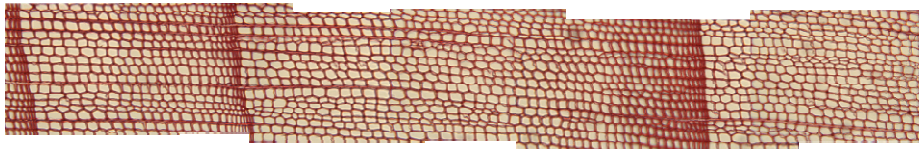


Fig. 1. Seven color images mosaic of Black Pinus overlapping a $2 \times 0.25 \text{ cm}^2$ area

We wish to avoid images stitching difficulties related to the high resolution and the radiative corrections. We propose an approach mixing gradually the processing of images and the progressive merging of the produced results [16-17]. The merging process takes benefits from the result description using a graph. Such a model is well adapted to cells organization. The results merging of two adjacent images consist thus of their two respective graphs amalgamation.

The paper extends a preliminary work published in *Fourth International Symposium on Plant Growth Modeling* [14] presenting radial files identification in wood slice image. The following section describes the notion of simple-graph. Section 3 reminds the automated cell files identification principles in the microscopic images. Section 4 presents the graph merging and the fusion kernel notion. Results and perspectives are presented and discussed in the section 5 before conclusion.

Our contribution lies in the graphs fusion formalism and its application:

- We introduce the Simple Graph notion to describe cell organization and the fusion kernel used to amalgamate two graphs;
- We introduce a similarity function in order to define the fusion kernel, i.e. the vertices of the graphs which should be amalgamated.
- We introduce a local blur indicator to define the best representative vertex in vertices fusion.

2 Graph Theory Notion

This section gives some general definitions about graphs merging mechanism. The only considered specificity is the nature of the graph, linked to the properties of cell files. We present here the formalization of the graphs amalgamation, i.e. when two distinct graphs are gathered by merging of one of their vertices.

2.1 Simple Graph or S-Graph

A graph $G=(V,E)$ is a mathematical structure composed of two finite sets V and E . The elements of V are called vertices, and the elements of E are called edges. A simple graph or S-Graph is an undirected graph that has no loops and no more than one edge between any two different vertices. In a simple graph, the edges of the graph form a set and each edge is a distinct pair of vertices. In a simple graph of n vertices, every vertex has a degree that is less than n (the converse, is not true — there exist non-simple graphs with n vertices in which every vertex has a degree smaller than n).

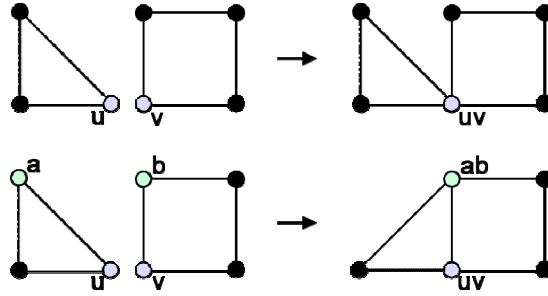


Fig. 2. Top, simple fusion of two nodes of disjoint S-graphs. Bottom, complex fusion of four nodes and two edges of disjoint S-graphs.

2.2 Multiple S-Graph Amalgamation

Let $G_i = (V_i, E_i)$ and $G_j = (V_j, E_j)$ be two disjoint S-graphs, with $K_i = (A_i, B_i)$ a sub S-graph of G_i and $K_j = (A_j, B_j)$ a sub S-graph of G_j . Let $f : K_i \rightarrow K_j$ be an involution without fixed point between these subgraphs.

$$\forall u \in K_i \quad f(u) \in K_j \quad \text{and} \quad f(f(u)) = u \quad \text{with} \quad f(u) \neq u$$

The amalgamation of G_i and G_j modulo the isomorphism f is the S-graph obtained from the union $G_i \cup G_j$ by merging each vertex a and each edge b in the subgraph K_i with their respective images $f(a)$ and $f(b)$ in subgraph K_j . The amalgamated vertex is noted $a f(a)$, and the amalgamated edge is noted $b f(b)$.

The vertex set of this new graph is $(V_i - A_i) \cup (V_j - B_j) \cup \{a f(a) \mid a \in A_i\}$, and the edge set is $(E_i - B_i) \cup (E_j - B_j) \cup \{e f(e) \mid e \in B_i\}$; any edge which had $a \in A_i$ or $f(a) \in A_j$ as endpoint now has the amalgamated vertex $a f(a)$ as endpoint. The amalgamation result is a S-graph: each edge appears one and only time in the edges set as shown in Fig. 2.

2.3 Amalgamation Kernel

The amalgamation kernel is the set of the merged vertex couples defined as

$$UV = \{uv \mid u \in A_i \text{ and } v \in A_j \text{ and } v = f(u) \}$$

3 The Cell File Identification Principle

The cell files identification process [14] leads on the following principles: (i) the files are alignment of two by two similar cells, i.e. close in terms of size, shape and dynamic; (ii) the cells alignment are independent of the image orientation: the method have to be reversible.

From methodological point of view, the contrast between cell wall and lumen is enhanced by a difference of Gaussians on the color image (Fig. 3.). Then, the image is converted in grey levels by the simple mean of the color channels.

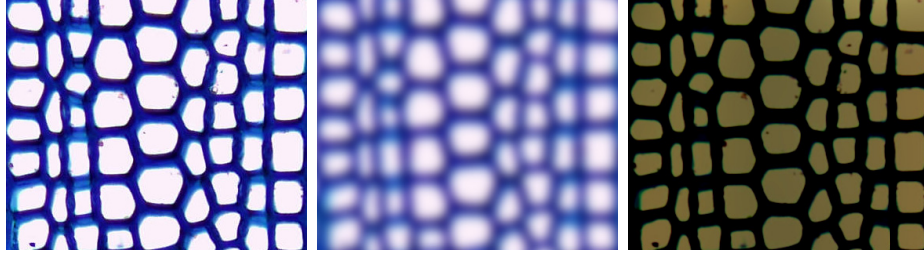


Fig. 3. Image contrast enhancing by a Difference of Gaussian filtering on color images of Pinus. Left: the image obtained by applying a small Gaussian blurring ($\sigma = 3$ pixels). Center: the image resulting from a strong Gaussian blurring ($\sigma = (\text{image width})/10$). Right: subtraction of previous images which increases the contrast between wall and lumen.

The cells individualization is built from watershed process [18]; the crest lines define regions whose limits correspond to the inter-cells separation (Fig. 4.). The cells adjacency graph is built from these lines [19]. The cell files are identified by reversible breadth-first search exploration of the graph under spatio-geometrical constraints: they are defined as the most linear paths whose adjacent vertices present similar characteristics, the vertex taking attributes describing shape, size and aspect of the cell.

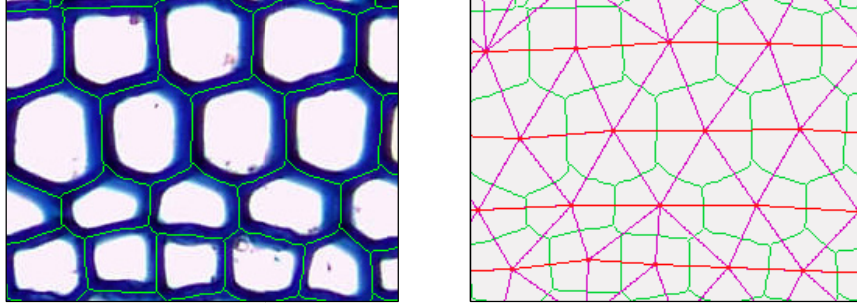


Fig. 4. Cells identification and its adjacency graph. Legend: in green, the crest lines; in red and pink, the adjacency graph; in red, the cell files. Left: the crest lines produced by the watershed algorithm define a good approximation of the cell contours. Right: the adjacency graph is built from the crest lines; the cell files are defined by the most linear paths of the graph.

4 The Graphs Fusion Implementation

This section specifies how the general definitions are implemented in our study context, in particular the way we define the amalgamation kernel.

Let $G=(V,E)$ be a given simple graph. At each vertex v of V is associated a set of spatio-geometrical attributes describing the cell c represented by v :

- r_v , the radius of the circumscribed circle of c ;
- x_v , the cell geometrical centre abscissa;
- y_v , the cell geometrical centre ordinate;
- s_v , the cell area.

4.1 Similarity Coefficient

The similarity coefficient estimates the likeness between vertices of distinct S-graphs. It is defined by a single product of attribute differences: each attribute has the same weight in the similarity evaluation. This coefficient tends toward 1 for similar vertices, towards 0 for too different or distant vertices.

Let $G_0 = (V_0, E_0)$ and $G_1 = (V_1, E_1)$ be two given S-graphs.

$\forall v \in G_i, \forall w \in G_{1-i}, i \in \{0,1\}$, the coefficient is defined by:

$$sim_{v,w} = \begin{cases} 0 & \text{if } |v, w| > \min(r_v, r_w) \\ \left(1 - \frac{|x_v - x_w|}{x_v + x_w}\right) \left(1 - \frac{|y_v - y_w|}{y_v + y_w}\right) \left(1 - \frac{|s_v - s_w|}{s_v + s_w}\right) & \text{if } |v, w| \leq \min(r_v, r_w) \end{cases}$$

This definition distinguishes the case of too distant vertices from the case of potentially fusion-able vertices. Each term is a Bray-Curtis dissimilarity estimator based on the spatial position or the size of cells. If necessary, it can be completed with other criteria; shape, dynamic...

4.2 Best Candidate Function

This function gives the best vertex fusion-able to a given vertex when it exists. We propose here the following expression:

$$\beta(v) \begin{cases} \emptyset & \text{if } \forall w \in G_{1-i}, \text{sim}_{v,w} = 0 \\ \{w\} \text{ where } w \in G_{1-i} \text{ and } \text{sim}_{v,w} \text{ is maximal, else} \end{cases}$$

Our study case concerns successive images analysis where the overlapped areas are identical without geometrical offsets. The function β is commutative: it is independent of the terms order and of the indexes i and j . Reduced to the subset of vertices having a not empty image, β becomes an involution without fixed points which allows defining the amalgamation kernel.

$$\forall v \in G_i, i \in \{0,1\} \mid \beta(v) \neq \emptyset \Rightarrow \beta(v) \in G_{1-i} \text{ and } \beta(\beta(v)) = v$$

The key point is the area overlapping (at least one vertex satisfies the upper property). It is defined as such that bigger structure of the image is wholly overlapped. This point is developed in the result section.

5 Results and Perspective

Several aspects are presented to illustrate the robustness of the amalgamating method. Points in the study are discussed.

Amalgamation of different adjacency graphs of mosaic images allows tracking the radial lines over large areas of observation, as shown on two rings in (Fig. 5).

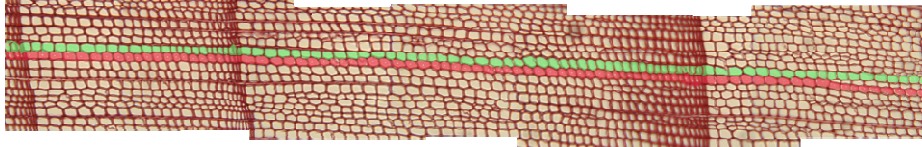


Fig. 5. Cell files through the mosaic shown in Fig 1

Transition between two successive rings is characterized by a cells morphological rupture. Note that this rupture does not affect the graphs amalgamation process, but the file tracking in image. The approach is stable regarding semi global cell property evolution.

5.1 Influence of Blur on the Constitution of Cell Files

Several studies have been initiated to check the blur independence of watershed which is the origin of adjacency graphs. Experiment consisted in comparing the lines of crest and the cells area on an image progressively blurred.

No significant gap is observed, illustrating good correlation between cell areas (Fig. 6. Left). The colored lines illustrate the well-known over-segment linked to the

watershed algorithm. The Waterfall method [20] could contribute to reduce the observed differences. This method uses the result of the watershed transform on gradients of the original image to produce a sequence of images by merging some adjacent regions. Otherwise, the compared fiftieth areas are well adjusted (Fig. 6 right): the coefficient of determination and the slope of linear line tend towards 1.

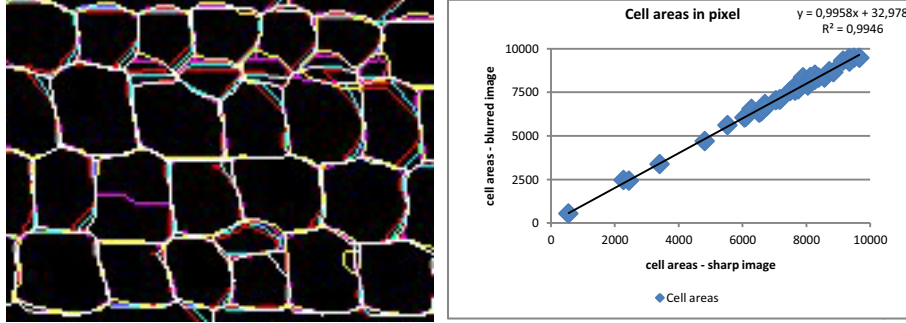


Fig. 6. Left: superposition of crest lines produced by a watershed applied to a gradually image blurred: crest lines are displayed in different color, and colors are added in case of superposition; are thus displayed in white of the image points common to all crest lines; differences are due to the over-segmentation. Right: comparison of cell area a gradually image blurred. The variation is smaller when the cell surface is higher.

5.2 Fusion Function

The fusion method is nearly independent from a covering rate and a blur value in the mosaic images result process (Fig. 7). Invariance rises from the robustness of the amalgamating function using only geometric criteria, insensitive to blur.

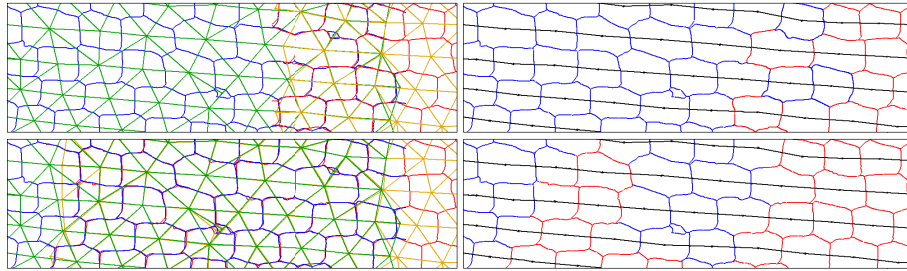


Fig. 7. Left, different overlapping of two same distinct graphs. Right, the result of two fused graphs, the cell lines are displayed in black. Bottom, with a high overlapping the results are essentially identical: difference comes from cell chosen (the least blurred).

We merge two cells by choosing one of the vertices of the fusion kernel (see previous §), and more specifically the less blurred cell. The local blur estimator, based on the relationship between local dynamics and intensity differences [21], is used to finalize the choice. In addition, knowledge of stage displacement is an advantage,

especially to identify the potentially cells at merging, i.e. close enough. We are currently studying the possibility of replacing in the function fusion the position terms by a term of the dynamics of cell: this point requires characterizing the local blur in image to normalize the signature of each cell. Such an indicator is presented in the definition of reliability coefficients [22] and used to validate the produced quantitative results. Use this coefficient to correct measurements is seriously considered, particularly to homogenize the intensities cells signing.

5.3 Overlapping of Images

The overlapping rate of images depends on the size of the bigger structure of the image. The gymnosperms (Fig. 5.) are composed of very regular cells, with weak size variation. It is easy to fix the optimal overlap during the images shooting so that at least one cell is wholly in overlapping area. The overlapping value is experimentally set to thirty micrometers, i.e. close to the three biggest cell sizes. This gives a safety margin to process the disturbed biological configurations. On the other hand, the Angiosperms (Fig. 8) present very different sizes of structures: set the overlapping size is difficult and is actually left to the biologist. Here, the overlapping width is 400 micrometers so that the bigger cell is wholly overlapped.

These big cells, called vessels, cause progressive or brutal cell deformations. Images contain cells with size and shape strongly variable. It is usual to find vessels 10 to 20 times larger than fibers.

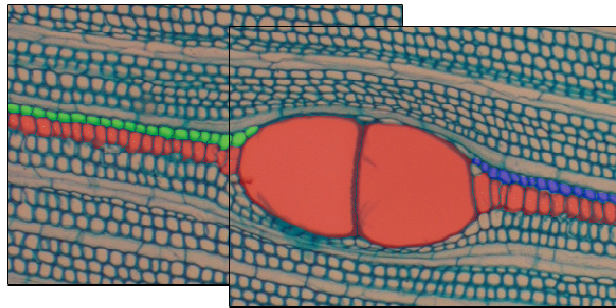


Fig. 8. Case of high structure geometry changes: graphs fusion still works when the biggest structure is totally recovered

In Fig. 8, the central file (in red) is properly assembled, thanks to the full vessel structure coverage. Squashing of fibers produced by vessel development causes discontinuities in peripheral files (in green and blue), discontinuities that fusion method cannot correct. It is necessary to implement a method of files concatenation from study of files relationship. Complementary work will soon be engaged to automatically set the optimal covering by studying the distribution of cell sizes. Study will seek a solution for cases where the covering of vessels can only be partial.

6 Conclusion

Image mosaic processing is a classical research thread in vision. We presented a generic method for aggregating results by merging disjoint single-graphs, i.e. undirected graphs without loops; as an alternative to image stitching.

We defined a fusion function of vertices based on the evaluation of a similarity coefficient. This function is an involution without fixed point allowing the determining of the amalgamating kernel, i.e. the set of points that are two by two merged. The similarity coefficient does not depend on the graph structure, but only on attributes contained by vertices of the graphs. So it is adaptable to other attributes in regard to the study object.

We have developed and implemented this method in the context of tracking cell files in image mosaics describing large viewing zones. Cells individualization in each image was obtained by watershed algorithm; the obtained crests lines are used to define the cells adjacency graph, which allows to identify cell files. It is defined as the straightest paths whose adjacent nodes have similar attributes. Aggregation of files on consecutive images in the mosaic is realized by fusing graphs in order to avoid the inconvenience related to the stitching: radiometric equalization, image repositioning, blur correction and local deformation compensation...

Applying the method requires only to define explicitly the similarity coefficient and involution. The similarity function is defined by a product of normalized Bray-and-Curtis dissimilarity estimators describing the geometrical aspects -position and size- of cells which can potentially be merged. In case of merging, only the less blurred cell is retained and the local blur estimator is based on the relationship between local dynamics and intensity differences. A study is engaged to introduce a component characterizing the dynamics of the cell, and so to be free of positioning criteria. Finally, the fusion method is insensitive to image blurring.

This generic method could be applied on other kinds of images of very high definition when we wish to avoid stitching as for example satellite images.

Acknowledgment. The authors gratefully acknowledge Yves CARAGLIO and Christine HEINZ for their involvement in this work, their advices and availability. A big thank to Michael GUEROULT without whom the laboratory work would not have been possible.

This work is jointly funded by a doctoral fellowship of the Labex NUMEV, by the SIBAGHE Graduate School of the University Montpellier 2 and by the Scientific Council of the University Montpellier 2.

References

1. Lézoray, O., Grady, L.: Image processing and Analysis With Graphs. CRC Press (2012)
2. Cousty, J., Bertrand, G., Couprie, M., Najman, L.: Fusion graphs: merging properties and watersheds. *Journal of Mathematical Imaging and Vision* 30, 87–104 (2008)

3. Baldevbhai, P.J., Anand, R.S.: Review of graph, medical and color Image base Segmentation Techniques. *Journal of Electronical and Electronics Engineering* 1, 1–19 (2012)
4. Peng, B., Zhang, L., Zhang, D., Yang, J.: Image segmentation by iterated region merging with localized graph cuts. *Pattern Recognition* 44, 10–11 (2013)
5. Brown, M., Lowe, D.G.: Automatic Panoramic Image Stitching using Invariant Features. *International Journal of Computer Vision* 74, 59–73 (2007)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
7. Clenti, C.: Architectures flot de données dédiées au traitement d'images par Morphologie Mathématique, Thèse de doctorat en morphologie mathématique. In: ENSMP, vol. (431) (2009)
8. Sun, C., Beare, R., Hilsenstein, V., Jackway, P.: mosaicing of microscope images with global geometric and radiometric corrections. *Journal of microscopy* 224, 158–165 (2006)
9. Vercauteren, T., Perchant, A., Malandain, G., Pennec, X., Ayache, N.: Robust mosaicing with correction of motions distortions and tissue deformations for In Vivo Fibered Microscopy. *Medical Image Analysis* 10, 673–692 (2006)
10. Thévenaz, P., Unser, M.: User-friendly Semiautomated Assembly of Accurate Image Mosaics in Microscopy. *Microscopy Research And Technique* 70, 135–146 (2007)
11. Liang, C., Filion, L., Cournoyer, L.: Wood structure of biotically and climatically induced light rings in eastern larch (*Larix laricina*). *Canadian Journal of Forest Research* 27, 1538–1547 (1997)
12. Wilson, B.F.: *The Growing Tree*. The University of Massachusetts Press, Amherst (1970)
13. Gindl, W.: Cell-wall lignin content related to tracheid dimensions in drought-sensitive austrian pine (*pinus nigra*). *Iawa Journal* 22, 113–120 (2001)
14. Brunel, G., Borianne, P., Subsol, G., Jaeger, M., Caraglio, Y.: Automatic characterization of the cell organization in light microscopic images of wood: application to the identification of the cell files. In: *Plant Growth Modeling, Simulation, Visualization and Applications*, vol. 4, pp. 58–65. IEEE Press (2012) ISBN 978-1-4673-0070-4
15. Kennel, P., Subsol, G., Guérout, M., Borianne, P.: Automatic identification of cell files in light microscopic images of conifer wood. In: *2nd International Conference on Image Processing Theory Tools and Applications*, pp. 98–103 (2010)
16. Sjodahl, M., Oreb, B.: Stitching interferometric measurement data for inspection of large optical components. *Optical Engineering* 41, 403–408 (2002)
17. Wyant, J.C., Schmit, J.: Large field of view, high spatial resolution, surface measurements. In: *International Conference on Metrology and Properties of Engineering Surfaces*, vol. 38, pp. 691–698 (1998)
18. Vincent, L., Soille, P.: Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEE Transactions on Pattern Analysis and Machine Intelligence* 13, 583–598 (1991)
19. Tremeau, A., Colantoni, P.: Regions adjacency graph applied to color image segmentation. *IEEE Transactions on Image Processing* 9, 735–744 (2000)
20. Beucher, S.: Watershed, Hierarchical Segmentation and Waterfall Algorithm. In: *Proc. Mathematical Morphology and its Applications to Image Processing*, pp. 69–76 (1994)
21. Ladjal, S.: Blur estimation in Natural Images. In: *15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle*, pp. 112–124 (2006)
22. Brunel, G., Borianne, P., Subsol, G., Jaeger, M., Caraglio, Y.: Results reliability of the automated identification of cell files in microscopic images of gymnosperms. In: *7th International Conference on Fuctional-Structural Plant Models* (2013) (in submission)

