



HAL
open science

Analyse statistique de populations pour l'interprétation d'images histologiques

Maya Alsheh Ali

► **To cite this version:**

Maya Alsheh Ali. Analyse statistique de populations pour l'interprétation d'images histologiques. Traitement du signal et de l'image [eess.SP]. Université Sorbonne Paris Cité, 2015. Français. NNT : 2015USPCB003 . tel-01191479

HAL Id: tel-01191479

<https://theses.hal.science/tel-01191479>

Submitted on 4 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Thèse de doctorat

présentée pour obtenir le grade de docteur
de l'UNIVERSITÉ PARIS-DESCARTES

École Doctorale EDITE

Spécialité : Traitement du signal et des images

Maya ALSHEH ALI

Analyse statistique
de populations pour
l'interprétation
d'images histologiques

Soutenue publiquement le 19/02/2015 devant le jury composé de :

Xavier DESCOMBES	Président
Xavier DESCOMBES	Rapporteurs
Nicolas PASSAT	
Vannary MEAS-YEDID	Examineurs
Benoît NAEGEL	
Farida CHERIET	
Laurent WENDLING	Directeur de thèse
Thomas HURTUT	Encadrant

Cette thèse s'est déroulée au laboratoire LIPADE (Laboratoire d'Informatique de PARIS DEscartes) dans l'équipe SIP (Systèmes Intelligents de Perception) de l'Université Paris Descartes.

Laboratoire LIPADE
Equipe SIP
Université Paris Descartes
45 rue des Saints Pères
75270 Paris cedex FRANCE

Téléphone : (+33)1 83 94 57 41

Référence Bib_TE_X :

```
@PHDTHESIS{AlshehAli2015,  
  author = {Alsheh Ali, M.},  
  title = {Analyse statistique de populations pour  
l'interprétation d'images histologiques},  
  school = {Universit\`e {P}aris-{D}escartes},  
  year = {2015}  
}
```


Résumé

Au cours de la dernière décennie, la pathologie numérique a été améliorée grâce aux avancées des algorithmes d'analyse d'images et de la puissance de calcul. Néanmoins, le diagnostic par un expert à partir d'images histopathologiques reste le *gold standard* pour un nombre considérable de maladies notamment le cancer. Ce type d'images préserve la structure des tissus aussi proches que possible de leur état vivant. Ainsi, cela permet de quantifier les objets biologiques et de décrire leur organisation spatiale afin de fournir une description plus précise des tissus malades. L'analyse automatique des images histopathologiques peut avoir trois objectifs : le diagnostic assisté par ordinateur, l'évaluation de la sévérité des maladies et enfin l'étude et l'interprétation des mécanismes sous-jacents des maladies et leurs impacts sur les objets biologiques.

L'objectif principal de cette thèse est en premier lieu de comprendre et relever les défis associés à l'analyse automatisée des images histologiques. Ensuite, ces travaux visent à décrire les populations d'objets biologiques présents dans les images et leurs relations et interactions à l'aide des statistiques spatiales et également à évaluer la significativité de leurs différences en fonction de la maladie par des tests statistiques. Après une étape de séparation des populations d'objets biologiques basée sur la couleur des marqueurs, une extraction automatique de leurs emplacements est effectuée en fonction de leur type, qui peut être ponctuel ou surfacique. Les statistiques spatiales, basées sur la distance pour les données ponctuelles, sont étudiées et une fonction originale afin de mesurer les interactions entre deux types de données est proposée. Puisqu'il a été montré dans la littérature que la texture d'un tissu est altérée par la présence d'une maladie, les méthodes fondées sur les motifs binaires locaux sont discutées et une approche basée sur une modification de la résolution de l'image afin d'améliorer leur description est introduite. Enfin, les statistiques descriptives et déductives sont appliquées afin d'interpréter les caractéristiques extraites et d'étudier leur pouvoir discriminant dans le cadre de l'étude des modèles animaux de cancer colorectal.

Ce travail préconise la mesure des associations entre différents types

d'objets biologiques pour mieux comprendre et comparer les mécanismes sous-jacents des maladies et leurs impacts sur la structure des tissus. En outre, nos expériences confirment que l'information de texture joue un rôle important dans la différenciation des deux modèles d'implantation d'une même maladie.

Mots-clés : *pathologie numérique, modèles animaux de cancer colorectal, détection des objets biologiques, organisation spatiale, interaction spatiale de structures hétérogènes, analyse de la texture, aide au diagnostic, glaucome, statistiques déductives.*

Abstract

Title : Statistical Analysis of Populations for Histological Images Interpretation

During the last decade, digital pathology has been improved thanks to the advance of image analysis algorithms and calculus power. However, the diagnosis from histopathology images by an expert remains the gold standard in a considerable number of diseases especially cancer. This type of images preserves the tissue structures as close as possible to their living state. Thus, it allows to quantify the biological objects and to describe their spatial organization in order to provide a more specific characterization of diseased tissues. The automated analysis of histopathological images can have three objectives : computer-aided diagnosis, disease grading, and the study and interpretation of the underlying disease mechanisms and their impact on biological objects.

The main goal of this dissertation is first to understand and address the challenges associated with the automated analysis of histology images. Then it aims at describing the populations of biological objects present in histology images and their relationships using spatial statistics and also at assessing the significance of their differences according to the disease through statistical tests. After a color-based separation of the biological object populations, an automated extraction of their locations is performed according to their types, which can be point or areal data. Distance-based spatial statistics for point data are reviewed and an original function to measure the interactions between point and areal data is proposed. Since it has been shown that the tissue texture is altered by the presence of a disease, local binary patterns methods are discussed and an approach based on a modification of the image resolution to enhance their description is introduced. Finally, descriptive and inferential statistics are applied in order to interpret the extracted features and to study their discriminative power in the application context of animal models of colorectal cancer.

This work advocates the measure of associations between different types

of biological objects to better understand and compare the underlying mechanisms of diseases and their impact on the tissue structure. Besides, our experiments confirm that the texture information plays an important part in the differentiation of two implemented models of the same disease.

Keywords : *digital pathology, colorectal cancer animal models, biological object detection, spatial organization, spatial interaction of heterogeneous structures, compute-aided diagnosis, texture analysis, glaucoma, inferential statistics.*

Remerciements

Il me sera très difficile de remercier tout le monde parce que c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme. Je tiens avant tout à remercier Laurent Wendling et Thomas Hurtut pour m'avoir donné l'opportunité d'avoir cette expérience importante.

Je remercie tous les membres de projet SPIRIT pour avoir fourni les problématiques médicales et les images, en particulier Johanne Séguin et Nathalie Mignet.

Je tiens également à remercier Xavier Descombes et Nicolas Passat pour avoir accepté de rapporter mon travail ainsi que pour leurs remarques constructives. Je remercie également Vannary Meas-Yedid, Benoît Naegel et Farida Cheriet pour avoir accepté de participer à mon jury de thèse.

Mes remerciements sincères vont également à Georges Stamon pour ses conseils et son soutien tout au long de mon passage dans l'équipe SIP.

Je remercie également tous mes collègues de l'équipe SIP : Hassan, Marwen, Num-Jun, Pierre-Edouard, Adrien, Adam, Sameh, Soumaya, Halima et Michaël pour les nombreuses discussions et pour les bons moments passés ensemble. Merci à Camille Kurtz et Nicole Vincent. Je remercie aussi Nicolas Loménié pour son aide et ce qu'il a apporté à ma thèse.

Mes remerciements s'adressent également aux membres de l'équipe du LIV4D de l'École Polytechnique de Montréal pour les moments agréables passés ensemble, dans et hors du laboratoire, en particulier à Farida Cheriet encore une fois pour m'avoir accueillie, ainsi qu'au personnel de l'entreprise DIAGNOS inc., notamment Timothée Faucon pour avoir fourni les problématiques et les images, pour les nombreuses discussions et le foot. Merci à Philippe, Sébastien, Cosmin, Lama, Mahnaz et son mari Behzad, Argyrios, Cédric et Jérémy, pour les débats sur l'actualité et les activités mémorables tous ensemble.

Je tiens également à remercier Thibault Lagache pour son aide et le temps qu'il a pris afin de répondre à mes questions.

Merci à mes amis qui sont en Syrie ou en France pour leur soutien et leurs encouragements ainsi que d'être toujours là pour moi.

Merci à tous les membres de ma famille, surtout merci à ma soeur Inès,

qui m'ont toujours encouragée et soutenue moralement.

Finalement, je tiens à remercier Mickaël, mon complice dans cette aventure, sans qui rien n'aurait été possible. Depuis le début, il a toujours été là pour me réconforter, m'encourager, m'aider, discuter et m'écouter patiemment répéter mes présentations. Cette thèse est aussi la tienne.

J'en oublie certainement encore et je m'en excuse. Encore un grand merci à tous pour m'avoir conduite à la fin de ma thèse.

Table des matières

Résumé	iii
Abstract	v
Remerciements	vii
Table des matières	xi
1 Introduction	1
1.1 Motivation	2
1.2 L’histopathologie et son intérêt	4
1.3 Procédures pré-analytiques	5
1.3.1 Introduction à la préparation des échantillons	5
1.3.2 Marquage du tissu	6
1.3.3 Stratégies d’acquisition des images en histologie	10
1.4 Procédures analytiques et interprétatives	11
1.4.1 Analyse manuelle	11
1.4.2 Analyse assistée par ordinateur	12
1.4.3 État de l’art de l’analyse assistée par ordinateur	13
1.5 Matériel biologique à disposition	19
1.5.1 Contexte d’application	19
1.5.2 Matériel	20
1.6 Contribution personnelle et organisation de ce manuscrit	22
2 Détection des objets biologiques dans les images histologiques	23
2.1 Motivation	24
2.2 État de l’art	26
2.2.1 Séparation	26
2.2.2 Segmentation d’images histologiques	30
2.3 Séparation des marqueurs	33
2.4 Extraction des différents objets biologiques	39
2.4.1 Définitions des catégories d’objets utilisées	39
2.4.2 Extraction des objets ponctuels	39
2.4.3 Extraction des objets surfaciques	41
2.5 Application	43
2.5.1 Sur des images hématoxyline et DAB	43
2.5.2 Sur des images hématoxyline et éosine	46
2.6 Résumé et perspectives	49

3	Statistiques spatiales	51
3.1	Motivation	52
3.2	État de l'art pour les données ponctuelles	55
3.2.1	Notion de processus ponctuels	55
3.2.2	Statistiques descriptives fonctionnelles	56
3.2.3	Autres fonctions plus génériques dans la littérature	61
3.2.4	Effet de bord et correction	62
3.2.5	Test de Monte Carlo	64
3.3	Méthode proposée pour la description de processus spatiaux	66
3.3.1	Données ponctuelles	66
3.3.2	Données mixtes	68
3.3.3	Regrouper les statistiques de plusieurs sources	71
3.4	Application aux images histologiques	73
3.4.1	Statistiques pour les processus ponctuels	73
3.4.2	Statistiques pour les données mixtes	77
3.4.3	Agrégation des fonctions issues de plusieurs images	80
3.5	Conclusion et perspectives	85
4	Analyse de la texture dans les images biomédicales	87
4.1	Motivation	88
4.2	État de l'art de l'analyse de la texture	90
4.2.1	Statistiques du premier ordre	90
4.2.2	Statistiques du second ordre	90
4.2.3	Statistiques d'ordre supérieur	91
4.2.4	Caractéristiques des filtres de Gabor	92
4.2.5	Dimension fractale	94
4.2.6	Caractéristiques sur la transformée en ondelettes	94
4.3	Extraction de l'information de texture par motifs binaires locaux	96
4.3.1	Motifs binaires locaux	96
4.3.2	Analyse multi-échelles	100
4.4	Applications	103
4.4.1	Paramètres et méthodes de comparaison	103
4.4.2	Aide au diagnostic du glaucome sur photographies de rétine	105
4.4.3	Applications aux images histologiques du <i>Challenge ICPR 2014 : Nuclear Atypia</i>	111
4.4.4	Application aux données histologiques de modèles animaux de cancer colorectal	116
4.5	Conclusion et perspectives	119
5	Prise de décision et interprétation des mesures	121
5.1	Motivation	122
5.2	Statistiques pour l'interprétation des données	124
5.2.1	Statistiques descriptives	124
5.2.2	Statistiques déductives	125
5.3	Sélection de caractéristiques et classification des données	132
5.4	Applications	134
5.4.1	Densité	134
5.4.2	Corrélation	135
5.4.3	Application de l'ANOVA	135
5.4.4	Sélection de caractéristiques et classification	138
5.5	Conclusion et perspectives	141
6	Conclusion générale et perspectives	143

Table des figures

1.1	Exemples d'images histologiques présentant divers marquages.	7
1.2	Illustration des deux méthodes de marquage immunohisto-chimique.	9
1.3	Images histologiques montrant les variations au sein d'une même maladie pour un même organe.	13
1.4	Variations issues des procédures pré-analytiques.	14
1.5	Vue d'ensemble des différentes étapes de l'analyse histolo-gique assistée par ordinateur.	15
1.6	Exemples d'images histologiques des deux modèles capturées à différents jours après implantation des tumeurs.	20
2.1	Exemples d'images marquées par l'hématoxyline et le DAB illustrant les variations d'apparence des marqueurs.	24
2.2	Exemple de séparation chimique des marqueurs.	25
2.3	Comparaison visuelle des différentes méthodes de séparation du DAB.	29
2.4	Illustration des différents canaux couleurs RGB et modèle de couleur invariant sur une image H & DAB.	35
2.5	Illustration des différents canaux couleurs RGB et modèle de couleur invariant sur une image H & E.	36
2.6	Illustration de la séparation des marqueurs H & DAB proposée.	37
2.7	Illustration de la séparation des marqueurs H & E proposée.	38
2.8	Exemples de la séparation des objets biologiques.	44
2.9	Extraction des noyaux sur une image tronquée.	45
2.10	Illustration des différences dans l'extraction automatique et manuelle des cellules.	46
2.11	Segmentation des objets biologiques sur une image représen-tative teintée par de l'hématoxyline et de l'éosine.	48
3.1	Exemples de description de processus ponctuels avec la fonc-tion \tilde{K} robuste.	66

3.2	Exemple d'un ensemble d'évènements utilisé pour estimer les statistiques spatiales.	74
3.3	Exemple de deux fonctions F et G tracées en fonction de leurs valeurs théoriques de référence avec les enveloppes de Monte Carlo correspondantes	74
3.4	Fonctions F et G mesurées selon les modèles et des jours après implantation.	75
3.5	Fonctions J mesurées selon les modèles et des jours après implantation.	76
3.6	Comparaison entre les enveloppes issues du test de Monte Carlo et leurs valeurs obtenues analytiquement pour la fonction \tilde{K} robuste.	77
3.7	Fonctions \tilde{K} robustes mesurées selon les modèles et les jours après implantation.	78
3.8	Comparaison entre les enveloppes obtenues analytiquement pour la fonction IPA et 100 simulations d'un processus de Poisson indépendant des données surfaciques.	79
3.9	Exemple de l'estimation de la fonction IPA sur deux images représentatives.	81
3.10	Fonctions IPA mesurées selon les modèles et les jours après implantation.	82
3.11	Regroupement des fonctions F par souris et affichage selon les jours et les modèles.	82
3.12	Regroupement des fonctions G par souris et affichage selon les jours et les modèles.	82
3.13	Regroupement des fonctions \tilde{K} par souris et affichage selon les jours et les modèles.	83
3.14	Regroupement des fonctions IPA par souris et affichage selon les jours et les modèles.	84
4.1	Méthode de calcul d'un motif binaire local avec un voisinage circulaire de huit éléments.	97
4.2	Exemples de primitives de textures extraites par les LBP	98
4.3	Exemples de motifs regroupés en appliquant l'invariance en rotation.	98
4.4	Ensemble des dix motifs uniformes deux et invariants en rotation dans le cas d'un voisinage de huit points.	99
4.5	Exemples d'yeux sains et glaucomateux.	106
4.6	Illustration d'une image glaucomateuse en couleur et sans le canal rouge.	107
4.7	Images extraites de la base de photographies de fonds d'œil.	108

4.8	Détection des résolutions optimales d'image suivant les deux méthodes proposées.	109
4.9	Exemples d'images illustrant l'atypie nucléaire.	112
4.10	Exemples d'images illustrant les variations dans la base du concours ICPR 2014.	113
4.11	Distribution des grades dans les données d'apprentissage et de test du concours ICPR 2014.	114
4.12	Taux de reconnaissance en fonction de la résolution de l'image sur l'application ICPR 2014.	114
4.13	Illustration de l'impact du changement de résolution sur les données du concours ICPR 2014.	115
4.14	Taux de reconnaissance en fonction de la résolution de l'image sur les images de cancer colorectal.	118
5.1	Diagramme pour sélectionner un test statistique pour la comparaison de groupes dans le cas de données continues. .	130
5.2	Illustration de la variabilité de la structure des vascularisations.	134
5.3	<i>Boxplots</i> des densités des noyaux et des vascularisations pour chaque modèle et chaque jour après implantation. . .	135
5.4	<i>Scatterplots</i> des densités des cellules et des vascularisations pour les deux modèles d'implantation pour chaque jour. . .	136

Liste des tableaux

1.1	Exemples de caractéristiques de dimensions et de formes couramment utilisées en analyse d’images histologiques. . .	16
2.1	Comparaison entre le nombre de noyaux extraits automatiquement et manuellement par un expert.	46
3.1	Caractéristiques de relations spatiales couramment extraites des structures de graphes et des arbres.	54
3.2	Tableau répertoriant les différentes fonctions descriptives. .	60
4.1	Caractéristiques usuelles calculées sur les matrices de co-occurrences.	91
4.2	Caractéristiques usuelles calculées sur une matrice de longueurs de plage.	93
4.3	Performances des méthodes d’analyse de la texture pour la détection du glaucome.	110
4.4	Influence du choix des opérateurs de <i>CLBP</i>	111
4.5	Comparaison des performances de la méthode proposée avec les approches classiques de l’analyse de la texture sur les données du concours ICPR 2014.	113
4.6	Taux de reconnaissance de chaque lame sur l’application ICPR 2014.	116
4.7	Comparaison des performances de la méthode proposée avec les approches classiques de l’analyse de la texture sur les données de SPIRIT.	117
5.1	Exemples de méthodes de classification et de sélection de caractéristiques à partir d’images histologiques.	123
5.2	Erreurs de types 1 et 2 ainsi que leurs relations avec le degré de signification α et la puissance du test $1 - \beta$	126

5.3	Coefficients de corrélation de Spearman entre les densités des noyaux et des vaisseaux ainsi que les valeurs p correspondantes.	136
5.4	Valeurs p obtenues lors des tests de H_0 pour les différentes caractéristiques suivant les deux facteurs et leur interaction.	138
5.5	Valeurs p corrigées obtenues lors des test deux à deux de H_0 pour les différentes caractéristiques.	139
5.6	Taux de reconnaissance des modèles suivant les caractéristiques employées.	140

Introduction

L'objectif principal de cette thèse est le développement et la discussion de techniques pour l'analyse et l'interprétation des images histologiques. Bien que ce domaine a vu un fort intérêt de la part de la communauté scientifique et le développement de nombreuses méthodes, beaucoup de défis restent à relever et de nombreux champs de recherche doivent encore être explorés. Ce chapitre introduit tout d'abord brièvement les concepts de l'histopathologie dans les parties 1.2, 1.3 et 1.4, allant de la préparation des échantillons à l'état de l'art de l'analyse des images histologiques, ce qui est nécessaire afin de comprendre ce domaine particulier. Un ensemble de données histologiques mis à disposition dans le cadre du projet ANR SPIRIT est également introduit dans la section 1.5. Finalement, la partie 1.6 présente une vue d'ensemble de la structure des recherches effectuées lors de cette thèse accompagnée d'un résumé des principales contributions qui en découlent.

1.1 Motivation

L’histologie traite de l’étude des structures microscopiques des cellules et des tissus des organismes vivants. La connaissance de ces structures biologiques et de leurs fonctions aux échelles des organes, des tissus, des cellules et des molécules est primordiale pour l’étude de la prolifération et de l’évolution des maladies. La pathologie numérique a fait passer la microscopie optique traditionnelle dans l’ère numérique en parallèle avec le développement des capacités de calculs informatiques.

Au cours de la dernière décennie, la pathologie numérique a rendu possible l’entreposage virtuel et le partage rapide de données biologiques tout en permettant aux pathologistes de tirer un diagnostic à partir d’un large ensemble de mesures quantitatives en plus de leurs observations visuelles. Les lames histologiques offrent une visualisation unique des maladies et de leurs effets sur les tissus et les cellules puisque leur procédé d’acquisition préserve l’architecture du tissu. Pour cela, l’analyse quantitative de ces données est nécessaire non seulement pour assister au diagnostic mais également afin de comprendre les raisons sous-jacentes justifiant un diagnostic et pour comprendre les mécanismes biologiques liés aux maladies.

Grâce à l’avancée des travaux de recherche effectués pour l’analyse de ces images particulières, de nombreux logiciels spécialisés¹ ont pu être développés. Malgré l’attention reçue actuellement par les images histologiques, de nombreux axes de recherches restent encore à explorer afin de tirer des connaissances de toutes les informations pertinentes qu’elles contiennent. La plupart des applications intègrent les mêmes étapes d’analyse, ce qui forme un schéma d’étude caractéristique. Ce cheminement passe par une étape de segmentation afin d’identifier des régions ou des objets d’intérêts, suivie par une quantification sous la forme de l’extraction de caractéristiques. Une grande majorité des méthodes proposées cherchent à reproduire les analyses faites par les experts afin d’établir un diagnostic. De ce fait, ces méthodes ne permettent pas d’extraire des connaissances dans les cas où les pathologistes n’y parviennent pas eux-mêmes.

1. Voici quelques exemples ainsi que les sites internet associés :

- Fiji (<http://fiji.sc/Fiji>)
- Cell Profiler (<http://www.cellprofiler.org/>)
- SigmaScan (<http://www.sigmaaldrich.com/catalog/product/sigma/z361747?lang=fr®ion=FR>)
- ePathology (<http://www.leicabiosystems.com/pathology-imaging/epathology/>)
- HistoQuest (<http://www.tissuegnostics.com/EN/software/histoquest.php>)

Les travaux réalisés au cours de cette thèse visent à proposer de nouveaux outils exploitant l'information issue des images histologiques pour aider à comprendre les maladies et améliorer leurs diagnostics. De nombreuses informations ne sont actuellement pas utilisées, en particulier les relations spatiales entre des objets de natures différentes, il est donc intéressant de mesurer leur influence et leur impact dans ce domaine spécifique.

1.2 L’histopathologie et son intérêt

Le terme histopathologie est composé de trois mots grecs désignant le tissu (*histos*), la maladie ou la souffrance (*pathos*) et l’expression ou les oracles (*logia*). Les travaux présentés dans cette thèse traitent de l’analyse de données histopathologiques et donc de l’examen microscopique de tissus biologiques afin d’étudier les manifestations d’une maladie. A l’opposé de la cytologie, qui réfère à l’observation des cellules isolées, l’objectif est ici d’analyser les interactions entre les différentes cellules et les autres éléments constituant le tissu.

Les organes du corps sont composés de quatre tissus de base² :

Épithélial tissus composés de cellules jointives, ils servent à séparer les tissus conjonctifs de l’extérieur, des cavités et des canalisations.

Conjonctif tissus composés de cellules séparées par de la matrice extracellulaire qui offrent un soutien structurel et métabolique pour les autres tissus et organes.

Muscle tissus formés de cellules pouvant se contracter afin de permettre le mouvement.

Nerveux tissus formés de cellules organisées en réseau pour contrôler l’organisme.

Tous les tissus sont formés de cellules, de matrice extracellulaire, et de fluides corporels. La matrice extracellulaire est fabriquées par les cellules et sa composition dépend de la fonction du tissu, pouvant former par exemple des os ou des tendons.

Il est important de bien reconnaître les différents types de tissus et leurs composants afin de bien comprendre les différents éléments constituant l’échantillon observé et contribuant à sa fonction.

La recherche médicale et biologique est étayée par la connaissance de la structure et de la fonction normales des cellules et des tissus ainsi que des organes qu’ils composent. Les cellules et tissus sains sont arrangés suivant des formes régulières et connues. En revanche, les maladies sont révélées par des modifications de la structure à une échelle microscopique et la plupart montrent des déformations caractéristiques. L’identification de ces altérations et leur association à des maladies particulières représentent la base de l’histopathologie et de la cytopathologie.

2. Ces informations proviennent du site : <http://www.histology.leeds.ac.uk/index.php> © Faculty of Biological Sciences, University of Leeds

1.3 Procédures pré-analytiques

Cette première étape nécessaire intègre la préparation des échantillons de tissus et l'acquisition des images. La forme de microscopie la plus couramment employée pour l'analyse histopathologique est la microscopie en lumière visible où un tissu est traversé par de la lumière. Pour être utilisée, un échantillon doit avoir été préparé de sorte à respecter quatre conditions nécessaires :

- Les cellules et autres éléments doivent être préservés dans un état proche du vivant.
- L'échantillon doit être transparent pour que la lumière puisse le traverser.
- L'échantillon doit être fin et plat de sorte qu'une unique couche de cellules soit présente.
- Certains composants doivent être teintés avec des marqueurs différents afin d'être facilement distingués.

D'autres modalités d'imageries microscopiques peuvent être utilisées, notamment la fluorescence, l'immunofluorescence et les microscopies confocale, multi-spectrale et électronique. Pour plus de détails sur ces modalités, il convient de se référer aux travaux de [He et al. \[2010\]](#) et de [Holmes et Cheng \[2006\]](#). La suite de ce chapitre traite uniquement de la microscopie en lumière visible.

1.3.1 Introduction à la préparation des échantillons

Afin de suivre les conditions imposées par la microscopie, les échantillons peuvent être préparés selon quatre méthodes d'après [Rolls \[2011\]](#) :

Whole-mount quand l'organisme ou la structure est suffisamment petit ou fin pour être placé directement sur une lame microscopique.

«**Écrasée**» quand les cellules sont concassées sur une lame afin de révéler leurs contenus.

Prélèvements quand l'échantillon est constitué de cellules en suspension dans un liquide, ou quand des cellules individuelles ont été prélevées depuis une surface ou un organe.

Section quand l'échantillon est préparé dans le but de découper de fines tranches, de les placer sur les lames et de les marquer.

Seules les méthodes *whole-mount* et sections préservent l'organisation du tissu. La préparation des sections est la plus utilisée mais également la plus complexe à réaliser puisqu'elle demande des équipements spécialisés et une expertise spécifique. Cette méthode est donc détaillée ci-dessous.

Préparation des sections histologiques

Trois types de préparation des sections peuvent être réalisées pour la microscopie en lumière visible : l'inclusion dans la paraffine, les sections congelées et les sections semi-fines (Winsor [1994]).

Inclusion dans la paraffine. La technique de la paraffine est la plus couramment employée et consiste à fixer les tissus avant de les intégrer dans de la cire. La fixation consiste à déshydrater et à ajouter des produits chimiques principalement afin de désactiver certaines enzymes préservant ainsi le tissu avec un aspect le plus proche possible du vivant. Finalement, de la cire de paraffine est utilisée pour prendre la place de l'eau dans le tissu. En refroidissant, le tissu devient solide et peut être découpé en fines sections avec un microtome, allant de 2 à 25 μm d'épaisseur. Les sections sont ensuite placées sur des lames microscopiques et prêtes à être marquées.

Sections congelées. Les tissus sont congelés rapidement avec de l'azote liquide avant d'être découpés dans un appareil réfrigéré (un cryostat) par un microtome à lames froides. Les sections sont alors prêtes à être placées sur une lame et marquées. Cette procédure est plus rapide et permet de préserver certains détails du tissu qui peuvent être perdus avec la technique de la paraffine mais peut compromettre son organisation. Les sections ainsi obtenues peuvent être de 5 à 10 μm d'épaisseur.

Sections semi-fines. Les méthodes précédentes produisent des sections parfois trop épaisses pour permettre d'observer certains détails. Pour contourner ce problème, les échantillons peuvent être inclus dans de l'époxy ou de la résine d'acrylique, ce qui permet de découper des sections plus fines que 2 μm .

1.3.2 Marquage du tissu

À l'exception de quelques pigments naturels tels que la mélanine, les cellules ainsi que les autres éléments constituant les échantillons de tissus sont dénués de couleurs. Afin de révéler les détails structurels en utilisant la microscopie en lumière visible, il est nécessaire d'utiliser un marquage. Les différentes techniques peuvent être spécifiques, au sens où elles ne colorent que des groupes chimiques particuliers ou des molécules précises à l'intérieur des cellules ou du tissu, et non-spécifiques, quand elles teignent la plupart des cellules de la même façon. Le marquage est généralement fait

en utilisant un marqueur teintant quelques composants des cellules d'une couleur claire conjointement avec une contre-teinte marquant le reste des cellules d'une couleur différente (Winsor [1994]). Des exemples de plusieurs marqueurs sont montrés dans la figure 1.1.

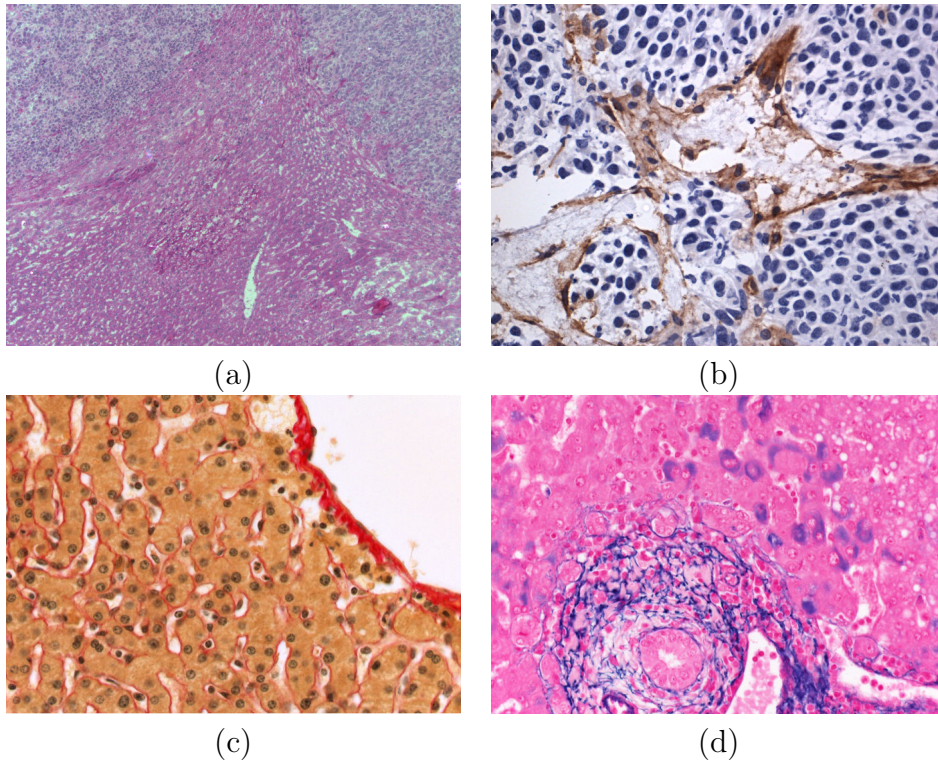


FIGURE 1.1 – Exemples d'images histologiques présentant divers marquages. (a) Une image marquée par l'hématoxyline, qui teinte les noyaux en bleu-violet, et de l'éosine, qui révèle les cytoplasmes en rose. (b) Une image avec du DAB marron, pour les vascularisations, et de l'hématoxyline, pour les noyaux. (c) Une image avec marquage trichromatique nommé rouge picrosirius, avec de l'hématoxyline pour les noyaux, de l'acide picric jaune pour les cytoplasmes et les noyaux, et du rouge sirius ciblant les vascularisations et le collagène. (d) Un marquage bleu victoria révélant les hépatocytes avec contre-teinte rouge. Les images (a) et (b) proviennent des données associées au projet ANR SPIRIT alors que (c) et (d) sont issues de l'article de Tadrous [2010].

Procédure de marquage

La plupart des produits permettant de marquer les tissus sont aqueux, ce qui impose donc de dissoudre la cire, le cas échéant, et de la rempla-

cer par de l'eau. Les lames sont baignées dans du xylène, puis dans des concentrations de plus en plus faibles d'éthanol, allant de 100 % à de l'eau pure. Une fois la teinte apposée puis rincée, l'échantillon est de nouveau déshydraté et baigné dans du xylène.

Hématoxyline et éosine

Le marquage de routine généralement employé afin de fournir des informations essentielles quant à la structure du tissu se base sur l'hématoxyline et l'éosine (H&E). L'éosine est un marqueur acide qui colore donc les structures basiques en rose. L'hématoxyline est en revanche basique et teinte donc les objets acides dans le tissu en bleu-violet.

L'ADN dans les noyaux, l'ARN dans les ribosomes et le réticulum endoplasmique rugueux, qui permet de régénérer les membranes, sont acides et donc marqués par l'hématoxyline. Quelques matériaux extracellulaires, tels que ceux présents dans le cartilage, sont également ciblés. À l'opposé, la plupart des protéines dans le cytoplasme sont basiques et donc teintées par l'éosine en rose, cela inclut les filaments cytoplasmiques dans les cellules de muscle, les membranes intracellulaires et les fibres extracellulaires. Un exemple d'un tel marquage est fourni en figure 1.1 (a).

En histopathologie, de nombreuses maladies peuvent être diagnostiquées en examinant des lames H&E. Cependant, il arrive que des informations supplémentaires soient nécessaires afin de produire un diagnostic fiable et cela implique d'utiliser des méthodes de coloration plus spécialisées. Elles peuvent être obtenues par l'utilisation de marqueurs spéciaux se basant sur des teintures ou des apports de métaux afin de définir des structures ou des micro-organismes particuliers.

Immunohistochimie (IHC)

Les marquages immunohistochimiques servent à localiser des antigènes, donc des protéines, dans les tissus. Contrairement aux marqueurs usuels, cette localisation exploite le principe des anticorps qui se lient spécifiquement à certains antigènes (Ramos-Vara [2005]). Le tissu est finalement marqué par les anticorps étiquetés. Ces marqueurs sont très utilisés afin de diagnostiquer les cellules anormales telles que celles présentes dans les tumeurs. Des marqueurs moléculaires spécifiques sont caractéristiques d'événements cellulaires particuliers comme la prolifération ou l'apoptose, qui est la mort programmée des cellules. D'après Ramos-Vara [2005], l'IHC est également utilisée en recherche dans le but de comprendre la distribution

et la position de biomarqueurs et de protéines dans différentes parties d'un tissu.

Les marquages IHC peuvent être de types direct ou indirect ([Ramos-Vara \[2005\]](#)). La méthode directe est une détection en une étape qui se sert d'un marqueur visuel couplé à un anticorps capable de se lier directement avec l'antigène recherché. La méthode indirecte est réalisée en deux étapes. Premièrement, un anticorps primaire sans étiquette est lié à l'antigène ciblé. Après rinçage de l'excès, un anticorps secondaire avec marqueur visuel est ajouté afin de se lier au premier. La méthode directe est simple et rapide tout en étant très sensible ([Kårsnäs \[2014\]](#)) alors que la méthode indirecte permet d'être encore plus sensible puisque chaque anticorps primaire peut être lié à plusieurs secondaires. Les deux approches sont illustrées dans la figure 1.2.

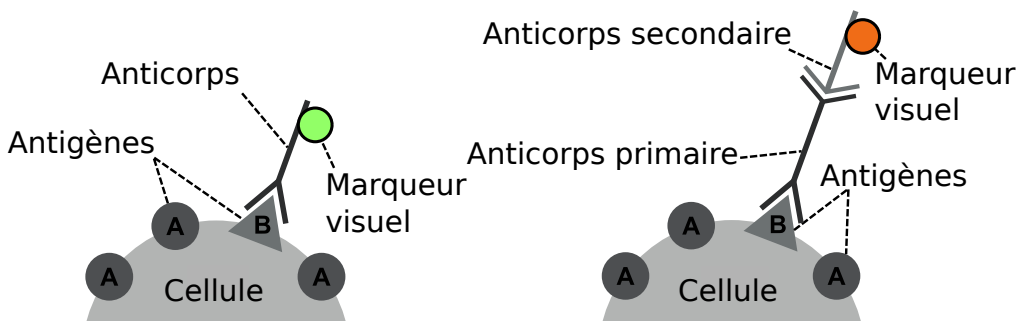


FIGURE 1.2 – Illustration des deux méthodes de marquage immunohisto-chimique. A gauche la méthode directe où un anticorps marqué est direc-tement lié à un antigène de l'objet ciblé. A droite la méthode indirecte où un anticorps primaire est lié à l'antigène ciblé puis un anticorps secondaire avec marqueur visuel est lié au premier.

Le marqueur visuel le plus couramment employé en IHC est le DAB (3,3'-Diaminobenzidine) qui produit un marquage marron intense, voir figure 1.1 (b). Optionnellement, un contre marquage peut être réalisé avec une teinte contrastante afin de souligner les composants du tissus non marqués par le marqueur principal. Généralement, ce contre marquage est réalisé avec de l'hématoxyline.

Pour plus d'informations sur l'IHC, il est possible de se référer à l'article de [Di Cataldo et al. \[2012\]](#).

1.3.3 Stratégies d'acquisition des images en histologie

Il existe plusieurs modalités permettant de capturer les images histologiques par microscopie en lumière visible, notamment les microscopes avec caméras et les scanners de lames complètes. Une dernière stratégie consiste à utiliser des micro-vecteurs de tissus afin de limiter les variations.

Microscopes avec caméras

Les images histologiques sont classiquement obtenues en montant une caméra numérique sur un microscope optique traditionnel. Deux approches différentes existent pour ce qui est de la sélection des régions à numériser. Dans le premier cas, seules les régions d'intérêts, nommées *hot spot*, sont photographiées puisqu'elles sont supposées contenir le plus d'informations pertinentes. A l'opposé, l'autre méthode consiste à prendre en photo des régions choisies aléatoirement sur la lame dans le but d'avoir les images les plus représentatives et les moins subjectives possibles.

Scanners de lames complètes

Une récente avancée permet de numériser les lames histologiques complètes de façon automatique. C'est un outil robuste et performant qui prend un rôle important dans de nombreux laboratoires d'étude des pathologies.

Deux types principaux de scanners ont été développés, les scanners par zones et les scanners en lignes. Les scanners par zones capturent des sous-images microscopiques puis les alignent afin de fournir une image complète de la lame. A l'opposé, les scanners en lignes utilisent des capteurs linéaires combinés avec un système mécanique permettant de déplacer la lame en continu durant l'acquisition, limitant ainsi le nombre d'images à coller les unes aux autres. Ces deux techniques sont en cours de développement afin d'améliorer la vitesse d'acquisition, la qualité de l'image et les méthodes de marquages supportées.

Les lames numériques ainsi obtenues sont construites d'images de diverses résolutions organisées de façon pyramidale. Ces images permettent donc aux pathologistes de naviguer dans la lame et de changer de grossissement de façon similaire à ce qui est possible avec un microscope.

Micro-vecteurs de tissus

La préparation des échantillons est une étape laborieuse et sujette à de nombreuses variations qui ressortent dans les images finalement générées. La miniaturisation et les approches à forte cadence de production ont été employées pour formaliser les micro-vecteurs de tissus dans le but de limiter au mieux ces nombreuses variabilités. Cette méthode permet de représenter des centaines de petits échantillons dans un unique bloc de cire de paraffine. Ainsi, tous les échantillons subissent exactement le même protocole et sont ensuite découpés et marqués simultanément.

Les principaux avantages de cette technique sont l'homogénéisation des procédures, la rapidité de production et l'économie de produits chimiques. En revanche, le petit diamètre des échantillons fait qu'ils ne sont potentiellement pas représentatifs des tissus d'origine. Pour éviter cela, le vecteur doit être conçu avec un soin particulier d'après [Packeisen et al. \[2003\]](#).

1.4 Procédures analytiques et interprétatives

1.4.1 Analyse manuelle

Le *gold standard* de l'analyse histopathologique pour l'établissement d'un diagnostic est l'observation visuelle de lames issues de plusieurs biopsies par un pathologiste expert. L'examen est basé sur la détection de motifs, de formes ou de dimensions des structures pertinentes présentes dans le tissu ou d'autres caractéristiques propres à certaines maladies. L'expert se doit donc de pouvoir différencier les tissus sains et malades et de repérer les plus subtiles variations afin de fournir un diagnostic fiable.

Du fait du nombre usuel de patients, des dimensions des lames histologiques et du nombre de cellules, l'évaluation visuelle effectuée par un expert est nécessairement longue et fastidieuse. De plus, cet examen se base sur des relevés qualitatifs, faisant de ce *gold standard* un procédé subjectif enclin aux erreurs et sujet aux variations inter- et intra-experts ([Al-Kofahi et al. \[2011\]](#); [Rizzardi et al. \[2012\]](#)). De ce fait, malgré les efforts pour homogénéiser l'évaluation histopathologique et réduire les différentes sources de variations, [Stenkvist et al. \[1983\]](#) et [Jain et al. \[2011\]](#) soulignent que l'examen par un pathologiste expert reste une méthode non-optimale.

1.4.2 Analyse assistée par ordinateur

Objectifs

En raison de la demande grandissante et irréalisable pour les experts afin d'interpréter les images histologiques, les systèmes de diagnostics assistés par ordinateur deviennent de plus en plus cruciaux. L'informatisation de cette analyse permet d'extraire des mesures quantitatives afin de fournir des indications objectives en se basant sur des informations difficilement accessibles par un examen visuel. Contrairement aux méthodes manuelles, les systèmes informatiques offrent des performances rapides et invariantes dans le temps, permettant donc d'obtenir des résultats reproductibles.

De plus, l'utilisation d'ordinateurs ouvre la possibilité de développer la télépathologie qui permet de se soustraire du besoin d'avoir un pathologiste dans la région, économisant ainsi du temps et offrant des possibilités de diagnostic dans les zones reculées. Une autre possibilité apportée par les mesures quantitatives est de permettre la caractérisation objective des structures du tissu, ce qui est une étape nécessaire dans le cadre de la recherche biologique pour comprendre les mécanismes de croissance et de propagation des maladies.

Principaux défis

Variations naturellement présentes dans les tissus. Par nature, tous les organes sont composés de tissus différents dépendant de leurs fonctions. Il n'est donc pas envisageable qu'une méthode automatique universelle soit développée pour analyser toutes les maladies ciblant tous les organes. De plus, deux échantillons extraits d'un même organe sont susceptibles de présenter des structures différentes tel qu'illustré dans la figure 1.3. De ce fait, la standardisation de la localisation des biopsies est une nécessité. La présence de maladies affectant les tissus fait que les différents objets biologiques ne respectent plus les structures et morphologies normales. Cependant, ces modifications dépendent de chaque individu. Les méthodes d'analyse se doivent donc d'essayer de prendre en compte toutes ces sources de variations pour pouvoir fournir des diagnostics fiables.

Variations liées aux procédures pré-analytiques. En plus des variations issues de la biologie, chacune des étapes de la préparation des échantillons est une source possible de variations se retrouvant dans les images à analyser. Les variations issues de chaque étape ainsi que leurs impacts sont discutés par [Diniz \[2010\]](#). Plus d'informations sur les différences liées

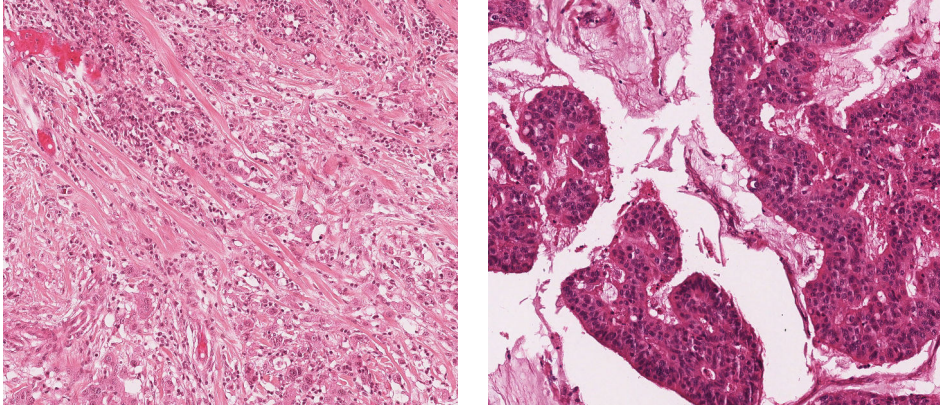


FIGURE 1.3 – Images histologiques montrant les variations au sein d’une même maladie pour un même organe. Ces deux images sont issues du *Challenge ICPR 2014 : nuclear atypia* et présentent des tissus de sein humain atteint d’un cancer de grade 2 (modéré). Elles soulignent les différences de structures pouvant être rencontrées.

au marquage des lames sont données dans la partie 2.1. D’autres variations peuvent surgir lors de l’acquisition de l’image. Elles sont généralement liées soit au microscope utilisé, chaque appareil ayant ses spécificités, soit aux réglages effectués par l’utilisateur. Certaines de ces variations issues des procédures pré-analytiques sont illustrées dans la figure 1.4.

Problème de la vérité terrain. Un autre défi important lié à l’analyse d’images histologiques concerne le manque de vérités terrains permettant l’identification et la classification des tissus. Ce manque et le fait que le *gold standard* est réalisé par des pathologistes entraînent que les méthodes sont subjectives puisque sujettes à l’appréciation d’un nombre réduit de pathologistes. Les algorithmes développés peuvent donc au mieux faire comme le pathologiste et les objectifs des méthodes informatiques, en particulier l’objectivité et la reproductibilité, ne sont pas évalués et la méthode échoue quand l’expert échoue également.

1.4.3 État de l’art de l’analyse assistée par ordinateur

Les méthodes d’analyse d’images histologiques exploitent généralement une première étape de prétraitements, qui inclut tout d’abord un contrôle de la qualité des données, avec correction le cas échéant puis de la segmentation si nécessaire. La deuxième étape concerne l’extraction de l’information. Les différentes techniques peuvent être regroupées selon trois

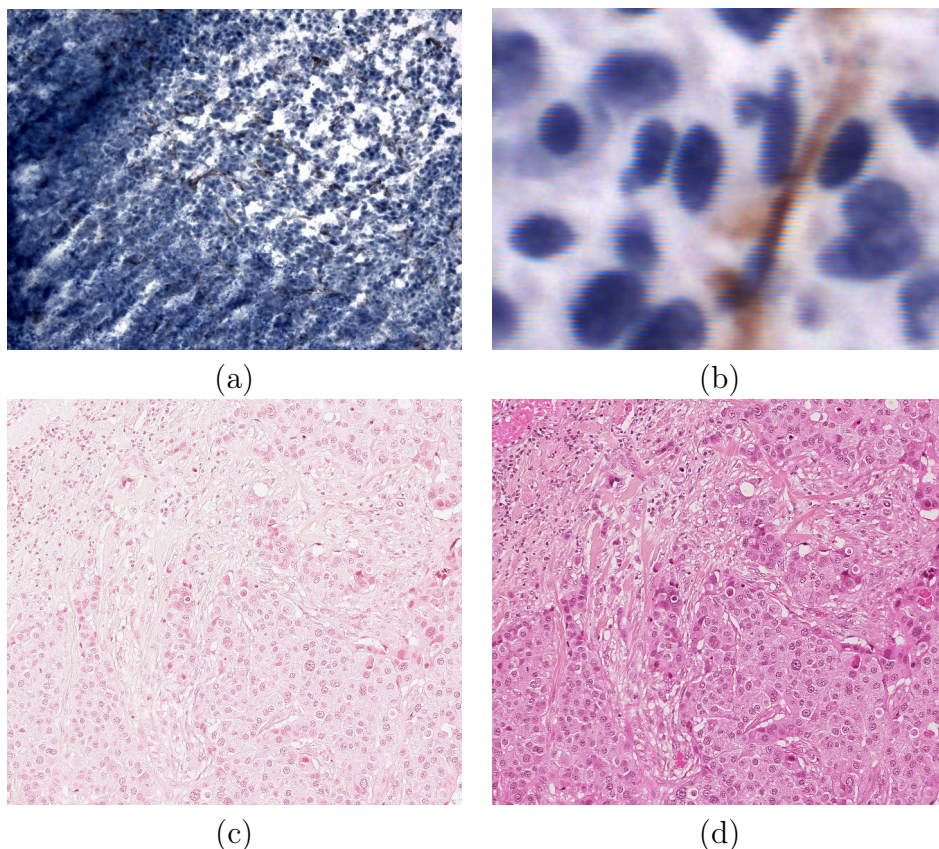


FIGURE 1.4 – Variations issues des procédures pré-analytiques. (a) présente des problèmes liés à l'absorption des marqueurs et des traces de découpe. (b) montre des défauts issus de l'acquisition de l'image. (c) et (d) montrent la même région d'intérêt d'un même échantillon capturée par deux microscopes différents, un Aperio pour (c) et un Hamamatsu pour (d).

échelles de capture : celle des pixels, celle des objets et celle de la sémantique. Finalement, un test statistique ou une étape de classification permet de diagnostiquer l'échantillon. Ces trois étapes sont résumées dans le diagramme de la figure 1.5.

Pré-traitements

La notion de la qualité d'une image n'est pas triviale. En effet, la visualisation d'une image induit, en plus de ses propriétés optiques, un ressenti subjectif qui n'est pas entièrement lié à ses caractéristiques physiques. Ainsi, la capacité d'extraire des informations d'une image dépend également de données liées à son contenu sémantique. Cette étape est primordiale mais peu de travaux dans la littérature l'abordent alors qu'il est

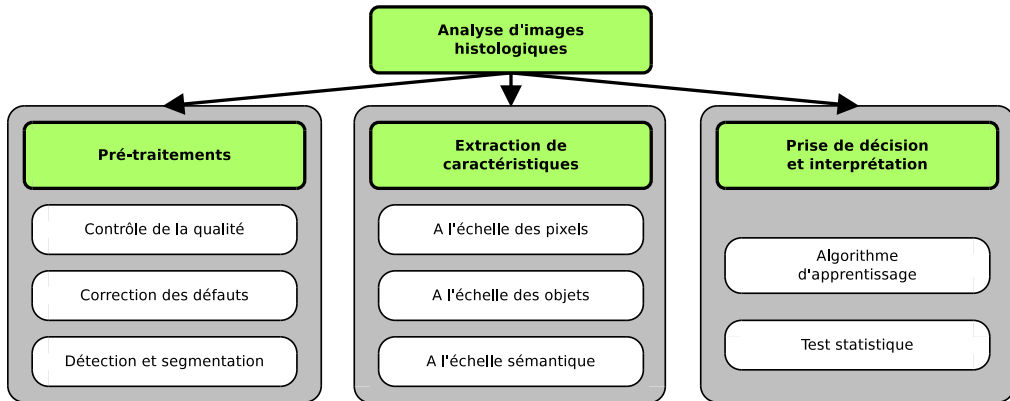


FIGURE 1.5 – Vue d’ensemble des différentes étapes de l’analyse histologique assistée par ordinateur. Bien qu’aucune méthode universelle n’existe, la plupart des approches suivent ce schéma d’analyse.

nécessaire de s’assurer que l’image renferme suffisamment d’informations afin de permettre l’établissement d’un diagnostic.

Selon [Kayser et al. \[2008\]](#), la qualité d’une image peut être estimée par des opérations de filtrage locales afin de détecter des variations d’illuminations, associées à l’analyse du rapport signal sur bruit, des gradients maximums, de la distribution des niveaux de gris et des extrema d’intensité et de teinte. Une autre approche proposée par [Wang et al. \[2011\]](#) consiste à exploiter une carte de similarité des couleurs. Les deux méthodes se basent sur la comparaison avec une image de référence sélectionnée de façon subjective.

La correction des défauts présents dans les images est réalisée avec les méthodes classiques du traitement de l’image. La technique la plus souvent utilisée reste cependant la normalisation des couleurs afin de réduire les variations issues du marquage. Une étude de cette approche ainsi qu’un état de l’art des différentes techniques sont fournis par [Khan et al. \[2014\]](#). Une discussion et analyse des méthodes de détection et de segmentation des objets dans les images histologiques est présentée dans la partie 2.2.

Extraction de caractéristiques

À l’échelle des pixels. Les caractéristiques extraites à partir des pixels ne se concentrent pas sur des groupes de pixels particuliers mais sur la totalité de l’image en capturant des propriétés telles que la texture ou la couleur. La texture est utilisée afin de segmenter plusieurs objets ou types de tissus, par exemple les méningiomes par [Al-Kadi \[2010\]](#) ou encore la

séparation de l'épithélium tumoral du stroma par [Linder et al. \[2012\]](#). Une autre utilisation par [Mahmoud-Ghoneim \[2011\]](#) de l'estimation de la texture est de caractériser la fibrose du foie. Également, [Sertel et al. \[2009\]](#) exploitent la quantification couleur non linéaire et les cartes de Kohonen pour évaluer le grade des lymphomes folliculaires. Les approches permettant de mesurer la texture d'une image sont discutées dans le chapitre 4. L'information de couleur mesurée à partir de huit canaux différent est utilisée afin de classer les mélanomes de la peau par [Celebi et al. \[2007\]](#). Bien que ces mesures ne sont pas aisément interprétables biologiquement, elles sont très utilisées dans la littérature en raison de leur simplicité d'extraction et de leurs performances.

À l'échelle des objets. L'analyse au niveau des objets dépend fortement d'une première étape de segmentation qui permet de définir ce qui représente un objet dans l'image. Un objet correspond généralement à un ensemble de pixels connexes réunis par un critère de similarité, tels que les noyaux marqués par l'hématoxyline.

Les caractéristiques ainsi extraites peuvent être définies soit objet par objet, en observant leurs dimensions, formes ou encore leurs textures, soit en mesurant les relations spatiales entre ces objets. Les caractéristiques les plus courantes sont regroupées dans le tableau 1.1. Les mesures spatiales sont généralement capturées sous la forme de graphes à partir des objets ou de fonctions statistiques décrivant leurs distributions en les considérant comme des processus ponctuels; ces méthodes sont décrites dans le chapitre 3. Les mesures de texture basées sur les objets sont extraites uniquement sur les groupes de pixels composant les objets. Plus d'exemples et de détails sur les caractéristiques de forme, de dimensions et de texture peuvent être trouvés dans le chapitre 5 de la thèse de [Boucheron \[2008\]](#).

TABLE 1.1 – Exemples de caractéristiques de dimensions et de formes couramment utilisées en analyse d'images histologiques.

Catégorie	Caractéristiques
Dimensions	Surface, périmètre et rayon équivalent
	Longueurs des demi-grand et demi-petit axes elliptiques
Formes	Surface convexe et solidité
	Orientation et excentricité
	Diamètre équivalent, compacité et forme d'inertie

Des caractéristiques de formes sont exploitées par [Kothari et al. \[2013b\]](#) dans le cadre d'une recherche exploratoire afin de reconnaître les différents

sous-types de tumeurs rénales. De même, la texture des noyaux a été reconnue par [Krishnan et al. \[2012\]](#) comme étant porteuse d'informations afin de discriminer les régions malignes et bénignes.

À l'échelle sémantique. À l'opposé des caractéristiques extraites aux deux échelles précédentes, les mesures sémantiques permettent de capturer des concepts de plus haut niveau facilement interprétables tels que la présence ou l'absence de certains objets ou types de tissus dans les images. Une information sémantique est généralement issue de la classification ou d'une règle statistique d'un sous-ensemble de caractéristiques de plus bas niveau.

Les systèmes d'aide au diagnostic se basent donc souvent sur des méthodes analysant l'espace des caractéristiques. Dans ce contexte, [Cruz-Roa et al. \[2011\]](#) utilisent les sacs de caractéristiques afin d'associer chaque concept sémantique avec des groupes de mesures pour annoter les images. Également, [Caicedo et al. \[2011\]](#) exploitent de nombreuses caractéristiques et une étape de sélection afin d'apposer des concepts sémantiques pour la recherche d'images histologiques dans les grandes bases de données. Le développement de ces méthodes permet de se soustraire de variations issues de la biologie des cancers mais cela demande un nombre important de données d'entraînement annotées par un expert. Selon [Kothari et al. \[2013a\]](#), peu de méthodes se penchent sur ce type de caractéristiques en raison du manque de vérité terrain.

Approches multi-échelles. Les méthodes multi-échelles peuvent être regroupées en deux familles, celles qui capturent les informations à partir d'une même image à plusieurs résolutions, et celles qui exploitent les capacités descriptives des caractéristiques issues des échelles des pixels, des objets et de l'information sémantique.

Une approche multi-résolutions est employée par [Signolle et al. \[2010\]](#) afin de segmenter les différents tissus présents sur des images de lames complètes. Leur méthode analyse la texture des images à l'aide d'arbres cachés de Markov et associe chaque portion de l'image avec un type de tissu. Également, [Qureshi et al. \[2008\]](#) utilisent le regroupement de plusieurs décompositions en paquets d'ondelettes discriminantes et adaptatives afin de prendre en considération les variations au sein d'une même classe de tissu pour reconnaître les différents types de méningiomes.

La combinaison d'informations liées aux pixels et aux objets est réalisée par [Kothari et al. \[2013b\]](#) sous la forme de caractéristiques morphologiques des noyaux et de la texture des images pour différencier les tumeurs des

reins. De même, [Tabesh et al. \[2007\]](#) regroupent les informations de couleurs et de texture de l'image avec les formes et l'organisation des différents objets biologiques identifiables dans le but d'évaluer le cancer de la prostate.

Diagnostic et interprétation

Pour cette dernière étape, un premier groupe d'études utilise des tests statistiques afin de déterminer si des différences significatives existent entre les caractéristiques correspondantes aux différentes classes. En particulier, [Mattfeldt et al. \[2006\]](#) se basent sur des t-tests pour vérifier la capacité discriminative des mesures de relations spatiales extraites de lésions bénignes et malignes.

Un second groupe d'études utilise les algorithmes d'apprentissage automatique afin de former le classifieur à distinguer les différentes classes. Par exemple, [Oztan et al. \[2012\]](#) se servent de SVM, de classifieurs bayésiens et de k plus proches voisins pour reconnaître les différents grades de lymphomes folliculaires à partir de caractéristiques structurelles.

Ces deux catégories d'approches sont traitées et discutées en détails dans le chapitre 5.

1.5 Matériel biologique à disposition

Dans le cadre du projet ANR SPIRIT, un ensemble de données a été fourni par Johanne Séguin et Nathalie Mignet de l'UPCGI afin de tester et valider une partie des méthodes proposées tout au long de cette thèse. D'autres ensembles de données sont également utilisés dans un objectif de validation mais sont décrits uniquement dans la section expérimentale correspondante.

1.5.1 Contexte d'application

Le cancer colorectal attire fortement l'attention des biologistes puisqu'il s'agit d'un des types de cancer les plus répandus au monde. Quant à la mortalité, il occupe la seconde place après le cancer des poumons. Afin d'examiner les effets de thérapies pharmacologiques et anti-cancer, les biologistes réalisent des études pré-cliniques en se basant généralement sur des modèles d'implantation de tumeurs sur les animaux. Dans le cas où les cellules cancéreuses sont injectées dans la même région anatomique que celle d'où elles sont extraites, le modèle est dit orthotopique. À l'inverse, les modèles dits ectopiques supposent l'injection de tumeurs dans une région différente de celle d'origine, généralement sous la peau. Ce second modèle est plus économique et la tumeur est alors plus facile d'accès tout en permettant un examen rapide de son état de façon macroscopique mais ne reflète pas l'environnement d'origine des cellules cancéreuses. En revanche, ces différences font que les deux modèles sont utiles lors des tests pré-cliniques afin de développer de meilleures thérapies pour traiter les maladies (Ho et al. [2012]). En effet, ils permettent d'examiner l'impact des traitements thérapeutiques à la fois sur la tumeur principale (orthotopique) et sur les métastases (ectopique).

Les images histologiques issues des deux modèles sont visuellement très similaires, comme le montre la figure 1.6, bien que des études ont montré que les tumeurs orthotopiques ont une croissance plus rapide que les ectopiques. En effet, une tumeur d'un modèle orthotopique sera deux fois plus grande qu'une ectopique douze jours après implantation (Heldmuth [2012]). Comparer et différencier de telles images est une problématique pleine de défis mais de grand intérêt clinique afin d'interpréter les différences dans leurs mécanismes de croissance.

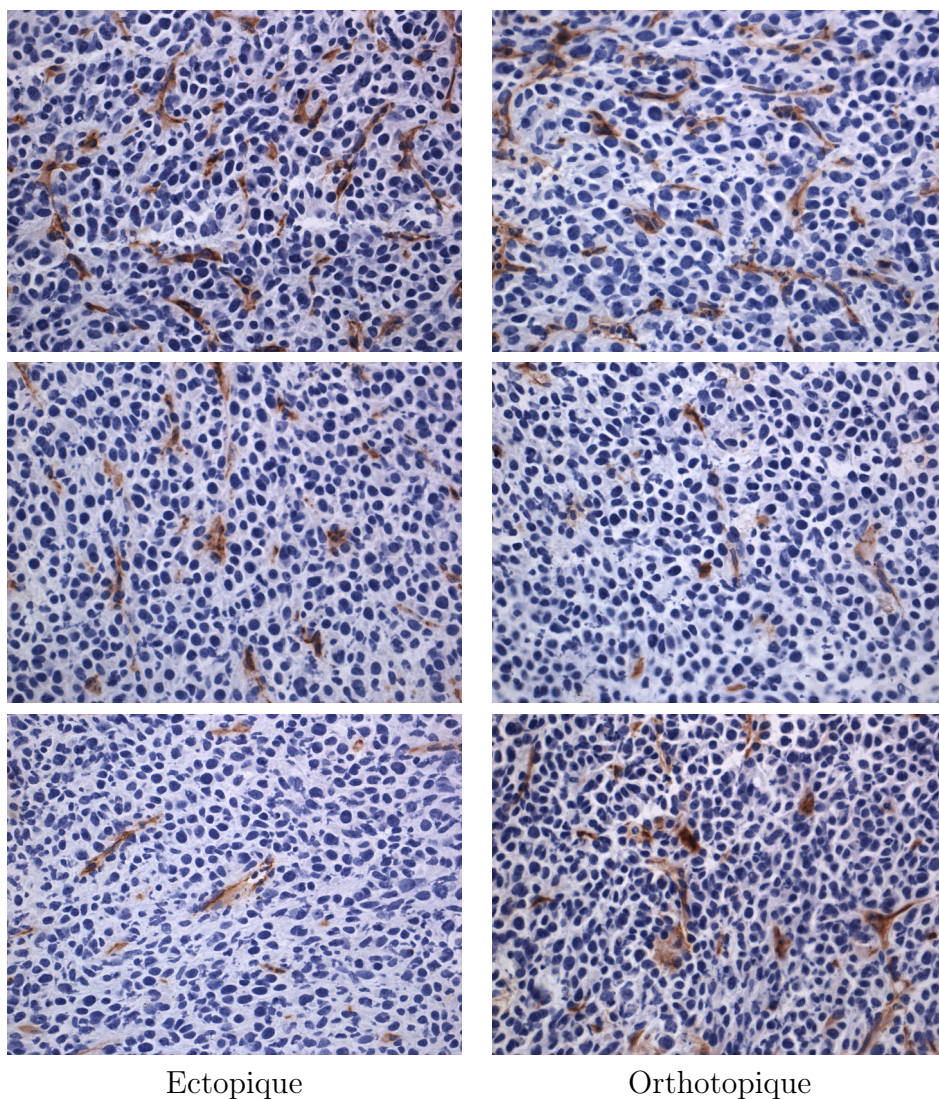


FIGURE 1.6 – Exemples d’images histologiques des deux modèles capturées à différents jours après implantation des tumeurs. Les première, deuxième et dernière lignes montrent respectivement des images issues des jours 11, 15 et 21. Ces images soulignent la forte similarité entre ces deux modèles à l’échelle du tissu.

1.5.2 Matériel

CT26 est une lignée de cellules d’adénocarcinome du colon murin non différenciée qui est syngénique de la souche de souris BALB/c présentée par [Corbett et al. \[1975\]](#). Dans cette étude, de telles tumeurs ont été implantées dans le cæcum de souris BALB/c pour le modèle orthotopique et sous la peau dans le cas ectopique. Plusieurs souris ont été sacrifiées aux jours

11, 15 et 21 après implantation des tumeurs. Dès le décès de la souris, les tumeurs sont retirées, congelées, découpées puis marquées selon un même protocole pour tous les animaux.

Chaque échantillon a été teinté suivant les procédés d'immunohistochimie par du 3,3'-Diaminobenzidine (DAB) afin de localiser les vaisseaux sanguins. Une contre-teinte d'hématoxyline est ensuite apposée afin de rendre visibles les noyaux des cellules. À partir de ce matériel biologique, 18 échantillons provenant de souris différentes ont été conservés, trois par jours et par modèles. Finalement, chaque échantillon est représenté par dix images, menant à la numérisation de 180 images histologiques via un microscope Leica DM6000B. Toutes les images ont une résolution de 2592×1944 pixels.

1.6 Contribution personnelle et organisation de ce manuscrit

La motivation des recherches liées à cette thèse a été introduite conjointement à une présentation du domaine d'étude. Cette partie résume brièvement les contributions scientifiques et l'organisation de la suite de ce manuscrit.

Le chapitre 2 traite de la séparation et de la détection des objets biologiques présents dans les images histologiques. Après présentation des défis associés et des méthodes de la littérature, une méthode simple et rapide permettant de séparer les objets biologiques est tout d'abord proposée. Cette technique s'appuie sur les marqueurs les plus couramment utilisés qui sont l'hématoxyline, l'éosine et le DAB. Une fois les objets séparés, une méthode automatique pour la détection d'objets ponctuels est introduite. Finalement ces approches sont testées et validées sur les données histologiques à disposition.

Les statistiques spatiales ainsi que leurs mises en œuvre dans le cadre des images histologiques sont étudiées dans le chapitre 3. Une description et une discussion des statistiques usuelles sont tout d'abord fournies puis une nouvelle fonction permettant l'analyse des interactions spatiales entre objets de natures différentes est proposée. Une étude expérimentale de toutes ces fonctions sur la base d'images liée à cette thèse est enfin réalisée.

Le chapitre 4 traite des techniques d'analyse de la texture et discute de leurs efficacités pour l'aide au diagnostic. Après un état de l'art des méthodes usuelles, une approche simple d'un point de vue calculatoire et permettant d'extraire la texture aux échelles les plus discriminantes est introduite. Finalement, deux applications issues de domaines médicaux différents sont considérées afin d'évaluer l'approche proposée comparativement aux méthodes classiques de la littérature.

Finalement, le chapitre 5 présente la dernière étape de traitement des images histologiques qui est la prise de décision et l'interprétation des caractéristiques extraites. Cette étude permet d'évaluer le pouvoir discriminant des différentes caractéristiques ainsi que leurs éventuelles corrélations via des tests statistiques ou par apprentissage suivi d'une classification. Ce qui clôt le cheminement classique de l'analyse d'images histologiques.

Détection des objets biologiques dans les images histologiques

La détection des objets biologiques présents dans les images histologiques est une étape fondamentale pour de nombreuses analyses. Les images histologiques sont acquises à partir de lames généralement teintées en utilisant plusieurs marqueurs ciblant différents objets. La séparation de ces marqueurs peut être considérée comme un pré-traitement afin de guider la détection et la segmentation. Dans ce chapitre, un état de l'art des méthodes les plus courantes pour la séparation des marqueurs et pour la segmentation des objets est présenté (partie 2.2). Deux méthodes sont ensuite proposées, une pour la séparation (partie 2.3) et l'autre pour la détection d'objets ponctuels (partie 2.4). Ces deux nouvelles approches sont ensuite validées sur les données présentées dans le chapitre 1 (partie 2.5).

2.1 Motivation

En tant que première étape pour analyser les maladies à l'échelle des cellules, la détection des objets biologiques revêt une importance considérable. Elle est nécessaire afin de pouvoir observer et quantifier les objets pour permettre d'étudier et de comprendre leurs comportements vis-à-vis des maladies ou plus simplement d'aider au diagnostic.

C'est une étape critique et pleine de défis parmi lesquels il convient de faire référence à la présence de composants du tissu non liés à l'analyse en cours, à la superposition de plusieurs coupes de tissus ou encore à la présence d'irrégularités dans la coupe dues aux imperfections de la préparation de l'échantillon. Également, la superposition de différents marqueurs, la présence de chevauchements de cellules ainsi que la morphologie imprévisible des variations du tissu induites par la maladie compliquent encore cette étape selon [Di Cataldo et al. \[2010\]](#).

Les variations de couleurs présentent elles aussi un problème récurrent en histopathologie par microscopie en lumière visible. En effet, la vaste variété de facteurs matériels pouvant influencer sur les couleurs comprend entre autres les différents modèles de scanners, l'utilisation de divers colorants ou réactions chimiques provenant de plusieurs fabricants ou simplement de multiples lots de produits, et des conditions d'entreposage puisque les marqueurs perdent de leur intensité s'ils sont exposés à la lumière. Les couleurs varient également en fonction de la procédure de marquage, en particulier selon la quantité de produit utilisée afin de marquer un échantillon, le temps de la procédure et l'homogénéité du procédé d'application sur la totalité du tissu. De même, l'épaisseur de la lame histologique influe puisque la transmission lumineuse est fonction de l'épaisseur de la section. La figure 2.1 met clairement en évidence la variabilité de l'apparence des marqueurs.

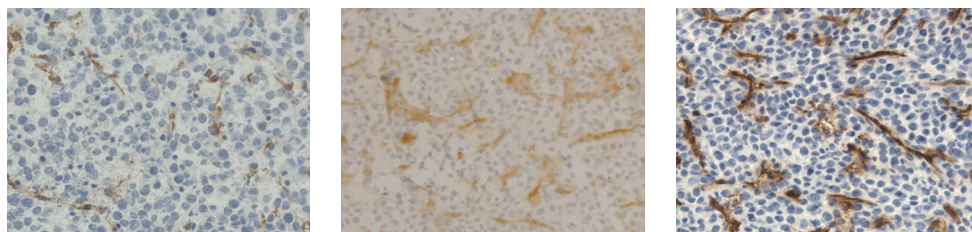


FIGURE 2.1 – Exemples d'images marquées par l'hématoxyline et le DAB. Cela illustre clairement la variabilité de l'apparence de ces marqueurs.

Afin de limiter au mieux toutes ces sources de variations, [Lyon et al. \[1994\]](#) soulignent le besoin de standardiser les produits et la marche à suivre

lors des procédures histologiques. Cependant, une standardisation complète est difficilement réalisable avec les technologies actuelles notamment à cause des découpes manuelles des échantillons et de l'estompage progressif des marqueurs dans le temps. Actuellement, les qualités de l'image et du marquage sont jugées de façon subjective et via des comparaisons entre laboratoires afin de minimiser les variations visibles et leurs impacts sur la qualité du diagnostic.

Outre ces difficultés liées aux contraintes physiques lors de l'acquisition des images histologiques, la détection et la segmentation par ordinateur des objets biologiques est également complexe à cause du manque de vérité terrain permettant une validation correcte. Par exemple, la validation de la séparation des marqueurs obtenue de manière mathématique n'est pas comparable avec une vérité terrain de type chimique. Comme le montre la figure 2.2, même chimiquement, il n'existe pas de séparation parfaite des objets. L'éosine utilisée afin de marquer le cytoplasme teinte également les noyaux des cellules, de façon similaire l'hématoxyline marque légèrement le cytoplasme.

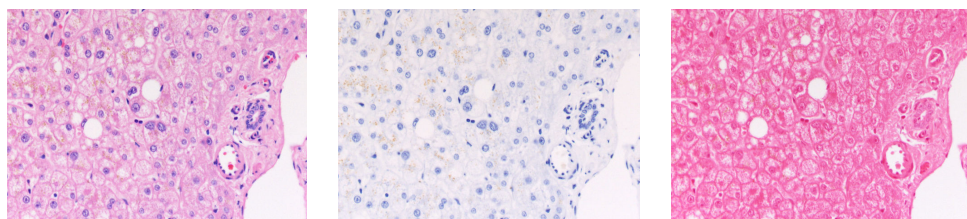


FIGURE 2.2 – Exemple de séparation chimique des marqueurs hématoxyline (H), ciblant les noyaux, et éosine (E), visant le cytoplasme. À gauche, l'image originale, au centre l'image H et à droite E. Cela montre bien que l'hématoxyline marque également le cytoplasme de même que l'éosine fait ressortir les noyaux. Les objets marrons, particulièrement visibles dans l'image H, sont des pigments de lipofuscine naturellement de couleur marrons. Ces images proviennent de l'article de [Tadrous \[2010\]](#).

2.2 État de l’art

La segmentation et la séparation des objets biologiques dans les images microscopiques sont une part importante du traitement de ces images particulières dans de nombreuses applications de recherche ou cliniques. Étant donné la large variabilité notamment de modalités d’images, de procédures de marquage et de conditions expérimentales, de nombreuses méthodes ont été développées et appliquées, voir notamment les travaux de [He et al. \[2012\]](#), [Boucheron \[2008\]](#) et [Gurcan et al. \[2009\]](#).

2.2.1 Séparation

Bien que des méthodes plus récentes soient présentées dans la littérature ([Gavrilovic et al. \[2013\]](#); [Tadrous \[2010\]](#)), la déconvolution des couleurs reste majoritairement utilisée pour la séparation des marqueurs dans les images histologiques. Toutes ces méthodes ne sont pas à même de permettre la quantification des marqueurs de type immunohistochimique tels que le DAB. La déconvolution des couleurs est donc présentée en détails ci-après suivie par les méthodes les plus courantes pour séparer le DAB.

Déconvolution des couleurs

La déconvolution des couleurs, proposée par [Ruifrok et Johnston \[2001\]](#), est une méthode très utilisée afin de transformer l’image couleur originale dans un espace couleur lié aux marqueurs. Chaque canal couleur de cet espace correspond à l’intensité optique d’un marqueur. Cette technique peut être utilisée afin de séparer jusqu’à trois marqueurs absorbant la lumière, dans le cas d’une séparation entre deux marqueurs, un troisième canal contenant l’erreur est généré, voir les figures 2.3, 2.6 et 2.7.

Soit une image \mathcal{I} définie comme un espace bidimensionnel de pixels C avec une fonction d’espace couleur associée ψ attribuant les intensités de rouge, vert et bleu à chaque pixel $c \in C$. La relation entre ψ et le nouvel espace couleur $\hat{\psi}$ correspondant aux marqueurs est alors définie selon la loi de Beer Lambert :

$$\psi = \psi_0 e^{-S\hat{\psi}}, \quad (2.1)$$

où ψ_0 est l’intensité de la lumière incidente, et S est la matrice définissant les facteurs d’absorption de chaque marqueur selon les composantes de ψ .

Elle s'écrit de la façon suivante :

$$S = \begin{bmatrix} r_{s1} & g_{s1} & b_{s1} \\ r_{s2} & g_{s2} & b_{s2} \\ r_{s3} & g_{s3} & b_{s3} \end{bmatrix}, \quad (2.2)$$

où r_{s1} , g_{s1} et b_{s1} correspondent respectivement aux valeurs prédéfinies et normalisées de rouge, vert et bleu du premier marqueur $s1$, il en est de même pour les marqueurs $s2$ et $s3$. La valeur d'un pixel c dans ce nouvel espace couleur est finalement définie comme :

$$\forall c \in C, \psi(c) \neq 0, \hat{\psi}(c) = S^{-1} \left[-\log(\psi(c)) \right]. \quad (2.3)$$

S^{-1} donne la matrice de déconvolution des couleurs et $-\log(\psi)$ est l'espace de densité optique où une combinaison linéaire de valeurs de marqueurs correspond à une combinaison linéaire de densités optiques.

La déconvolution des couleurs est très sensible à l'estimation des valeurs de S . Ainsi la calibration des vecteurs couleurs de chaque marqueur est une étape primordiale. Idéalement, ces vecteurs doivent être définis à partir d'un échantillon de contrôle marqué par un seul marqueur suivant un procédé manuel laborieux tel que décrit par [Ruifrok et Johnston \[2001\]](#). Puisque cette procédure est complexe et n'est pas toujours envisageable, de nombreuses recherches portent sur une automatisation de l'estimation de S . Ces méthodes ([Newberg et Murphy \[2008\]](#); [Rabinovich et al. \[2003\]](#); [Macenko et al. \[2009\]](#)) se basent sur une *non-negative matrix factorization*, une analyse en composantes indépendantes, la distribution du canal de teinte de l'espace couleur HSV ou une décomposition en valeurs singulières géodésique.

Puisque la déconvolution des couleurs se base sur la loi de Beer Lambert, elle ne convient pas à tous les marqueurs maniés en histopathologie. Notamment, les marqueurs d'immunohistochimie qui dispersent la lumière au lieu de l'absorber. Dans ce cas, la déconvolution des couleurs peut être utilisée afin de segmenter ou d'améliorer la visibilité de l'image mais pas pour quantifier l'intensité du marqueur. Le DAB, présent dans les données utilisées, est le marqueur le plus couramment employé de cette famille.

Méthodes spécifiques au DAB

Les méthodes proposées dans la littérature ([Brey et al. \[2003\]](#)) afin de séparer le DAB des autres marqueurs cherchent une nouvelle représentation de l'image dans laquelle la différence entre les pixels marqués par

le DAB et ceux du fond de l'image est maximale afin de pouvoir les extraire via une méthode de seuillage. Dans la littérature, le DAB est plus fréquemment utilisé conjointement avec de l'hématoxyline, comme c'est le cas dans l'application principale de cette thèse. Les techniques ici discutées se concentrent donc uniquement sur la séparation de ces deux marqueurs.

La méthode la plus simple consiste à n'utiliser que le canal bleu de l'image qui présente le plus grand contraste entre le bleu de l'hématoxyline et le marron du DAB, [King et al. \[2002\]](#). La sélection des pixels marqués par le DAB en utilisant un seuillage sur les valeurs du canal bleu est la meilleure option concernant les canaux RGB mais les interférences avec le fond de l'image peuvent être considérables. Afin de minimiser ces interférences, [Brey et al. \[2003\]](#) suggèrent d'utiliser le canal bleu normalisé (BN) selon la formule :

$$\forall c \in \mathbb{N}^2, BN(c) = \frac{255 \times Bleu(c)}{Rouge(c) + Vert(c) + Bleu(c)}, \quad (2.4)$$

où *Rouge*, *Vert* et *Bleu* correspondent aux valeurs du pixel considéré c selon les trois canaux variant de 0 à 255.

Le canal de teinte du format HSV, correspondant à une estimation de la longueur d'onde dominante de chaque pixel, fournit une méthode plus efficace pour estimer un changement de couleur et a donc été suggéré par [Ma et Lozanoff \[1999\]](#) pour améliorer la séparation du DAB.

Puisque le DAB est de teinte marron, [Ruifrok \[1997\]](#) propose de définir un canal marron où la valeur de chaque pixel c est calculée selon l'équation :

$$\forall c \in \mathbb{N}^2, Marron(c) = Bleu(c) - 0.3 \times (Rouge(c) + Vert(c)). \quad (2.5)$$

La figure 2.3 illustre les résultats obtenus pour ces différentes méthodes ainsi que la déconvolution des couleurs. Il est bon de relever que seuls les canaux bleu et marron ainsi que la déconvolution des couleurs permettent d'extraire le DAB malgré les superpositions avec de l'hématoxyline. Néanmoins, le résultat le plus satisfaisant visuellement reste celui obtenu par la déconvolution des couleurs. Une analyse quantitative de ces méthodes est faite par [Brey et al. \[2003\]](#) afin de fournir un point de départ pour les laboratoires de recherche désireux de proposer des méthodes automatiques pour quantifier la densité, et non l'intensité, des marqueurs immunohistochimiques.

Récemment, la compagnie Visiopharm[®] a breveté¹ la technologie de double marquage virtuel qui permet de correctement séparer et analyser

1. Feature-based registration of sectional images, brevet : EP 2095332 B1.

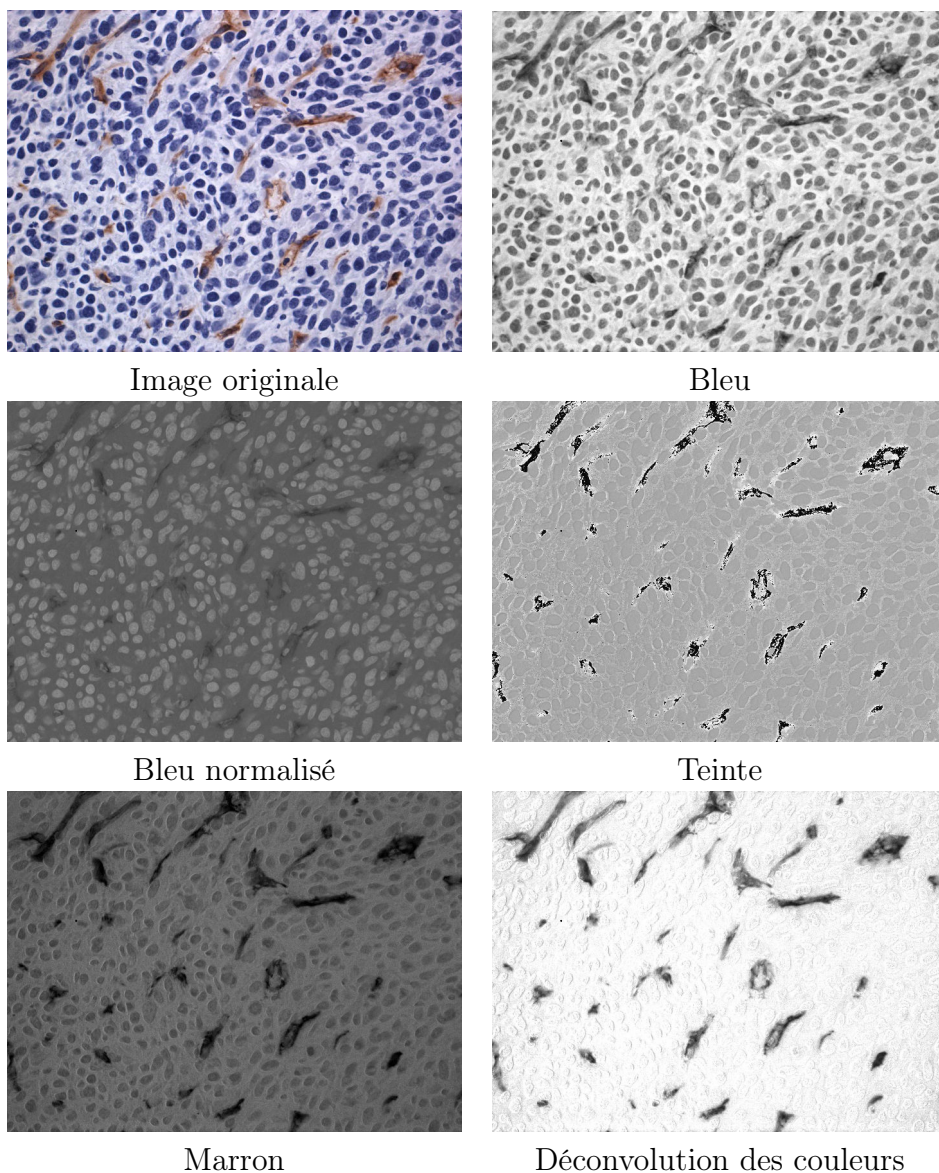


FIGURE 2.3 – Comparaison visuelle des différentes méthodes de séparation du DAB sur une image de bonne qualité.

les marqueurs immunohistochimiques, incluant une quantification de leur intensité. Aucune donnée publique n'est disponible à notre connaissance actuellement.

2.2.2 Segmentation d'images histologiques

La segmentation d'images médicales est une problématique complexe pour laquelle de très nombreuses méthodes existent. Bien souvent, ces techniques sont spécifiques à une unique application ou à une seule modalité d'imagerie. Certaines des approches les plus couramment utilisées en analyse d'image histologiques sont ici présentées et regroupées en trois familles selon qu'elles se basent sur les valeurs des pixels, des contours ou des régions. Certaines de ces approches utilisent des graines dont les méthodes de détection sont également discutées ci-après.

Méthodes basées pixels

Les méthodes les plus simples et rapides pour la segmentation sont les approches de type seuillage (Jiao et al. [2011]; Korzynska et al. [2013]). Sezgin et Sankur [2004] comparent plusieurs des plus couramment utilisées. Afin d'extraire plusieurs objets différents d'une même image, tels que les noyaux, le cytoplasme ou le fond, des méthodes multi-seuils sont utilisées, voir par exemple Loukas et Linney [2004]. Ces techniques ne sont cependant pas adaptées aux images contenant des variations d'intensités, beaucoup de bruit ou des superpositions d'objets. Afin de palier les variations de contraste, de luminance ou d'intensité, les images peuvent être découpées en plusieurs sous-images et un seuil adaptatif est alors défini pour chacune d'entre elles.

Lorsqu'une connaissance a priori des objets est disponible, il est possible de segmenter l'image via une méthode de classification par pixel, voir par exemple Doyle et al. [2006] et Setiadi et al. [2010]. Ces techniques regroupent des pixels connexes dans une même classe en fonction de descripteurs mesurés sur leurs voisinages. Le principal défaut de ces méthodes, en plus de la présence d'un apprentissage, apparaît en cas de colocalisation de marqueurs, et donc de superposition d'objets différents, puisqu'un unique label est donné à chaque pixel, comme le soulignent Ruifrok et al. [2003].

Méthodes basées contours

La détection des contours de l'image se fait par l'application de filtres analysant les intensités des pixels voisins ou les gradients afin de localiser les frontières entre objets et fond de l'image, par exemple le filtre de Canny dans Fuchs et al. [2008].

Les contours actifs sont un autre groupe de méthodes se basant sur

les contours des objets biologiques dans l'image, par exemple dans [Zeng et al. \[2013\]](#). Ces approches cherchent à minimiser itérativement une fonctionnelle énergétique afin de segmenter les objets recherchés ([Kass et al. \[1988\]](#)). Ces méthodes sont très sensibles à l'énergie considérée ainsi qu'à leur initialisation. Une amélioration des contours actifs, proposée par [Chan et Vese \[2001\]](#), permet de les utiliser en se basant sur les régions et non les contours. [Dzyubachyk et al. \[2010\]](#) les utilisent sur des images de cellules.

Méthodes basées régions

La méthode la plus connue dans cette famille est la ligne de partage des eaux (*Watershed*). Cette approche permet de segmenter des objets superposés ou adjacents mais l'utilisation directe de cet algorithme entraîne généralement des sur-segmentations selon [Malpica et al. \[1997\]](#). Pour contrer ce problème, des méthodes utilisant des graines ou des marqueurs ont été proposées ([Pinidiyaarachchi et Wählby \[2005\]](#); [Yang et al. \[2006\]](#)). En revanche, ces techniques sont très dépendantes de la détection des graines.

La croissance des régions (*Region growing*) est une approche regroupant en régions les groupes de pixels connexes présentant des caractéristiques similaires, telles que l'intensité ou la texture ([Peng et al. \[2011\]](#)). L'approche se base sur un ensemble de pixels servant de graines aux régions croissantes. Une région continue de grandir jusqu'à ce qu'aucun nouveau pixel ne respecte le critère de caractéristiques similaires. Le critère d'arrêt peut également être fondé sur des propriétés intrinsèques de la région telles que la taille ou la forme.

En plus de ces méthodes de segmentation courantes, d'autres approches existent et sont basées notamment sur les champs aléatoires de Markov ([Luck et al. \[2005\]](#)), le graph-cut ([Boykov et Funka-Lea \[2006\]](#)), et des méthodes d'apprentissage supervisé avec *template matching* ([Chen et al. \[2013\]](#)). Puisque les modalités d'images et de marqueurs évoluent continuellement, les méthodes efficaces pour une application spécifique peuvent ne pas être généralisables à d'autres jeux de données. Une étude comparative de plusieurs méthodes sur plusieurs applications différentes est présentée par [Irshad et al. \[2014\]](#).

Détection des graines

La détection des graines se base sur le repérage de tous les objets biologiques recherchés par un unique point, idéalement le barycentre. Les ap-

proches les plus classiques consistent à détecter les pics après une transformée de distance sur l'image binaire (Dalle et al. [2009]), à appliquer la transformée de Hough pour repérer les objets ayant une certaine forme (Cosatto et al. [2008]), ou encore la transformée H-minima (Jung et Kim [2010]). La transformée de distance requiert de seuiller l'image, rendant la méthode non robuste aux changements locaux de contraste. La transformée de Hough, lourde en termes de calculs, nécessite une forte similarité de la forme des objets, ce qui n'est pas toujours le cas notamment pour les cellules cancéreuses. La transformation H-minima a le revers de fournir un trop grand nombre de graines, entraînant des sur-segmentations.

Plus récemment, plusieurs méthodes ont été proposées afin de résoudre ces problèmes et améliorer la détection de graines. Par exemple, Veta et al. [2013] suggèrent de nettoyer l'image correspondant au marqueur recherché via de multiples opérations morphologiques puis d'appliquer une transformation en symétrie radiale et d'extraire les minimums locaux. Une autre méthode, proposée par Al-Kofahi et al. [2010], consiste en un filtrage par *Laplacian of Gaussian* multi-échelles contraint par un choix d'échelles adaptatif sur une carte de distances. Une autre approche, présentée par Qi et al. [2012], utilise un système de vote par pixel pondéré par une fonction gaussienne. Les graines sont finalement détectées en appliquant un algorithme de *mean-shift* sur les votes restants.

2.3 Séparation des marqueurs

Les images histologiques contiennent généralement plusieurs marqueurs faisant ressortir des objets biologiques différents. La séparation de ces marqueurs peut servir à segmenter les structures ciblées par un marqueur spécifique, à améliorer la visibilité de l'image pour les observateurs daltoniens ou encore à quantifier ces objets.

Dans cette partie, une méthode simple et rapide permettant de séparer l'hématoxyline du DAB ou de l'éosine est proposée. Cette méthode cherche à fournir une image par type d'objets biologiques, ainsi les objets teintés par le marqueur ciblé seront très contrastés par rapport au reste de l'image. Elle se base sur l'observation des différents canaux RGB ainsi que des canaux du modèle de couleur invariant proposé par Gevers et Smeulders [1999]. En effet, comme le montre les figures 2.4 et 2.5, ces différents canaux semblent contenir des informations complémentaires permettant de bien séparer les objets en fonction de leurs marqueurs. Les valeurs des canaux du modèle de couleur invariant $C1$, $C2$ et $C3$ sont définies comme :

$$\forall c \in \mathbb{N}^2, C1(c) = \arctan \left[\frac{Rouge(c)}{\max(Vert(c), Bleu(c))} \right] \quad (2.6)$$

$$\forall c \in \mathbb{N}^2, C2(c) = \arctan \left[\frac{Vert(c)}{\max(Rouge(c), Bleu(c))} \right] \quad (2.7)$$

$$\forall c \in \mathbb{N}^2, C3(c) = \arctan \left[\frac{Bleu(c)}{\max(Rouge(c), Vert(c))} \right] \quad (2.8)$$

Concernant l'hématoxyline, quelque soit l'autre marqueur utilisé, le canal rouge montre une teinte très sombre du marqueur opposée à un DAB ou une éosine grise et à un fond très clair. À l'opposé, le canal $C3$ du modèle de couleur invariant révèle de l'hématoxyline claire, un fond et une éosine gris et un DAB très sombre. Ainsi en divisant ces deux canaux, l'hématoxyline apparaît en noir alors que le reste de l'image, fond et éosine ou DAB inclus, est clair, y compris lorsque les deux marqueurs se superposent. L'hématoxyline est donc extraite de la façon suivante :

$$\forall c \in \mathbb{N}^2, H(c) = \frac{Rouge(c)}{C3(c)}, \quad (2.9)$$

où les valeurs nulles sont remplacées par un infiniment petit afin d'éviter toutes divisions par zéro.

De façon similaire pour le DAB, le canal bleu renferme un DAB sombre avec de l'hématoxyline grise et un fond clair. Après élévation au carré

afin de renforcer les différences et division avec le canal rouge, voir équation 2.10, une image en niveau de gris est obtenue dans laquelle le DAB est très contrasté.

$$\forall c \in \mathbb{N}^2, DAB(c) = \frac{Bleu^2(c)}{Rouge(c)}. \quad (2.10)$$

En appliquant l'équation 2.10 sur une image marquée par de l'éosine et de l'hématoxyline, l'image obtenue montre le tissu teinté par l'éosine uniquement. C'est à dire que les objets marqués par l'hématoxyline, tels que les noyaux, sont alors considérés comme appartenant au fond de l'image. Bien que cette séparation puisse être utile dans de nombreuses applications, via la détection du cytoplasme, elle ne correspond pas à la séparation chimique.

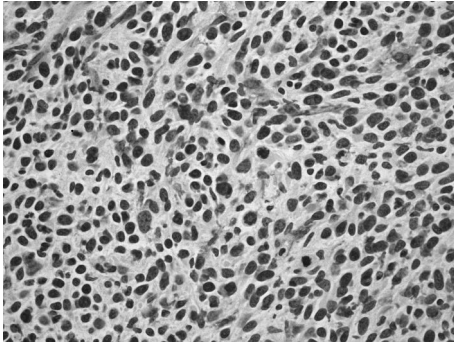
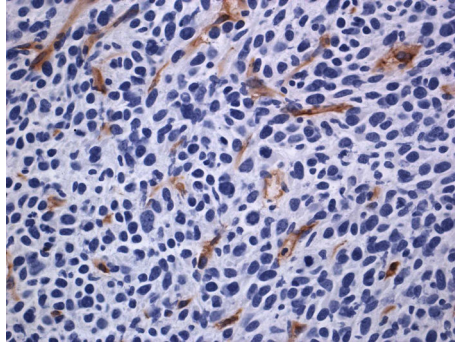
Une autre séparation de l'éosine est possible en considérant que le canal vert montre une éosine grise et une hématoxyline sombre alors que le canal rouge contient une éosine très claire et une hématoxyline sombre. Après division des deux canaux, voir équation 2.11, une image très proche du canal d'éosine obtenu par la déconvolution des couleurs est fournie.

$$\forall c \in \mathbb{N}^2, E(c) = \frac{Vert(c)}{Rouge(c)}. \quad (2.11)$$

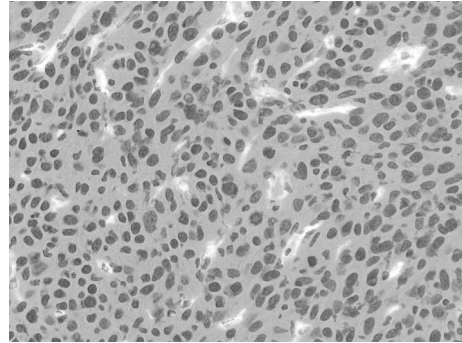
Les résultats de ces séparations pour deux image teintées par de l'hématoxyline et du DAB, et de l'hématoxyline et de l'éosine sont montrés respectivement dans les figures 2.6 et 2.7. Pour comparaison, les mêmes marqueurs extraits par la déconvolution des couleurs en utilisant les vecteurs couleurs fournis par [Ruifrok et Johnston \[2001\]](#) sont présentés.

La méthode proposée permet d'obtenir un canal d'hématoxyline similaire, plus contrasté et légèrement moins bruité que celui obtenu par la déconvolution des couleurs sur les images marquées par de l'hématoxyline et du DAB ou de l'éosine. L'extraction du DAB donne un résultat comparable tandis que celle de l'éosine permet d'obtenir sensiblement la même sortie mais en laissant le choix d'inclure ou non les objets marqués par l'hématoxyline. Bien qu'elle offre des résultats comparables à la déconvolution des couleurs, la méthode ici proposée est nettement plus simple et rapide puisqu'elle se résume à une simple division pixel par pixel.

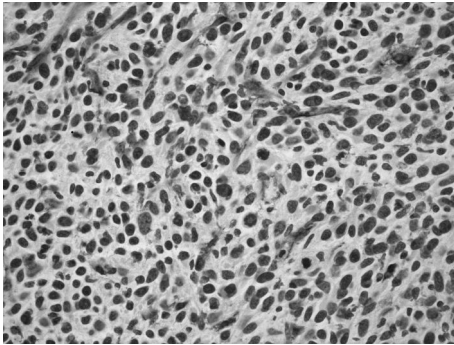
Cette méthode peut être considérée comme un pré-traitement à une technique de détection des objets biologiques mais ne permet pas de mesurer l'intensité des marqueurs. Contrairement aux autres méthodes visant à maximiser le contraste du DAB, cette approche minimise également le contraste des autres marqueurs présents dans l'image.



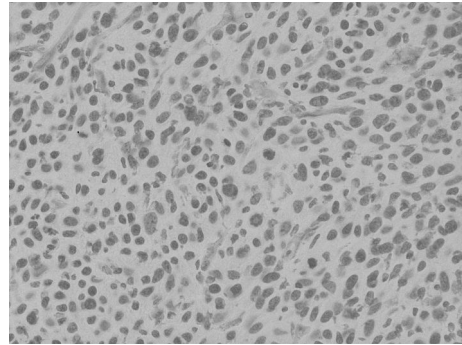
Rouge



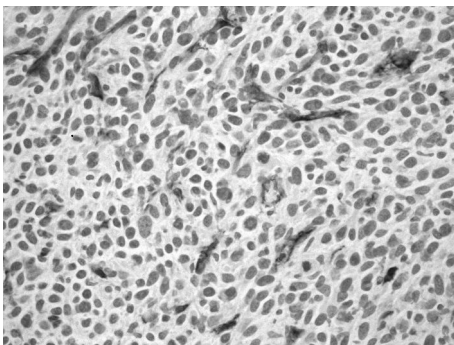
C1



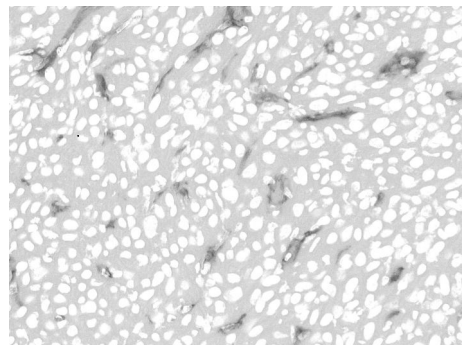
Vert



C2



Bleu



C3

FIGURE 2.4 – Illustration des différents canaux couleurs RGB (colonne de gauche) et modèle de couleur invariant (colonne de droite) d'une image marquée par de l'hématoxyline et du DAB (en haut).

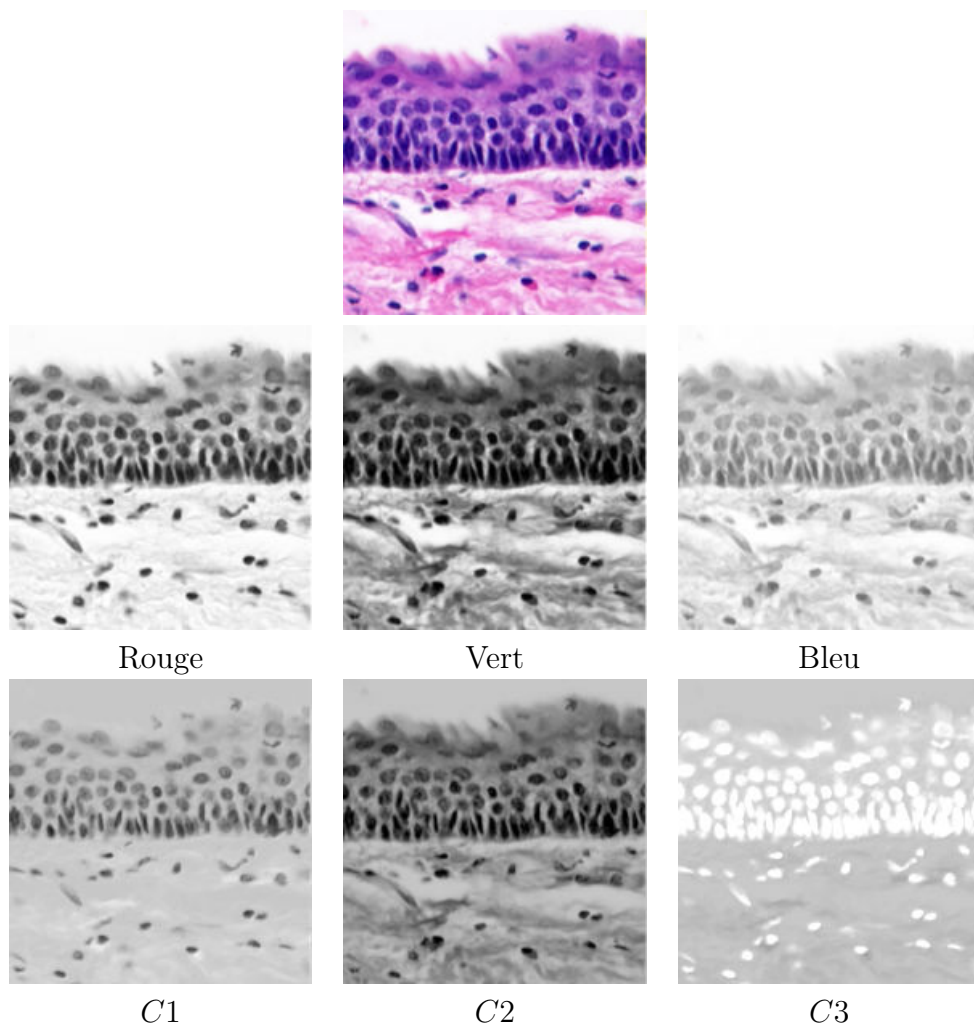
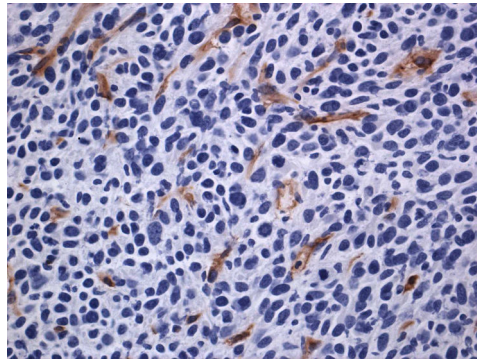
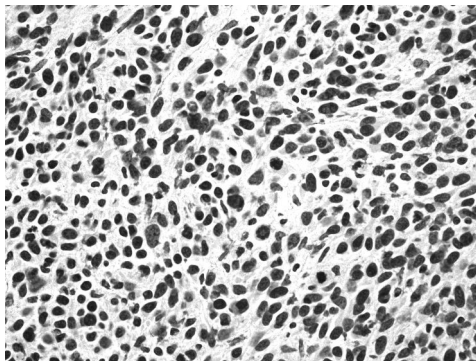


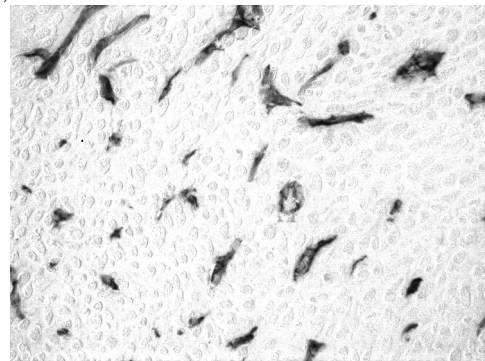
FIGURE 2.5 – Illustration des différents canaux couleurs RGB (deuxième ligne) et modèle de couleur invariant (troisième ligne) d’une image marquée par de l’hématoxyline et de l’éosine (en haut).



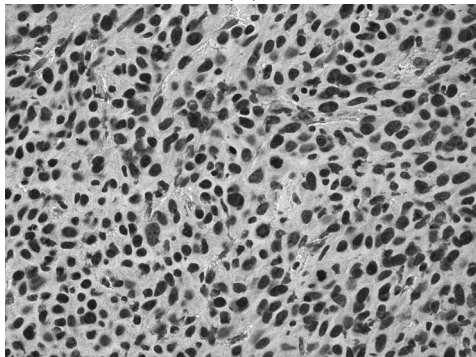
(a)



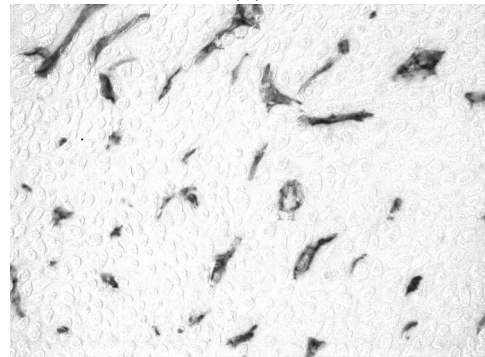
(b)



(c)



(d)



(e)

FIGURE 2.6 – Illustration de la séparation des marqueurs hématoxyline et DAB proposée. (a) l'image originale, (b) l'hématoxyline extraite et (c) le DAB. (d) et (e) montrent respectivement les mêmes marqueurs extraits par la déconvolution des couleurs.

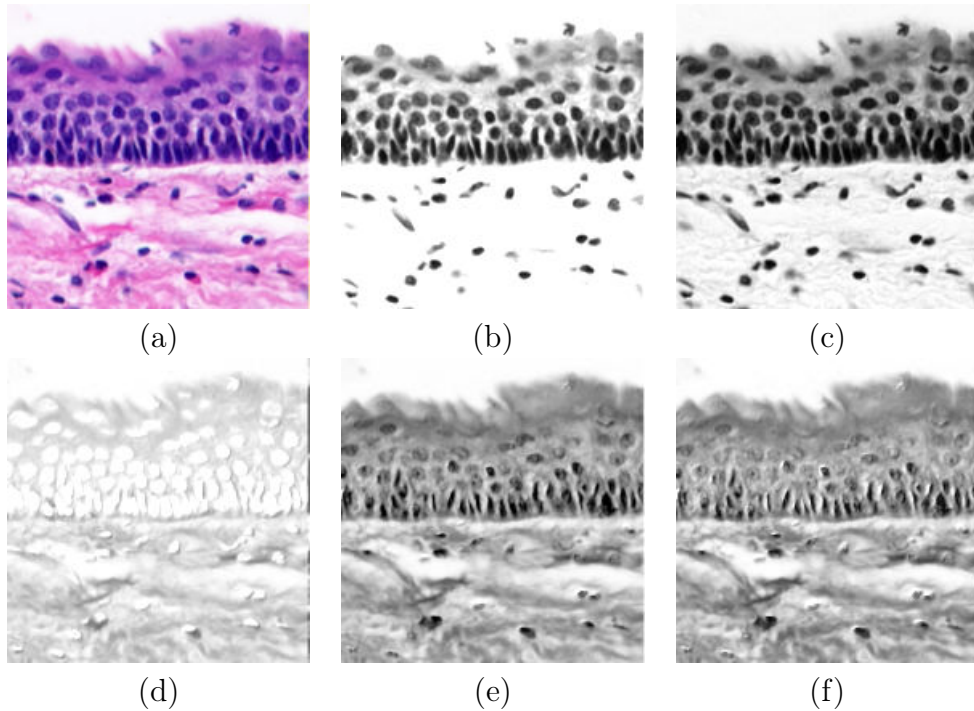


FIGURE 2.7 – Illustration de la séparation des marqueurs hématoxyline et éosine proposée. (a) l'image originale, (b) l'hématoxyline extraite, (d) l'éosine séparée sans superposition avec un autre marqueur, et (e) l'éosine par l'équation 2.11. (c) et (f) montrent respectivement l'hématoxyline et l'éosine extraites par la déconvolution des couleurs.

2.4 Extraction des différents objets biologiques

L'extraction des objets dépend des informations que l'étude souhaite prendre en compte par la suite. De ce fait, deux classes d'objets sont définies avant de présenter les méthodes d'extractions correspondantes.

2.4.1 Définitions des catégories d'objets utilisées

Une fois les marqueurs séparés, chaque type d'objets marqués est représenté sur une image en niveau de gris différente. À partir de là, on peut définir deux types d'objets.

Les objets ponctuels sont ceux qui peuvent être modélisés par un unique point par objet. Leurs informations de taille, de forme et de texture ne sont donc pas considérées pertinentes du point de vue de l'étude en cours.

Les objets surfaciques sont ceux qui ne peuvent pas être modélisés par un unique point et dont la taille, la forme ou la texture sont intéressantes pour l'étude considérée.

Compte tenu des attentes différentes selon le type d'objets considéré, les méthodes proposées dans la suite de ce chapitre sont séparées en fonction du type d'objets qu'elles concernent.

2.4.2 Extraction des objets ponctuels

La méthode d'extraction des objets ponctuels ici présentée se base sur le calcul de la carte topographique introduite par [Caselles et Morel \[1999\]](#). Le principe de cette approche, publiée dans [Alsheh Ali et al. \[2013\]](#), est de filtrer la carte topographique de l'image afin de n'avoir qu'un unique point par objet ciblé. La carte topographique est tout d'abord rappelée ainsi que ses propriétés avant de détailler les étapes de filtrages successives.

Carte topographique

Cet outil a été présenté comme étant un moyen efficace afin de représenter une image. Elle est composée des lignes de niveaux, elles-mêmes étant les bordures des ensembles de niveaux de l'image. Soit une image u , les ensembles de niveaux inférieurs χ^μ et supérieurs χ_λ de valeurs de niveau

de gris respectivement μ et λ sont définis comme :

$$\forall \lambda \in \mathbb{R}, \chi_\lambda = \left\{ x \in \mathbb{R}^2, u(x) \geq \lambda \right\}, \quad (2.12)$$

$$\forall \mu \in \mathbb{R}, \chi^\mu = \left\{ x \in \mathbb{R}^2, u(x) \leq \mu \right\}. \quad (2.13)$$

Les informations contenues dans l'ensemble des χ_λ pour tous les λ possibles (ou l'ensemble des χ^μ pour tous les μ possibles) sont suffisantes pour reconstruire l'image :

$$\begin{aligned} \forall x \in \mathbb{R}^2, u(x) &= \sup \left\{ \lambda \in \mathbb{R}, x \in \chi_\lambda : \chi_\lambda(x) \right\}, \\ &= \inf \left\{ \mu \in \mathbb{R}, x \in \chi^\mu : \chi^\mu(x) \right\}. \end{aligned} \quad (2.14)$$

La carte topographique offre donc une représentation complète de l'image. Cette technique est également invariante par changement de contraste global.

De par la définition des ensembles de niveaux, soient λ_1 , λ_2 , μ_1 et μ_2 quatre niveaux de gris, on a alors :

$$\forall (\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda_1 > \lambda_2 \Leftrightarrow \chi_{\lambda_1} \subseteq \chi_{\lambda_2}, \quad (2.15)$$

$$\forall (\mu_1, \mu_2) \in \mathbb{R}^2, \mu_1 > \mu_2 \Leftrightarrow \chi^{\mu_2} \subseteq \chi^{\mu_1}. \quad (2.16)$$

Cela implique que les lignes de niveaux des ensembles de niveaux supérieurs sont naturellement intégrées dans une structure en forme d'arbre (et respectivement pour les ensembles inférieurs). Ces deux arbres de lignes de niveaux sont alors utilisés pour représenter et manipuler les images de façon simple et efficace (Monasse et Guichard [2000]; Xia et al. [2010]).

Dans les études liées à cette thèse, la carte topographique est calculée en utilisant le code disponible dans la librairie C *Megawave2*².

Filtrage de la carte topographique

Les objets biologiques ciblés sont les plus sombres dans l'image de leur marqueur, seul l'arbre des lignes de niveaux inférieurs est donc considéré. Afin d'extraire ces objets, trois filtrages successifs sont donc effectués sur l'arbre.

Dimension. Il y a dans l'arbre de nombreuses lignes de niveaux trop petites pour correspondre aux objets biologiques. Ces lignes reflètent l'absorption non homogène du marqueur par les objets et sont donc considérées

2. <http://megawave.cmla.ens-cachan.fr/>

comme du bruit. Connaissant la dimension physique des pixels de l'image, puisque le niveau de zoom du microscope est connu, les dimensions typiques des objets biologiques recherchés dans l'image peuvent être estimées. Ainsi, un premier seuillage de l'arbre peut être effectué afin de supprimer toutes les lignes de niveaux trop petites pour correspondre à un objet. Cela peut être fait en enlevant les lignes ayant un périmètre ou une aire trop faible.

Feuilles. Afin de n'avoir qu'une unique ligne de niveaux par objet, les propriétés de la structure d'arbre de la carte topographiques sont exploitées pour ne garder que les feuilles de cet arbre.

Luminance. L'hétérogénéité du fond de l'image peut entraîner la formation de lignes de niveaux causant de fausses détections de structures similaires aux objets recherchés. Puisque l'arrière plan de l'image est plus clair que les objets ciblés, il est possible d'éliminer les lignes de niveaux dont le niveau de gris intérieur moyen est supérieur à un certain seuil, par exemple, le niveau de gris moyen de l'image.

Une fois ces trois filtrages exécutés, la carte topographique est réduite à un ensemble de lignes de niveaux chacune correspondant à un objet recherché. Puisque la modélisation désirée est de forme ponctuelle, les objets sont finalement représentés par le centre de gravité de la feuille qui leur est associée.

2.4.3 Extraction des objets surfaciques

Les objets surfaciques sont déjà bien contrastés sur l'image obtenue après séparation des marqueurs. Puisque leur forme est importante, il convient de l'altérer le moins possible lors de l'extraction, pour cela, cette étape peut se résumer à un débruitage suivi d'un seuillage de l'image. Dans le cas où ces données sont susceptibles de se chevaucher, il est recommandé d'utiliser une des méthodes de segmentation décrite dans la partie 2.2.2. Le bruit devant être supprimé est ici principalement dû aux variations dans la couleur du marqueur et à l'absorption de teinte par du tissu non ciblé.

Beaucoup de méthodes de segmentation, telles que les contours actifs ou les lignes de partage des eaux, imposent d'utiliser des graines comme initialisations. La méthode d'extraction d'objets ponctuels présentée dans la partie 2.4.2 offre une réponse à ce besoin en réduisant chaque objet à un unique point. Par exemple, si les cellules de l'image sont recherchées sous la forme d'un ensemble d'objets surfaciques, une extraction possible consiste

en la modélisation des cellules en objets ponctuels afin d'extraire un unique point par cellule dans le but d'initialiser une méthode de segmentation.

Ces deux méthodes d'extraction des objets biologiques proposent donc de représenter les objets ponctuels par un unique point et les objets surfaciques par un masque intégrant leurs formes. Ces deux représentations permettent de mesurer des statistiques du premier ordre, telles que la densité, ou du second ordre afin d'analyser l'organisation et le comportement de ces objets.

2.5 Application

2.5.1 Sur des images hématoxyline et DAB

Les méthodes de séparation et d'extraction des objets biologiques présentées précédemment sont ici validées expérimentalement sur l'ensemble de données histologiques de cancer colorectal de souris introduit dans la partie 1.5.

Séparation et extraction

La méthode de séparation présentée dans la section 2.3 a été appliquée sur toutes les images de la base de données. La figure 2.8 montre les résultats obtenus sur deux images, l'une bien teintée et avec un bon contraste et l'autre de moins bonne qualité visuelle. En particulier le DAB, marquant les vaisseaux sanguins, est brun foncé dans la première image et très clair dans la seconde. La méthode de séparation parvient à correctement isoler l'hématoxyline du DAB. Les images binaires des vaisseaux sanguins, ici considérés comme des objets surfaciques, sont obtenues en utilisant la méthode de seuillage d'Otsu avec deux classes sur l'image de DAB résultant de l'équation 2.10. Une ouverture morphologique est finalement appliquée, avec un élément structurant en forme de disque de rayon fixé expérimentalement à 5 pixels afin de supprimer le bruit.

Dans le cadre de cette étude, seuls le nombre et la localisation précise des noyaux des cellules sont recherchés. De ce fait, les noyaux sont donc modélisés comme des objets ponctuels. Afin d'obtenir l'emplacement de toutes les cellules, la carte topographique de chaque image d'hématoxyline est calculée puis filtrée par les critères proposés dans la section 2.4.2. La figure 2.9 montre le résultat sur une image tronquée. Puisque la carte topographique est invariante aux changements de contraste et que toutes les images sont capturées avec le même objectif optique, la procédure est applicable à l'ensemble de la base d'images avec le même jeu de paramètres.

Les deux paramètres ici pris en compte sont la taille physique des objets, nécessaire lors du filtrage de la carte topographique, et le pas de quantification utilisé lors du calcul de la carte. Le pas de quantification doit être sélectionné via un compromis entre l'efficacité de la méthode et sa rapidité ainsi que son coût en mémoire. En effet, dans le cas de ces images, mesurer la totalité de la carte topographique donne plus de deux millions de lignes de niveaux, ce qui n'est pas exploitable pour une application réelle. Afin de réduire la taille de la carte et puisque les images sont relativement

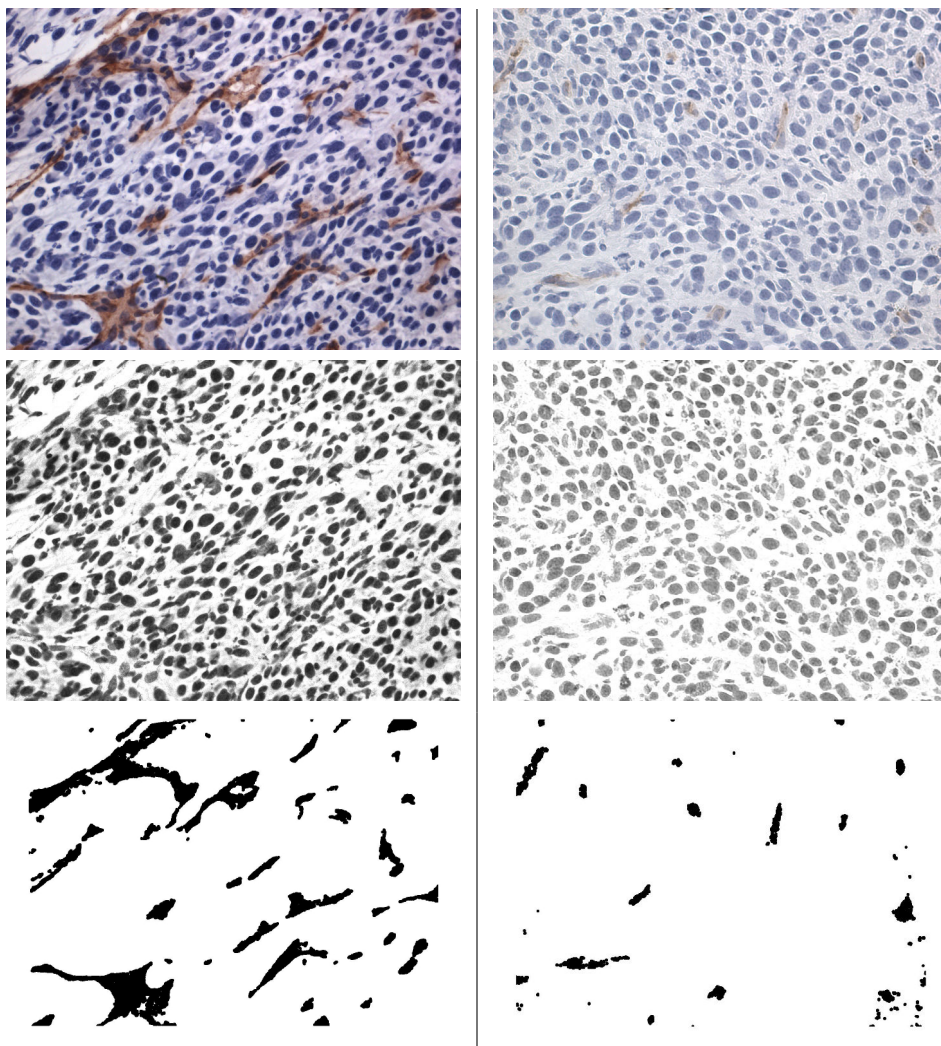


FIGURE 2.8 – Exemples de la séparation des objets biologiques. La première ligne présente les images originales. L'hématoxyline, qui marque les noyaux des cellules, et le DAB, qui cible les vaisseaux sanguins, sont respectivement dans la deuxième et la troisième ligne. L'image des vaisseaux, objets surfaciques, présente l'extraction complète, qui consiste en la séparation suivie d'un seuillage et d'un filtrage du bruit.

bien contrastées et de qualités similaires, un pas de cinq niveaux de gris a été fixé. Un tel pas permet de n'avoir *que* quelques centaines de milliers de lignes de niveaux à prendre en compte sans provoquer de pertes d'informations tangibles.

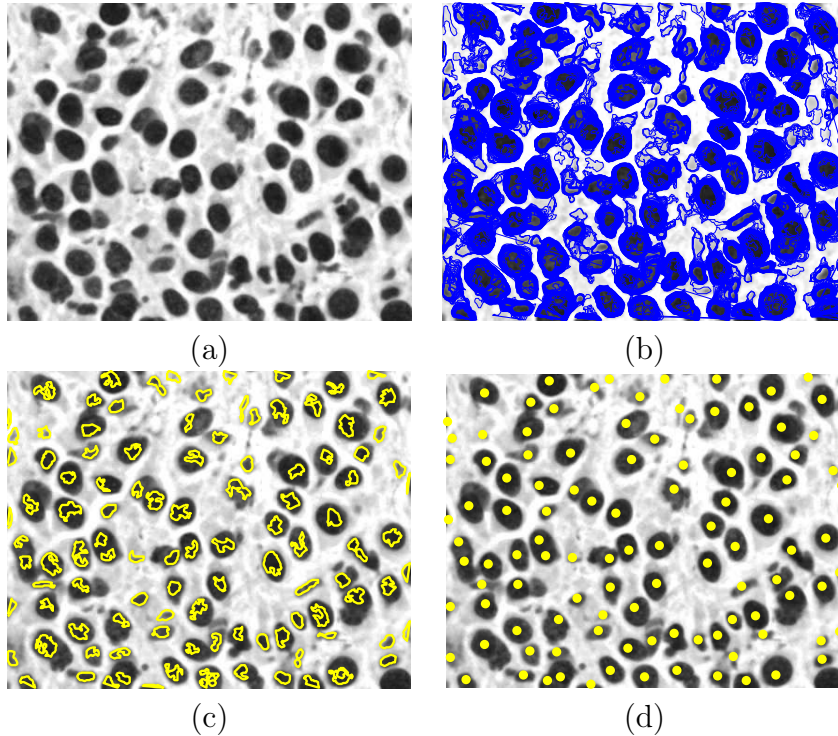


FIGURE 2.9 – Extraction des noyaux sur une image tronquée. La carte topographique est calculée sur l’image originale (a) et est filtrée par le critère de dimension, donnant ainsi de nombreuses lignes de niveaux concentrées autour des noyaux (b). Seules les feuilles de l’arbre sont conservées (c) et les noyaux sont finalement représentés par les centres de gravité des quelques lignes restantes (d).

Validation expérimentale

Afin de valider les méthodes de séparation et d’extraction proposées, les noyaux de six images, une par jour et par modèle, ont été localisés par un expert. Le tableau 2.1 montre le nombre de noyaux extraits par la méthode proposée et le compare à celui de l’expert tout en mesurant le taux d’erreur. La figure 2.10 montre une image tronquée avec les points correspondant aux noyaux détectés et ceux marqués par l’expert. Ceci afin d’illustrer à la fois la proximité entre les noyaux détectés automatiquement et manuellement, et les cas particuliers entraînant des erreurs dans la détection.

L’erreur de détection entre la méthode automatique et l’expert se situe entre 0,7 % et 4,6 %. Les erreurs sont susceptibles d’avoir deux raisons, d’une part les noyaux se chevauchant peuvent avoir une teinte uniforme et donc être difficiles à distinguer. D’autre part, les noyaux de cellules cancéreuses présentent des tailles et des textures très variables, entraînant

TABLE 2.1 – Comparaison entre le nombre de noyaux extraits automatiquement et manuellement par un expert.

Modèle	Ectopique			Orthotopique		
Jour	11 ^{ième}	15 ^{ième}	21 ^{ième}	11 ^{ième}	15 ^{ième}	21 ^{ième}
Expert	814	885	887	863	841	869
Algorithmme	851	898	875	850	847	829
Erreur	4,5%	1,5%	1,4%	1,5%	0,7%	4,6%

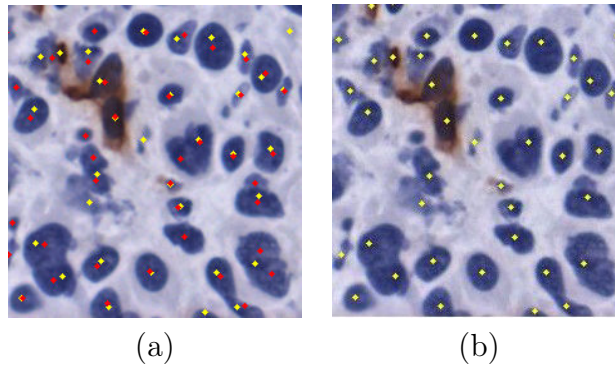


FIGURE 2.10 – Illustration des différences dans l'extraction automatique et manuelle des cellules. Les points rouges correspondent à l'extraction faite par un expert, alors que les points jaunes représentent l'extraction par la méthode proposée. (a) montre les deux groupes de points détectés alors que (b) montre uniquement les noyaux détectés automatiquement afin de permettre de bien visualiser les éventuelles superpositions.

parfois une forte hétérogénéité des marqueurs conduisant à la détection de deux noyaux au lieu d'un seul plus gros que la normale. Néanmoins, l'erreur est faible et atteste de la précision et de la robustesse de la méthode proposée.

2.5.2 Sur des images hématoxyline et éosine

De nombreuses applications utilisent de l'hématoxyline et de l'éosine afin de faire ressortir respectivement les noyaux et le stroma. Ces images présentent donc des noyaux teintés par les deux marqueurs et des membranes ainsi que le cytoplasme principalement teintés par l'éosine. Il est souvent important de décomposer l'image selon les objets biologiques afin de mesurer des statistiques descriptives ou inférentielles, telles que l'organisation des objets les uns par rapport aux autres ou leurs formes, voir par

exemple les travaux de [Kothari et al. \[2013b\]](#) et de [Garnier et al. \[2014\]](#).

La méthode de séparation et d'extraction des objets proposée dans ce chapitre permet d'accéder à ces informations. Pour cela, il convient de considérer tous les objets comme étant des données surfaciques et donc d'appliquer des opérations de seuillage après la séparation des marqueurs. L'image ici décomposée (voir figure 2.11) présente du tissu sain, avec des teintes relativement standard, une zone métastasée avec des cellules de formes et tailles variables agglomérées les unes aux autres, ce qui donne une teinte d'éosine plus terne et une hématoxyline plus marquée. L'image montre également une importante zone non marquée qui correspond à un vaisseau sanguin.

Du point de vue de l'application, il convient de segmenter les noyaux de cellules, le cytoplasme et les zones non marquées. L'éosine est ici séparée selon l'équation 2.11. Les seuillages sont effectués par la méthode d'Otsu avec trois classes, une pour les objets ciblés, une pour le fond de l'image et la dernière pour le bruit dû aux variations dans l'absorption des marqueurs. Puisque le cytoplasme cancéreux a une teinte d'éosine moins marquée, les trois sorties du filtrage d'Otsu correspondent au cytoplasme sain, au cancéreux et au fond de l'image. De ce fait, le cytoplasme est extrait en regroupant ces deux classes. Afin de valider qualitativement cette méthode, les résultats obtenus sur une image représentative sont comparés visuellement avec ceux issus de la déconvolution des couleurs, voir figure 2.11.

Les deux approches donnent des résultats très proches cependant la méthode ici proposée présente moins de bruit et des objets visuellement mieux définis. Le principal avantage de cette séparation vis-à-vis de la déconvolution des couleurs reste néanmoins sa simplicité et sa rapidité d'exécution.

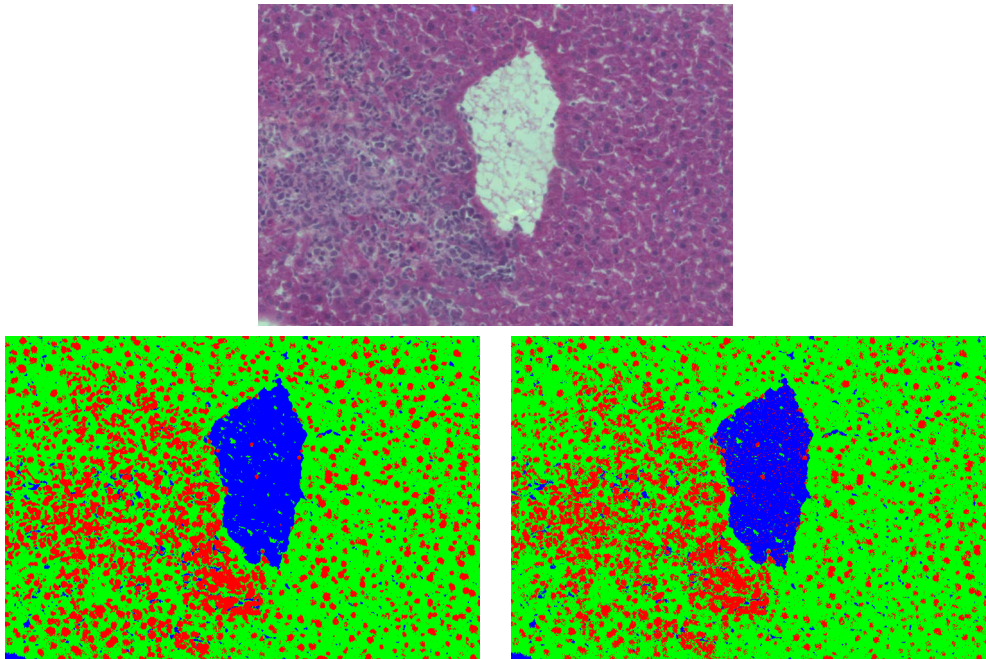


FIGURE 2.11 – Segmentation des objets biologiques sur une image représentative teintée par de l’hématoxyline et de l’éosine. L’image originale est en haut, l’image de gauche montre la segmentation fournie par la méthode proposée alors que celle de droite montre le résultat via la déconvolution des couleurs. Les canaux rouge, vert et bleu présentent respectivement les noyaux, le cytoplasme et le reste.

2.6 Résumé et perspectives

La détection des objets biologiques est une étape nécessaire à la plupart des analyses d'images histologiques. Ces images sont colorées par plusieurs marqueurs ciblant différents objets, la détection passe donc généralement par la séparation des marqueurs. Dans ce chapitre, une méthode simple et rapide de séparation est proposée et comparée qualitativement à la déconvolution des couleurs qui reste l'approche la plus utilisée dans la littérature. Une fois chaque type d'objet présent dans une image en niveaux de gris différente, deux méthodes de détection différentes sont introduites selon que les objets sont modélisables par des points ou non. Pour les objets ponctuels, la méthode de détection se base sur le calcul de la carte topographique et a été comparée à un expert pour la détection des noyaux de cellules dans les images de cancer colorectal de souris. Pour les objets surfaciques, une simple méthode de segmentation par seuillage est utilisée puisque les objets de notre étude, les vaisseaux sanguins, ne se superposent pas. Une validation qualitative sur des images d'hématoxyline et d'éosine a été étudiée. Cette détection permet finalement de mesurer des caractéristiques des premier et second ordres afin d'analyser le comportement de ces objets vis-à-vis de certaines maladies telles que le cancer.

À l'avenir, il peut être intéressant d'étudier le comportement de la méthode de séparation sur des images de qualité plus médiocre. De même, les techniques de détection doivent être validées quantitativement sur un nombre de données plus important et avec une vérité terrain définie par plusieurs experts à plusieurs reprises, ceci afin de minimiser l'impact des variabilités inter et intra experts. Il serait également intéressant d'étudier le comportement des techniques de segmentation des noyaux des cellules en appliquant la méthode de détection des objets ponctuels pour la localisation des graines. Finalement, une méthode pour régler automatiquement le pas de quantification de la carte topographique en fonction de la qualité de l'image permettrait de supprimer un paramètre tout en optimisant le temps de calcul et l'utilisation mémoire par rapport à la précision de la détection. Une première solution serait d'étudier l'histogramme de l'image pour adapter le pas à sa dynamique.

Statistiques spatiales

Les statistiques spatiales impliquent l'analyse de l'organisation de toutes sortes de données en donnant des informations quantitatives concernant leurs interactions entre elles et avec le reste de l'espace. Ces techniques sont utilisées dans beaucoup de domaines, allant de la géologie à la sociologie. Elles sont le plus couramment appliquées en sismologie, afin d'étudier la distribution spatiale des séismes et rechercher des motifs permettant de prédire leurs emplacements futurs, ou encore dans le domaine de la santé publique dans le but d'étudier les motifs spatiaux dans les épidémies ou d'analyser les corrélations entre les maladies et d'autres facteurs environnementaux. Ce chapitre traite des statistiques spatiales et de leur utilisation dans le cadre des images histologiques. Les statistiques usuelles pour les processus ponctuels sont décrites dans la partie 3.2. La méthode proposée pour l'analyse de données ponctuelles et surfaciques est présentée dans la partie 3.3 et est ensuite appliquée et discutée sur l'ensemble d'images histologiques en partie 3.4.

3.1 Motivation

Comprendre l'organisation spatiale des objets est essentiel dans de nombreux domaines scientifiques parce que cela permet d'obtenir des informations concernant leurs interactions entre eux et avec leur environnement. L'organisation des objets peut être étudiée à toutes les échelles envisageables, allant de celle des galaxies, à celle d'un pays pour étudier des phénomènes climatiques ou encore à celle de l'atome en physique.

Les images histologiques rendent possible l'observation de la structure et de l'organisation d'un échantillon de tissu sans l'altérer. Ainsi, elles permettent d'étudier les relations entre les différents objets biologiques présents afin d'apporter des connaissances quant aux mécanismes intrinsèques des maladies et des tissus biologiques. En reprenant les définitions des objets ponctuels et surfaciques du chapitre 2, il est donc possible de modéliser les objets pouvant être réduits à un simple point comme des réalisations de processus ponctuels. Ainsi leurs distributions spatiales peuvent être estimées en utilisant les nombreuses fonctions statistiques liées à ce formalisme mathématique. Une autre solution pour étudier les relations spatiales entre objets ponctuels est d'utiliser des caractéristiques estimées à partir de structures de graphes.

D'après [Gurcan et al. \[2009\]](#), les graphes sont construits pour modéliser différents états d'un tissu et permettre de distinguer entre deux états par le calcul de métriques sur les graphes et la classification de leurs valeurs. Le tableau 3.1 regroupe les caractéristiques couramment extraites à partir des structures de graphes. Plus d'informations sur les structures de graphes et les caractéristiques associées peuvent être trouvées dans la thèse de [Boucheron \[2008\]](#). Au final, près de 150 caractéristiques différentes concernant les relations spatiales peuvent être extraites de l'ensemble des structures de graphes. Ces caractéristiques fournissent des informations structurelles à même de décrire l'organisation d'un tissu telles que les distributions locales et globales autour d'un objet ou groupe d'objets ou encore la connectivité globale de la structure. Ces méthodes sont utilisées afin de distinguer entre des tissus sains, abîmés ou cancéreux dans le cerveau ([Gunduz et al. \[2004\]](#); [Demir et al. \[2005\]](#)), les seins ([Bilgin et al. \[2007\]](#)) ou encore les os ([Bilgin et al. \[2010\]](#)). De même, elles sont employées afin de reconnaître les différents types de lymphomes folliculaires par [Oztan et al. \[2012\]](#).

Ces structures se fondent uniquement sur la distribution des objets présents et ne prennent pas en compte les caractéristiques intrinsèques du tissu considéré telles que la distance à partir de laquelle les objets n'interagissent plus les uns avec les autres. [Acar et al. \[2012\]](#) suggèrent de prendre en compte cette information sémantique lors de la construction du graphe

afin d'identifier des tissus sains ou cancéreux provenant du cerveaux, des seins et des os. Ces méthodes ont cependant l'inconvénient d'être dépendantes de la densité des objets ponctuels dans l'image. Afin de comparer les relations spatiales dans des images présentant des densités significativement différentes, il est donc plus intéressant d'utiliser les statistiques du deuxième ordre des processus ponctuels selon [Fleischer et al. \[2006\]](#). De plus, ces approches permettent de mesurer les interactions entre deux populations d'objets différents telles que dans les travaux de [Lagache et al. \[2013b\]](#).

Les fonctions statistiques des processus ponctuels ont été utilisées dans des contextes biomédicaux variés comme la description de l'organisation de cellules cancéreuses dans les seins ([Mattfeldt et al. \[2009\]](#); [Setiadi et al. \[2010\]](#)), dans les tumeurs cérébrales ([Jiao et al. \[2011\]](#)) ou encore pour différencier les fibres nerveuses de l'épiderme entre patients diabétiques et non diabétiques par [Waller et al. \[2011\]](#). L'efficacité de ces statistiques a été montrée dans la littérature afin de comparer l'organisation de données malades par rapport à des données saines, ce qui amène à étudier des images présentant différentes densités et distributions spatiales.

Quant aux objets surfaciques, la littérature ne semble pas proposer d'approches permettant de modéliser leurs distributions spatiales ni même leurs interactions avec d'autres objets, qu'ils soient surfaciques ou ponctuels. Ces objets, tels que les vascularisations, sont généralement réduits à leurs centres de gravité ([Weyn et al. \[2004\]](#); [Balsat et al. \[2011\]](#)) ou alors leurs interactions sont estimées en prenant la plus courte distance entre les paires de points appartenant aux contours de deux objets voisins ([Bernsen et al. \[1999\]](#)). Ces deux approches ne prennent pas en compte la forme ni la taille des objets et donnent donc autant de poids aux petits objets qu'aux grands.

TABLE 3.1 – Caractéristiques de relations spatiales couramment extraites des structures de graphes et des arbres.

Structure	Caractéristiques
Tessellation de Voronoi	Nombre de nœuds, nombre d'arêtes, nombre de cycles, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Randic, aire, facteur de rondeur, homogénéité du facteur de rondeur, désordre de surface.
Triangulation de Delaunay	Nombre de nœuds, longueur d'arête, degré, nombre d'arêtes, nombre de cycles, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Wiener, excentricité, indice de Randic, dimension fractale.
Arbre couvrant de poids minimal	Nombre de nœuds, longueur d'arête, degré, nombre de voisins, indice de Wiener, excentricité, indice de Randic, indice de Balaban, dimension fractale.
Graphe de voisinage de O'Callaghan	Nombre de nœuds, nombre d'arêtes, nombre de cycles, nombre de voisins, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Randic, dimension fractale.
Graphe connecté	Longueur d'arête, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Wiener, excentricité, indice de Randic, dimension fractale.
Graphe de voisinages relatif	Nombre de nœuds, nombre d'arêtes, nombre de cycles, nombre de voisins, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Randic, dimension fractale.
Graphe des k plus proches voisins	Nombre de nœuds, longueur d'arête, degré, nombre de triangles, nombre de k-parcours, rayon spectral, exposant propre, indice de Wiener, excentricité, indice de Randic, dimension fractale.

3.2 État de l'art pour les données ponctuelles

3.2.1 Notion de processus ponctuels

Un processus ponctuel est une modélisation stochastique d'un ensemble irrégulier de points. La réalisation d'un tel processus est définie comme un nombre aléatoire de points distribués de façon aléatoire sur un sous-espace de \mathbb{R}^n dont seule une partie est observée via une fenêtre. L'ensemble des points compris dans la fenêtre d'observation est alors nommé motif ponctuel.

Dans ce cadre, un événement désigne l'occurrence, observée dans la fenêtre, d'un phénomène spatial. Un motif ponctuel désigne donc l'ensemble des événements enregistrés dans la fenêtre d'observation. Un point correspond à un emplacement arbitraire, différent de celui d'un événement, dans cette même région.

Un motif ponctuel \mathbf{X} est un ensemble d'événements dont le nombre et les emplacements sont supposés aléatoires. L'intensité d'un processus est la quantité théorique d'événements par unité de surface, elle est estimée empiriquement par la densité du motif observé. Deux hypothèses sont généralement faites vis-à-vis de ces données, la stationnarité et l'isotropie. Un processus stationnaire n'évolue pas selon sa position dans l'espace. De ce fait, il suit la même loi de distribution avec la même intensité dans toute la fenêtre d'observation. De même, un processus isotrope est invariant par rotation et ne présente donc aucune direction privilégiée. Un processus stationnaire et isotrope est dit homogène.

Il existe trois types de distributions spatiales selon lesquelles l'emplacement des événements peut être régi. Ils peuvent être distribués purement aléatoirement ou présenter une tendance à être regroupés ensemble, formant alors des agrégats, mais peuvent également être distribués plus régulièrement, il y a alors une forme d'inhibition des événements les faisant apparaître avec une distance minimale entre eux. La répartition de type purement aléatoire correspond à des événements dispersés selon une distribution de Poisson, et le processus est donc dit de Poisson. Ce type de processus est défini comme étant la référence pour l'étude de toutes les données ponctuelles puisque l'intérêt principal de l'analyse de ces processus repose dans l'estimation de leurs écarts par rapport à une distribution aléatoire.

Le but de l'analyse de ces processus est de décrire leurs distributions de façon paramétrique, et éventuellement d'estimer ces paramètres, le tout en se basant sur les événements présents dans la fenêtre d'observation. L'étude

statistique d'un processus ponctuel cherche donc à répondre à plusieurs questions :

- Est-ce que les événements sont distribués de façon strictement aléatoire ?
- Est-ce que les événements sont distribués plus régulièrement ?
- Est-ce que les événements sont distribués de sorte à former des agrégats autour de certains emplacements ?
- Est-ce possible d'identifier un modèle de génération des événements ?
- Si deux, ou plus, processus ponctuels sont observés, existe-t-il une interaction entre eux ?

3.2.2 Statistiques descriptives fonctionnelles

Les fonctions permettant d'estimer des statistiques spatiales descriptives des interactions entre événements se basent sur trois façons différentes de mesurer la distance :

Espace vide : la distance entre un emplacement fixe de référence dans la fenêtre d'observation et l'événement le plus proche.

Plus proche voisin : la distance entre chaque événement de la fenêtre d'observation et son voisin le plus proche.

Deux à deux : la distance entre chaque paire d'événements dans la fenêtre d'observation.

Selon la distance d'espace vide

La fonction F représente la distribution cumulative de la distance d'espace vide. Pour un processus ponctuel \mathbf{X} observé sur une fenêtre Ω , elle est définie comme :

$$\forall r \in \mathbb{R}^+, F_u(r) = P \left\{ \min_i (\|x_i - u\|, x_i \in \mathbf{X}) \leq r \right\} \quad (3.1)$$

où u est un emplacement de référence arbitraire et P désigne la probabilité. Si le processus est stationnaire, alors cette définition est indépendante de u .

La fonction F empirique des distances d'espaces vides observées à partir des emplacements arbitraires u_j pour $j \in \{1..m\}$ est définie comme :

$$\forall r \in \mathbb{R}^+, \hat{F}(r) = \frac{1}{m} \sum_{j=1}^m \mathbf{1} \left\{ \min_i (\|x_i - u_j\|, x_i \in \mathbf{X}) \leq r \right\} \quad (3.2)$$

Classiquement, cette estimation est comparée avec sa valeur théorique dans le cas où le processus suit strictement une loi de Poisson. Si l'estimation pour une distance r est plus faible que la valeur théorique associée, cela implique que le processus présente des agrégats. Le cas opposé suggère un phénomène de régularité. Cependant cette fonction se base uniquement sur une information locale.

Selon la distance du plus proche voisin

La fonction de distribution cumulative selon la distance du plus proche voisin est la fonction G . Pour un processus ponctuel \mathbf{X} observé sur une fenêtre Ω , elle est définie comme :

$$\forall r \in \mathbb{R}^+, G_x(r) = P \{ \Delta_{\mathbf{X}}(x) \leq r \}, \quad (3.3)$$

avec :

$$\forall x \in \mathbf{X}, \Delta_{\mathbf{X}}(x) = \min_i (\|x - x_i\|, x_i \in \mathbf{X}, x_i \neq x). \quad (3.4)$$

De même que pour la fonction F , l'estimation empirique de la fonction G pour N événements est définie comme :

$$\forall r \in \mathbb{R}^+, \hat{G}(r) = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \{ \Delta_{\mathbf{X}}(x_i) \leq r \}. \quad (3.5)$$

Cette estimation est également comparée à sa valeur théorique associée dans le cas d'un processus de Poisson. Une estimation plus grande que la valeur théorique suggère une agrégation des événements alors qu'une estimation plus faible que la valeur implique une distribution régulière.

Une autre fonction est utilisée qui combine les distances de plus proche voisin et d'espace vide, il s'agit de la fonction J de [van Lieshout et Baddeley \[1996\]](#). Elle est définie comme :

$$\forall r \in \mathbb{R}^+, J(r) = \frac{1 - G(r)}{1 - F(r)}, \quad (3.6)$$

avec $\forall r \in \mathbb{R}^+, F(r) < 1$. Cette fonction est un outil efficace afin de détecter les déviations d'un processus ponctuel par rapport à un processus de Poisson. Dans le cas d'un processus de Poisson, $\forall r \in \mathbb{R}^+, J(r) = 1$ mais la réciproque n'est pas nécessairement vraie, comme le montre [Bedford et Van den Berg \[1997\]](#). Des processus réguliers présentent des valeurs de J supérieures à 1, à l'opposé, la fonction J d'un processus agrégé est inférieure à 1. Cette fonction permet de caractériser l'interaction entre les événements mais également de déterminer son type, son amplitude et sa portée. De plus, elle ne dépend pas de l'intensité du processus contrairement aux fonctions F et G .

Selon la distance deux à deux

Toutes les fonctions associées à cette distance peuvent être construites empiriquement en suivant cinq étapes.

Tout d'abord, les voisins sont comptabilisés autour de chaque événements à et dans une distance r . Dans certains cas, des poids liés aux occurrences d'événements sont additionnés au lieu de simplement les compter. Ensuite, le nombre moyen de voisin est calculé. Troisièmement, ce nombre moyen est divisé par une mesure de référence. Puis, ce rapport local est comparé à sa valeur globale, souvent la densité. Finalement, la signification de cette valeur est testée par rapport à sa valeur théorique dans le cas d'un processus de Poisson.

D'après [Brühlhart et Traeger \[2005\]](#), il est possible d'utiliser trois types de mesures dépendamment de la référence utilisée lors de la troisième étape :

Topographique : en utilisant l'espace comme référence, le nombre de voisins est donc divisé par la surface d'un anneau ou d'un disque.

Relative : le nombre de voisins du processus considérés est divisé par le nombre total de voisins, indépendamment des processus auxquels ils appartiennent.

Absolute : en n'utilisant pas de mesures de référence.

Les fonctions les plus utilisées sont détaillées par la suite.

Fonction de corrélation entre paires. Cette fonction g est une mesure topologique du second ordre. Afin de l'estimer, il convient d'utiliser une fonction noyau donnant un poids à chaque voisin en fonction de sa distance par rapport à celle considérée. Le poids est plus important pour les événements les plus proches de la distance considérée et leur somme doit être égale à 1. La fonction noyau la plus couramment utilisée est également la plus simple ([Illian et al. \[2008\]](#), section 4.3.3), il s'agit d'un noyau uniforme avec un paramètre de bande passante h :

$$\forall h \in \mathbb{R}^{+*}, \forall (d, r) \in \mathbb{R}^{+2}, k(d, r) = \begin{cases} \frac{1}{2h} & \text{si } r - h \leq d \leq r + h \\ 0 & \text{sinon} \end{cases}, \quad (3.7)$$

où d est la distance entre deux événements et r la distance considérée. La fonction g est finalement estimée par :

$$\forall r \in \mathbb{R}^+, \hat{g}(r) = \frac{1}{2\pi r |\Omega| \hat{\lambda}^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N k(\|x_i - x_j\|, r), \quad (3.8)$$

où $\hat{\lambda}$ est la densité du processus dans la fenêtre d'observation Ω et N le nombre total d'événements. La valeur théorique de cette fonction pour un processus de Poisson est de 1. Une valeur plus faible que 1 indique un processus régulier alors qu'une valeur plus grande des agrégats.

Fonction K de Ripley. Ripley [1976] intègre la fonction g sur des distances allant de 0 jusqu'à la distance considérée afin de définir la fonction théorique :

$$\forall r \in \mathbb{R}^+, K(r) = \int_0^r g(\rho) 2\pi\rho \, d\rho. \quad (3.9)$$

Cette fonction peut être estimée empiriquement selon :

$$\forall r \in \mathbb{R}^+, \hat{K}(r) = \frac{1}{|\Omega|\hat{\lambda}^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \mathbf{1}_{\{\|x_i - x_j\| \leq r\}}, \quad (3.10)$$

où $|\Omega|$ est la surface de la fenêtre d'observation Ω . K est une fonction cumulative de g , elle donne donc des informations jusqu'à une certaine distance alors que g fournit uniquement des informations à une distance donnée.

Si le processus ponctuel étudié est homogène, sans dépendance entre ses événements, et est distribué selon une loi de Poisson, sa distribution théorique est appelée *Complete Spatial Randomness* (CSR) et donc :

$$\forall r \in \mathbb{R}^+, K(r) = \pi r^2. \quad (3.11)$$

Si l'estimation est supérieure à la valeur théorique à une distance donnée, le processus ponctuel est dit attractif à cette distance et l'on observe donc des concentrations spatiales. À l'opposé, si l'estimation est inférieure à la valeur théorique, les événements se repoussent les uns les autres menant à un phénomène de régularité.

Comme la fonction K n'est pas simple à visualiser graphiquement et que la référence (πr^2) n'est pas des plus pratiques en termes de comparaison, Besag [1977] a proposé la fonction de l'équation 3.12 de sorte à avoir une valeur de référence nulle quelque soit la distance considérée.

$$\forall r \in \mathbb{R}^+, \hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} - r. \quad (3.12)$$

Dans le cas où plusieurs processus ponctuels sont présents dans la fenêtre d'observation, Lotwick et Silverman [1982] proposent une fonction K permettant d'estimer la distribution d'un processus par rapport à un autre

en comptant les voisins de l'un autour des événements de l'autre, cette fonction est appelée *cross* ou *intertype* fonction K . L'hypothèse de base, qui est que le processus suit strictement une loi de Poisson, n'est pas applicable ici. Deux alternatives sont envisageables. La première est de faire l'hypothèse que les emplacements des événements sont indépendants, cette méthode est appropriée si l'emplacement d'un événement d'un processus dépend uniquement des autres événements de ce même processus. La seconde est de considérer que chaque événement est attribué aléatoirement à un processus, ce qui est approprié lorsque les emplacements des événements sont définis et les processus auxquels ils appartiennent sont attribués a posteriori. Cette hypothèse de remplacement doit être sélectionnée avec soin dépendamment de l'application considérée afin d'éviter de générer des erreurs (Goreaud et Pélissier [2003]).

Synthèse et interprétations de ces fonctions

Le Tableau 3.2 regroupe les méthodes d'interprétations des différentes fonctions descriptives présentées précédemment pour des processus d'intensité λ à une distance r .

TABLE 3.2 – Tableau répertoriant les différentes fonctions descriptives, leurs valeurs théorique en cas de *Complete Spatial Randomness (CSR)* et dans quels cas elles indiquent un phénomène d'agrégats ou de régularité.

Fonction	G	F	J	g	K	L
<i>CSR</i> (ref)	$1 - e^{-\lambda\pi r^2}$	$1 - e^{-\lambda\pi r^2}$	1	1	πr^2	0
Agrégat	> ref	< ref	< ref	> ref	> ref	> ref
Régularité	< ref	> ref	> ref	< ref	< ref	< ref

Les fonctions F et G ne sont pas évidentes à interpréter du fait de leurs valeurs de *CSR*. Afin de bien visualiser pour quelles distances quels phénomènes sont présents, il est recommandé de tracer leurs valeurs empiriques en fonction de leurs valeurs théoriques. Ainsi, l'interprétation est évidente :

- une ligne droite à 45° représente un processus sous *CSR*,
- les points en dessous de cette ligne indiquent un agrégat pour F et de la régularité pour G ,
- ceux au dessus dénotent de la régularité pour F et un agrégat pour G .

La fonction F , basée sur la distance entre événements et points, est efficace pour détecter des phénomènes d'agrégats alors que la fonction G , observant les distances entre événements, repère précisément les réguli-

tés. La fonction J permet de cumuler ces deux informations mais reste bornée à une distance du plus proche voisin. La fonction K a, quant à elle, presque la même efficacité que la fonction F pour les agrégats et que la G pour les régularités (Diggle [1979]) mais permet de les détecter à plusieurs échelles. L'aspect non-cumulatif de la fonction g impacte la précision des estimations, particulièrement pour les courtes distances, et les estimations sont donc considérées moins fiables que celles issues de la fonction K .

3.2.3 Autres fonctions plus génériques dans la littérature

Toutes les fonctions présentées ci-dessus imposent une hypothèse d'homogénéité des processus ponctuels. Identifier les interactions sous une hypothèse de non-homogénéité est une question toujours d'actualité malgré les différentes tentatives proposées jusqu'à présent. La première de ces méthodes, la fonction D proposée par Diggle et Chetwynd [1991], est définie comme étant la différence entre la fonction K estimée sur les événements (nommés *cases*) et la fonction K décrivant d'autres points (nommés *controls*). Cette mesure n'est pas entièrement satisfaisante du fait que les deux fonctions K sont calculées séparément et donc que les informations contenues dans les positions relatives des *cases* et des *controls* ne sont pas prises en compte.

Une autre tentative plus récente est proposée par Baddeley et al. [2000]. Ils généralisent la fonction K pour les processus ponctuels non homogènes mais cette méthode n'est pas très utilisée puisqu'elle impose d'estimer localement l'intensité du processus avec des fonctions noyaux. De ce fait, des hypothèses sont nécessaires quant aux échelles de l'hétérogénéité et des interactions. Or des choix erronés peuvent entraîner des résultats contradictoires avec la réalité, comme le souligne Diggle et al. [2007].

Marcon et Puech [2010] ont proposé une fonction cumulative permettant d'étudier plusieurs processus ponctuels relativement les uns par rapport aux autres sans imposer d'hypothèse d'homogénéité. La fonction M observe le rapport de voisins du type d'intérêt sur le nombre total de voisins. Elle donne la fréquence relative de voisins d'un certain type jusqu'à chaque distance considérée et compare cette fréquence à celle de tous les points, indépendamment de leurs types, à cette même distance :

$$\forall r \in \mathbb{R}^+, \hat{M}(r) = \frac{N - 1}{N_c(N_c - 1)} \sum_{i=1}^{N_c} \frac{\sum_{j=1, j \neq i}^{N_c} \mathbf{1} \{ \|x_j^c - x_i^c\| \leq r \}}{\sum_{l=1, x_l \neq x_i^c}^N \mathbf{1} \{ \|x_l - x_i^c\| \leq r \}}, \quad (3.13)$$

où c représente le type de processus d'intérêt, N_c et N sont respective-

ment le nombre d'événements appartenant à ce processus et le nombre total d'événements dans la fenêtre d'observation, x est un événement d'un type quelconque et x^c un événement du type d'intérêt. Dans le contexte des images histologiques, cette fonction présente l'avantage de pouvoir étudier les relations entre plusieurs populations de cellules sans imposer une hypothèse d'homogénéité, qui n'est pas toujours évidente pour toutes les populations présentes. De plus, puisque c'est une mesure relative, elle n'est pas affectée par l'effet de bord.

3.2.4 Effet de bord et correction

Un problème récurrent des statistiques spatiales est celui de l'effet de bord qui est dû au fait que les processus ponctuels sont analysés uniquement à l'intérieur d'une fenêtre d'observation. Une telle fenêtre correspond à une surface limitée qui peut être de forme et de taille variables. Les voisins d'un événement ayant lieu à proximité des bordures de la fenêtre vont être situés en dehors de la région observée impliquant que les distances entre un tel événement et son voisinage ne peuvent pas être évaluées avec certitude. En conséquence, les fonctions de distributions ne peuvent pas être estimées sans un certain biais introduisant ainsi un effet de bord.

Plusieurs méthodes ont été proposées afin de corriger cet effet, et peuvent être regroupées en deux catégories. Celles qui suppriment les bordures et celles qui opèrent une pondération permettant de prendre en compte la présence d'événements en dehors de la fenêtre d'observation. Cependant, les méthodes définissant une pondération ne sont pas génériques et chaque fonction statistique doit, selon sa définition, adapter le facteur de pondération. Il a été montré par [Barot et al. \[1999\]](#) que la correction de l'effet de bord améliore plus les performances statistiques de la fonction K que celle des autres types de distances. Pour cela, les méthodes avec pondération sont discutées par la suite uniquement pour la fonction de Ripley.

Une comparaison empirique des méthodes de correction de l'effet de bord appliquées à l'estimation de la fonction K peut être trouvée dans les travaux de [Yamada et Rogerson \[2003\]](#).

Zone tampon

Cette première méthode de correction de l'effet de bord présente le double avantage d'être générique et utilisable sur toutes les formes possibles de fenêtres d'observation. Elle consiste en la définition d'une zone tampon à l'intérieur ou à l'extérieur de la région observée. Les événements

contenus dans cette zone ne sont utilisés qu'en tant que voisinage. Typiquement, les zones tampons ont une largeur égale à la plus grande valeur de distance employée dans l'analyse. Une zone tampon intérieure permet d'éviter toutes formes de biais mais entraîne la perte d'informations d'autant plus grande que la distance étudiée est importante. À l'opposé, la version extérieure de cette méthode n'est réalisable que si des données en dehors de la fenêtre d'observation sont accessibles. Malgré ces défauts, cette technique est la seule à permettre de n'utiliser que des événements existant lors de l'estimation des statistiques spatiales tout en étant simple à mettre en œuvre.

Correction toroïdale

Cette deuxième approche générique de correction suggère de faire l'hypothèse que les bordures de la fenêtre d'observation sont connectées à celles qui leurs sont opposées, considérant donc que la région étudiée a la forme d'un tore. Puisque cette méthode se base sur la présomption que le processus ponctuel à l'extérieur de la fenêtre d'observation est le même que celui à l'intérieur, elle intègre un risque potentiel de biais, notamment lorsqu'un agrégat est présent proche de la bordure de la région étudiée. Ainsi, avant d'utiliser cette méthode, il convient de s'assurer que le phénomène étudié satisfait cette hypothèse. De plus, cette technique ne peut être appliquée que pour des fenêtres d'observation de forme rectangulaire. En pratique, elle est réalisée simplement en entourant la fenêtre de base par huit copies identiques et de n'utiliser les points à l'intérieur des duplicatas uniquement en tant que voisinage.

Pondération de Ripley

En parallèle à la méthode permettant d'estimer la fonction K sur une fenêtre d'observation, Ripley [1977] propose une correction de l'effet de bord appropriée et spécifique pondérant les paires d'événements selon leurs positions relatives. Le poids $\omega_{x_i x_j}$ pour une paire d'événements x_i et x_j est défini comme l'inverse de la proportion de la circonférence C d'un cercle centré sur x_i et passant par x_j , ci-après noté $\delta b(x_i, \|x_i - x_j\|)$, qui est incluse dans la fenêtre d'observation Ω :

$$\omega_{x_i x_j} = \frac{C_{\delta b(x_i, \|x_i - x_j\|)}}{C_{\delta b(x_i, \|x_i - x_j\|) \cap \Omega}}. \quad (3.14)$$

La fonction K estimée ainsi corrigée, avec les mêmes notations que l'équation 3.10, est donc définie comme :

$$\forall r \in \mathbb{R}^+, \hat{K}(r) = \frac{1}{|\Omega|\hat{\lambda}^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \omega_{x_i x_j} \times \mathbf{1} \{ \|x_i - x_j\| \leq r \}. \quad (3.15)$$

Cette correction permet d'éviter le biais causé par l'effet de bord lorsque la plus grande distance prise en compte est plus petite que le rayon du cercle inscrit dans la fenêtre d'observation. Cependant, elle ne prend en compte que la distance entre la paire d'événements et les bords de la fenêtre d'observation et non sa forme à proximité des événements.

Pondération de Besag

Sur le même principe que la correction de Ripley, celle de [Besag \[1977\]](#) pondère les paires d'événements par le rapport entre l'aire du disque correspondant au voisinage et celle de sa portion incluse dans la fenêtre d'observation. Elle suppose donc que la portion du disque à l'extérieur de la région étudiée contient la même intensité de voisins que la portion à l'intérieur. En reprenant les notations de l'équation 3.10, la fonction K avec la correction de Besag se note donc :

$$\forall r \in \mathbb{R}^+, \hat{K}(r) = \frac{1}{|\Omega|\hat{\lambda}^2} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \frac{\pi r^2}{|\delta b(x_i, r) \cap \Omega|} \times \mathbf{1} \{ \|x_i - x_j\| \leq r \}, \quad (3.16)$$

où $\delta b(x_i, r)$ est le disque de centre x_i et de rayon r . La correction de Besag intègre les avantages de celle de Ripley mais prend également en compte la forme de la fenêtre d'observation.

3.2.5 Test de Monte Carlo

La méthode de Monte Carlo est utilisée afin de déterminer l'écart entre la réalisation d'un processus ponctuel et son hypothèse nulle. En effet, dans une application réelle, la fonction descriptive estimée sur un processus ponctuel de Poisson n'est jamais égale à la valeur théorique. L'idée de la méthode de Monte Carlo est d'utiliser des données générées par ordinateur pour déterminer l'amplitude des variations prévisibles pour les mesures issues d'un tel échantillon. Cette procédure a été appliquées dans des contextes spatiaux variés par [Ripley \[1977\]](#), [Besag et Diggle \[1977\]](#) et [Diggle \[1983\]](#).

Le test de Monte Carlo est effectué en mesurant la même statistique sur m ensembles de données générés de sorte à respecter l'hypothèse nulle. Puisque ces mesures artificielles représentent la plage de variation attendue pour des données suivant le processus de référence, leurs enveloppes supérieure et inférieure définissent les valeurs limites au delà desquelles le processus dévie de l'hypothèse nulle. Suivant la signification désirée de l'étude et les capacités de calcul à disposition, les enveloppes supérieure et inférieure sont souvent remplacées respectivement par celles correspondant aux 2, 5^{ième} et 97, 5^{ième} centiles pour chaque distance analysée. Finalement la statistique mesurée sur les données réelles est comparée à ces deux enveloppes afin de déterminer si elles respectent ou non l'hypothèse nulle et vers quel type de processus elles dévient.

Plus de simulations sont réalisées et plus le test de Monte Carlo est significatif du fait que les incertitudes dans les enveloppes sont réduites et que ces dernières sont plus lisses. Ainsi, les valeurs des 2, 5^{ième} et 97, 5^{ième} centiles pour chaque distance seront plus précises mais cela ne réduira en rien la largeur des enveloppes. Le nombre de simulations est donc sélectionné par un compromis entre le niveau de précision désiré et le temps de calcul induit. Une recommandation de [Diggle \[1983\]](#) est que $m = 100$ simulations est suffisant dans la mesure où la signification du test au delà de ce point n'augmente que faiblement avec le nombre de simulations.

3.3 Méthode proposée pour la description de processus spatiaux

3.3.1 Données ponctuelles

Fonction statistique robuste

La description recherchée vise à tester si une distribution d'événements ponctuels est aléatoire, agrégée ou régulière en la comparant avec les valeurs des quantiles critiques lui correspondant sous l'hypothèse de *CSR*, voir figure 3.1. L'outil le plus utilisé et recommandé pour cela est la fonction K de Ripley. Le problème lié à l'utilisation de cette fonction est que sa moyenne et sa variance sous l'hypothèse de *CSR* varient avec la distance étudiée r ce qui complexifie son interprétation quantitative. La fonction L proposée par Besag étant centrée mais non normalisée, [Lagache et al. \[2013a\]](#) ont introduit une nouvelle statistique avec une moyenne nulle et une variance unitaire. Cette fonction est définie comme :

$$\forall r \in \mathbb{R}^+, \tilde{K}(r) = \frac{\hat{K}(r) - \pi r^2}{\sqrt{\text{var}\{\hat{K}(r)\}}}. \quad (3.17)$$

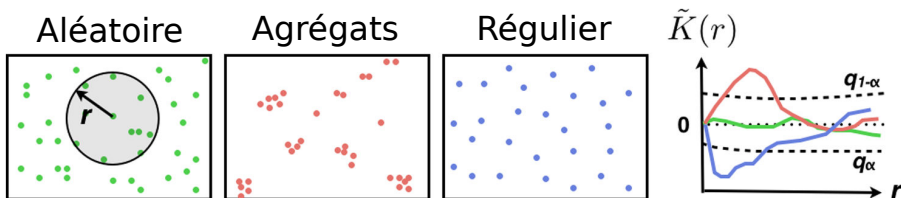


FIGURE 3.1 – Exemples de description de processus ponctuels avec la fonction \tilde{K} robuste. Les trois types de distribution de processus ponctuels sont représentées, à gauche en vert, un processus aléatoire, au centre en rouge, un processus avec agrégats, et à droite en bleu, un processus régulier. Les fonctions \tilde{K} avec les couleurs correspondantes sont représentées dans les courbes à droite ainsi que les deux enveloppes de Monte Carlo estimées par calcul analytique des quantiles critiques q_α et $q_{1-\alpha}$.

La variance de la fonction K sous hypothèse de *CSR* est rendue complexe par l'effet de bords. Cependant, en supposant que les bordures de la fenêtre d'observation Ω sont localement droites au niveau de l'intersection avec le cercle utilisé dans la correction de l'effet de bords de Ripley, une

solution analytique de cette variance a été obtenue par Ripley [1991] et est de la forme :

$$\forall r \in \mathbb{R}^+, \text{var}\{\hat{K}(r)\} = \frac{2|\Omega|^2\beta_1(r)}{N^2} \left(1+0, 305\beta_2(r)+\beta_1(r)(-1+0, 0132N\beta_2(r)) \right), \quad (3.18)$$

où :

$$\forall r \in \mathbb{R}^+, \beta_1(r) = \frac{\pi r^2}{|\Omega|}, \quad (3.19)$$

$$\forall r \in \mathbb{R}^+, \beta_2(r) = \frac{C_\Omega r}{|\Omega|}, \quad (3.20)$$

où C_Ω et $|\Omega|$ sont respectivement le périmètre et la surface de la fenêtre d'observation et N est le nombre d'événements.

Les déviations significatives de \tilde{K} de la valeur nulle sont caractéristiques d'événements agrégés à la distance r quand $\tilde{K}(r) > 0$ ou de régularité quand $\tilde{K}(r) < 0$. Afin de caractériser ces déviations de façon statistique, les enveloppes de Monte Carlo de \tilde{K} sous l'hypothèse de *CSR* doivent être définies. Ces enveloppes sont également appelées valeurs limites et sont ici calculées analytiquement.

Estimation des valeurs limites sous *CSR*

L'estimation empirique des enveloppes de Monte Carlo est très chronophage. Afin de passer outre ce problème, plusieurs travaux récents (Lang et Marcon [2013]) proposent de les déterminer de façon analytique. Ils ont prouvé que, pour N très grand, \tilde{K} peut être approximée, sous *CSR*, à une loi normale $\mathcal{N}(0, 1)$. Cela permet d'estimer q_α , le quantile de \tilde{K} au niveau α , à partir de z_α , le quantile correspondant pour la loi normale $\mathcal{N}(0, 1)$. Cette approximation n'est cependant pas valide pour les valeurs de N qui ne sont pas très grandes, ni pour les petites distances d'analyse. Lagache et al. [2013a] ont proposé une approximation générale des quantiles de \tilde{K} qui est valide pour la plupart des valeurs de distance et de nombre d'événements. Cette approximation se base sur l'approximation de Cornish Fisher (Cornish et Fisher [1938]) et est définie comme :

$$q_\alpha \approx z_\alpha + \frac{1}{6}(z_\alpha^2 - 1)\gamma\{\tilde{K}\} + \frac{1}{24}(z_\alpha^3 - 3z_\alpha)(\kappa\{\tilde{K}\} - 3) - \frac{1}{36}(2z_\alpha^3 - 5z_\alpha)\gamma^2\{\tilde{K}\}, \quad (3.21)$$

où $\gamma\{\tilde{K}\}$ et $\kappa\{\tilde{K}\}$ sont respectivement les coefficients de dissymétrie (*skewness*) et d'aplatissement (*kurtosis*) de \tilde{K} , et z_α est le quantile de la loi normale $\mathcal{N}(0, 1)$ et ses valeurs sont connues (notamment $z_{0,01} = -2,32$ et

$z_{0,99} = 2,32$). Les expressions analytiques du *skewness* et du *kurtosis* en prenant en compte la correction de l'effet de bord de Ripley sont fournies par [Lagache et al. \[2013a\]](#) :

$$\forall r \in \mathbb{R}^+, \gamma\{\tilde{K}(r)\} = \frac{4|\Omega|^3\beta_1(r)}{N^4\text{var}\{K(r)\}^{\frac{3}{2}}} \left[1 + 0,76\beta_2(r) \right. \\ \left. + N\beta_1(r)(1,173 + 0,414\beta_2(r)) \right. \\ \left. + N\beta_1^2(r)(-2 + 0,012N\beta_2(r)) \right], \quad (3.22)$$

$$\forall r \in \mathbb{R}^+, \kappa\{\tilde{K}(r)\} = \frac{|\Omega|^4\beta_1(r)}{N^6\text{var}\{K(r)\}^2} \left[8 + 11,52\beta_2(r) \right. \\ \left. + N\beta_1(r)((104,3 + 12N) + (78,7 + 7,32N)\beta_2(r)) \right. \\ \left. + 1,116N\beta_2^2(r) + N\beta_1^2(r)((-304,3 - 1,92N) \right. \\ \left. + (-97,9 + 2,69N + 0,317N^2)\beta_2(r) + 0,0966N^2\beta_2^2(r)) \right. \\ \left. + n^2\beta_1^3(r)(-36 + 0,0021N^2\beta_2^2(r)) \right]. \quad (3.23)$$

β_1 et β_2 sont définis comme dans l'équation 3.19 et la variance est celle calculée analytiquement dans l'équation 3.18.

3.3.2 Données mixtes

Dans de nombreuses applications, plusieurs types de données sont présents et tous ne sont pas nécessairement modélisables par un processus ponctuel. Ces données particulières sont donc considérées comme étant surfaciques, tel que présenté dans la partie 2.4. Cela implique que les données présentes dans la fenêtre d'observation peuvent être mixtes, au sens où les deux types, ponctuelles et surfaciques, sont représentés. La littérature comporte de nombreuses méthodes pour étudier l'organisation spatiale des processus ponctuels mais aucune à notre connaissance pour les données surfaciques ou pour analyser les interactions entre les deux types. Pourtant des méthodes à même de modéliser les relations complexes entre des structures biologiques et des objets ponctuels sont nécessaires afin de prendre en compte toutes les informations contenues dans les images histologiques. L'un des rares exemples de la littérature proposant de mesurer les interactions entre des données mixtes est appliqué aux images de rétines par [Ruttenberg et al. \[2013\]](#). Cependant, leur méthode impose d'avoir un point de référence pour définir les objets surfaciques tel que le disque optique pour les vaisseaux sanguins dans le fond de l'œil, or ce type de repère n'est généralement pas disponible pour les images histologiques. Une approche similaire, proposée par [Salama et al. \[2013\]](#), mesure le ratio de deux types

de cellules autour des vascularisations à plusieurs distances afin d'observer l'évolution d'une maladie ou sa réaction face à un traitement.

Pour fournir une réponse à ce problème, une fonction inspirée par celle de Ripley est proposée pour l'analyse des interactions entre deux types de données, les processus ponctuels et les données surfaciques. Le principe de base est le même que la fonction d'inspiration mais la forme des données surfaciques est conservée tout au long de l'analyse. L'idée est alors de déterminer le nombre d'événements étant compris à une distance r d'un élément des données surfaciques. Afin de conserver une mesure comparable avec d'autres échantillons de données, la fonction est alors normalisée par rapport à la surface observée et à l'intensité des données ponctuelles dans la fenêtre d'observation. La fonction, baptisée IPA (*Interaction Point Area*) est alors définie comme :

$$\forall r \in \mathbb{R}^+, IPA(r) = \frac{N_e(r)}{S_o(r) \times \lambda_e}, \quad (3.24)$$

où $N_e(r)$ est le nombre d'événements situés dans une distance r des objets surfaciques, $S_o(r)$ est la surface observée et correspond à l'aire du masque des objets surfaciques dilatés par un élément structural circulaire de diamètre r , et λ_e est l'intensité du processus ponctuel. Pour une distance suffisamment grande, la surface observée correspond à la fenêtre d'observation, ce qui, ajouté au fait que c'est une mesure relative, empêche la fonction *IPA* d'être sujette à l'effet de bords. Cette fonction impose d'avoir un processus ponctuel homogène et permet d'avoir des superpositions entre les deux types de données. De plus, elle considère les objets surfaciques comme un tout, ils ne sont donc pas pris en compte séparément, ce qui est approprié dans le cas où les objets surfaciques sont difficilement séparables comme les vascularisations dans les images histologiques. Cette fonction peut être estimée empiriquement selon :

$$\forall r \in \mathbb{R}^+, \widehat{IPA}(r) = \frac{1}{|S_o(r)| \hat{\lambda}_e} \sum_{i=1}^N \mathbf{1} \left\{ \min_{x_s \in S_o(0)} \{\|x_i - x_s\|\} \leq r \right\}. \quad (3.25)$$

De même que pour la fonction \tilde{K} robuste présentée ci-dessus, les valeurs des enveloppes de Monte Carlo sont déterminées analytiquement. Dans le cas où le processus ponctuel X est homogène et respecte l'hypothèse de *CSR*, la probabilité p qu'un événement intervienne à l'intérieur de la surface observée S_o dans la fenêtre d'observation Ω est définie par :

$$\begin{aligned} \forall x \in X, p &= \mathbb{P}(x \in S_o), \\ &= \frac{|S_o|}{|\Omega|}. \end{aligned} \quad (3.26)$$

Ainsi, la probabilité d'avoir un certain nombre d'événement N_e dans S_o parmi les N présents dans la fenêtre d'observation peut être déterminée par la loi binomiale :

$$\forall (N_e, N) \in \mathbb{N}^2, N_e \leq N, \mathbb{P}(N_e \in S_o) = \binom{N}{N_e} p^{N_e} (1-p)^{N-N_e}. \quad (3.27)$$

Pour un N suffisamment grand du point de vue de l'inégalité de Berry Esseen (Berry [1941]), cette loi binomiale tend vers :

$$\mathcal{N}(\mu, \sigma) = \mu + \sigma \times \mathcal{N}(0, 1), \quad (3.28)$$

où :

$$\mu = Np, \quad (3.29)$$

$$\sigma = \sqrt{Np(1-p)}. \quad (3.30)$$

Ainsi, le quantile q_α est ici défini comme :

$$\mathbb{P}(\mu + \sigma \mathcal{N}(0, 1) > q_\alpha) = \alpha. \quad (3.31)$$

Ce qui donne donc :

$$\mathbb{P}(\mu + \sigma \mathcal{N}(0, 1) > q_\alpha) = \mathbb{P}\left(\mathcal{N}(0, 1) > \frac{q_\alpha - \mu}{\sigma}\right). \quad (3.32)$$

La valeur $\frac{q_\alpha - \mu}{\sigma}$ correspond alors au quantile connu d'ordre α de la loi normale $\mathcal{N}(0, 1)$, ci-après noté z_α . On a donc :

$$\frac{q_\alpha - \mu}{\sigma} = z_\alpha \Leftrightarrow q_\alpha = \mu + \sigma z_\alpha. \quad (3.33)$$

Avec les équations 3.26, 3.29 et 3.30 , le quantile q_α est finalement défini comme :

$$q_\alpha = N \frac{|S_o|}{|\Omega|} + z_\alpha \times \sqrt{N \frac{|S_o|}{|\Omega|} \left(1 - \frac{|S_o|}{|\Omega|}\right)}. \quad (3.34)$$

Les deux quantiles critiques q_α et $q_{1-\alpha}$, avec α dépendant de la précision recherchée pour l'étude, correspondent aux enveloppes de Monte Carlo définissant l'écart minimal pour que le nombre d'événements $N_e(r)$ situés à une distance r des objets surfaciques respecte l'hypothèse de *CSR*. Les quantiles de la fonction *IPA* sous hypothèse de *CSR* sont finalement définis comme :

$$\begin{aligned} \forall r \in \mathbb{R}^+, q_{\alpha IPA}(r) &= \frac{q_\alpha(r)}{|S_o(r)|\lambda_e}, \quad (3.35) \\ &= 1 + \frac{z_\alpha}{|S_o(r)|\lambda_e} \times \sqrt{\lambda_e |S_o(r)| \left(1 - \frac{|S_o(r)|}{|\Omega|}\right)}. \end{aligned}$$

Si les deux types de données sont indépendants, la valeur de IPA vaut théoriquement 1. D'un point de vue pratique cela correspond à des valeurs comprises entre les deux quantiles critiques q_α et $q_{1-\alpha}$. Autrement, il existe une dépendance de type attraction pour des valeurs supérieures et de type répulsion dans le cas contraire.

3.3.3 Regrouper les statistiques de plusieurs sources

De nombreux cadres applicatifs fournissent plusieurs échantillons d'une même réalisation d'un processus, c'est notamment le cas dans l'analyse d'images histologiques, en particulier avec les micro-vecteurs de tissus ou lorsque qu'une même lame histologique est représentée par un groupe d'images capturées aléatoirement ou non. Dans ces cas de figure, plusieurs fenêtres d'observation sont donc disponibles et deux stratégies peuvent être considérées afin de les prendre en compte. Soit les différentes fenêtres sont prises en compte indépendamment les unes des autres et donc traitées comme montrant des processus différents, soit elles doivent être agrégées afin de cumuler les informations qu'elles contiennent concernant la réalisation d'un même processus. Le but est alors d'obtenir une description globale en cumulant les mesures statistiques de chaque échantillon.

Soient m fenêtres d'observation $(\Omega_i)_{i \in \{1..m\}}$ de formes similaires d'un même processus. Les mesures statistiques sont calculées sur chaque fenêtre séparément puis regroupées pour obtenir une estimation globale. La densité $\hat{\lambda}$ d'un processus ponctuel homogène est obtenue à partir des m densités $\hat{\lambda}_m$ issues des m fenêtres d'observation selon l'équation :

$$\hat{\lambda} = \sum_{i=1}^m \hat{\lambda}_i \frac{|\Omega_i|}{|\Omega|}, \quad (3.36)$$

où :

$$|\Omega| = \sum_{i=1}^m |\Omega_i|. \quad (3.37)$$

De façon similaire, la fonction \tilde{K} globale peut être retrouvée à partir des estimations \tilde{K}_i mesurées sur les m fenêtres d'observation suivant la formule :

$$\forall r \in \mathbb{R}^+, \tilde{K}(r) = \sum_{i=1}^m \tilde{K}_i(r) \frac{N_i}{N}, \quad (3.38)$$

où N_i est le nombre d'événements dans la fenêtre d'observation Ω_i , et N est le nombre total d'événements observés dans l'ensemble des fenêtres.

De même, les fonctions \hat{F} et \hat{G} peuvent être agrégées respectivement à partir des \hat{F}_i et \hat{G}_i estimées sur les m fenêtres d'observation :

$$\forall r \in \mathbb{R}^+, \quad \hat{F}(r) = \sum_{i=1}^m \hat{F}_i(r) \frac{|\Omega_i|}{|\Omega|}, \quad (3.39)$$

$$\forall r \in \mathbb{R}^+, \quad \hat{G}(r) = \sum_{i=1}^m \hat{G}_i(r) \frac{N_i(r)}{N(r)}, \quad (3.40)$$

où $N_i(r)$ est le nombre d'événements dans le voisinage de rayon r dans la fenêtre d'observation Ω_i , et $N(r)$ est le nombre total d'événements dans ce même voisinage sur l'ensemble des fenêtres.

Les justifications mathématiques de ces formules sont fournies dans le livre de [Illian et al. \[2008\]](#) au chapitre 4.

En répétant les mêmes procédures, la fonction \widehat{IPA} peut être retrouvée suivant :

$$\forall r \in \mathbb{R}^+, \quad \widehat{IPA}(r) = \sum_{i=1}^m \widehat{IPA}_i(r) \frac{N_i}{N}. \quad (3.41)$$

3.4 Application aux images histologiques

Les méthodes proposées pour l'analyse spatiale statistique de données ponctuelles et surfaciques ont été appliquées sur les images histologiques de la base de données introduite dans la partie 1.5. De même que dans les expérimentations de la partie 2.5, les noyaux des cellules sont considérés comme la réalisation d'un processus ponctuel alors que les vascularisations sont prises en compte comme des données surfaciques. Les fenêtres d'observation ici envisagées correspondent aux images. D'autres fonctions statistiques pour les processus ponctuels sont également mesurées afin de comparer les capacités des différentes approches.

3.4.1 Statistiques pour les processus ponctuels

Protocole expérimental

L'objectif biomédical est ici d'étudier l'organisation spatiale des noyaux cellulaires afin d'analyser les différences éventuelles entre les modèles d'implantation ectopique et orthotopique de même que les mécanismes d'évolution du cancer en observant les possibles variations selon les jours après implantation. Le but mathématique est d'observer les informations capturées par chacune des fonctions mesurées et d'interpréter leurs différences.

Les positions des noyaux servent de processus ponctuels et sont extraites avec la méthode proposée dans la partie 2.4.2. La figure 3.2 montre un exemple de cette détection sur une image représentative. L'observation visuelle de la position des événements sur les détections issues de l'ensemble de la base d'images rend plausible les hypothèses d'isotropie et de stationnarité.

Fonctions F , G et J

Les deux fonctions F et G sont estimées suivant les équations 3.2 et 3.5 sur l'ensemble des images de la base. Leurs interprétations n'étant pas triviales, la figure 3.3 montre les valeurs des deux fonctions issues de l'image représentative de la figure 3.2 en fonction de leurs valeurs théoriques pour un processus aléatoire avec les enveloppes de Monte Carlo réalisées avec 100 simulations selon la méthode discutée dans la partie 3.2.2. La fonction F est estimée en générant 500 points de contrôle selon une distribution de Poisson. Les valeurs de F sont au dessus des enveloppes de Monte Carlo tandis que celles de G sont en dessous, ainsi les deux fonctions indiquent

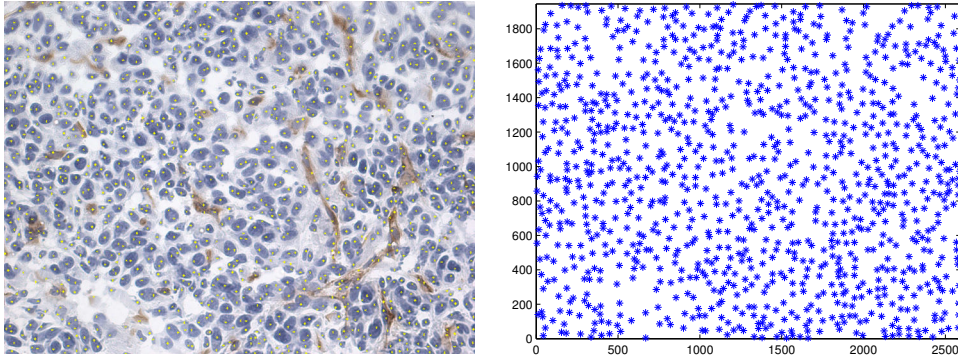


FIGURE 3.2 – Exemple d’un ensemble d’évènements utilisé pour estimer les statistiques spatiales. L’image de gauche montre les données originales avec superposition des points détectés alors que celle de droite montre uniquement le nuage d’évènements.

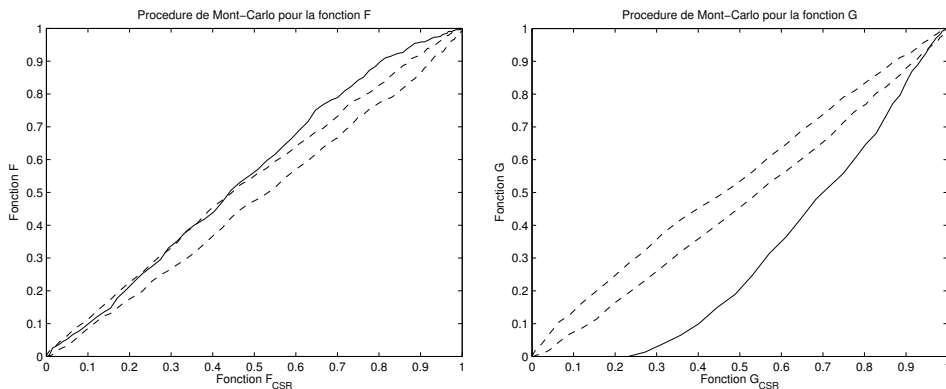


FIGURE 3.3 – Exemple de deux fonctions F et G tracées en fonction de leurs valeurs théoriques de référence dans le cas de la réalisation d’un processus de Poisson avec les enveloppes de Monte Carlo correspondantes dessinées en pointillés.

la présence de régularités dans la distribution des noyaux cellulaires. Ces deux courbes sont cohérentes avec le fait que la fonction G est plus à même de détecter les phénomènes de régularité (et F les agrégats) puisque la fonction G ici mesurée indique un écart plus important avec l’hypothèse de CSR alors que le résultat de la fonction F est plus mitigé.

Cette technique de présentation permet de visualiser rapidement l’écart entre un processus ponctuel et l’hypothèse de CSR mais n’est pas utile pour comparer plusieurs mesures de processus différents en fonction de la distance considérée. Pour cela, la figure 3.4 montre les fonctions F et G estimées sur l’ensemble de la base d’image selon le modèle et le jour après implantation.

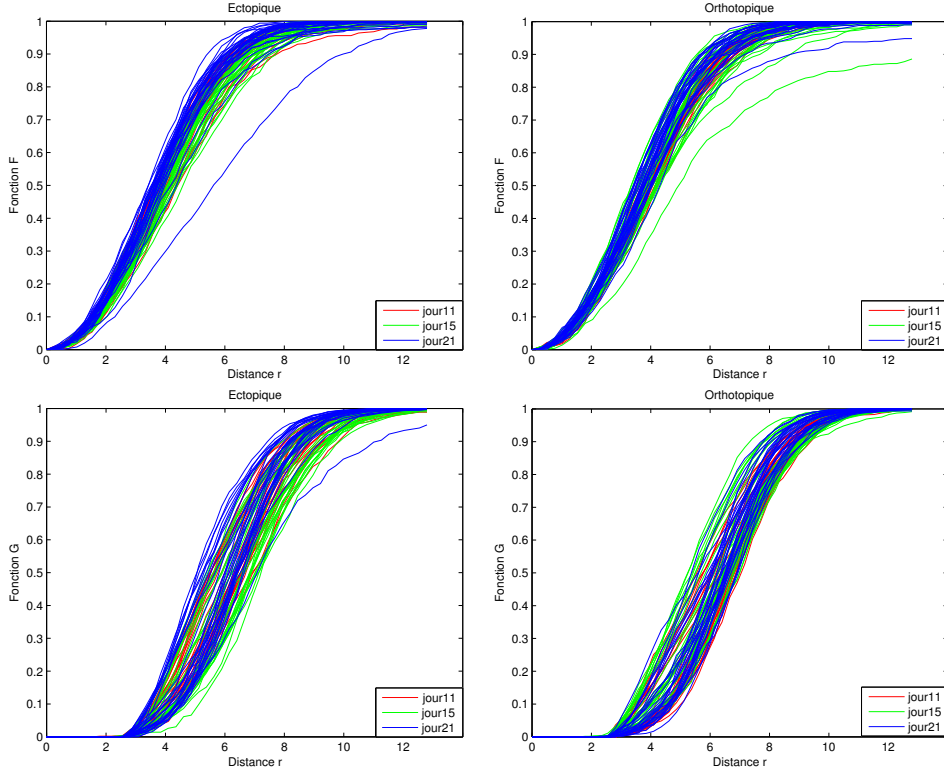


FIGURE 3.4 – Fonctions F et G mesurées selon les modèles et les jours après implantation. La première ligne montre les fonctions F tandis que la seconde contient les fonctions G . La distance r est donnée en micromètres.

Ces figures montrent un très fort chevauchement des courbes indépendamment du modèle et du jour après implantation. De ce fait, elles ne permettent pas d'établir simplement la présence de différences dans la distribution des noyaux des cellules. Un test statistique est nécessaire afin de décréter si des différences sont présentes ou non.

Puisque les deux fonctions F et G sont estimées sur la base, il est possible de calculer les fonctions J correspondantes. La figure 3.5 montre les fonctions J mesurées pour les différents modèles et jours après implantation. De même que les mesures précédentes, les fonctions J révèlent une régularité dans la distribution des événements. Contrairement aux fonctions F et G et bien qu'elles montrent de nombreuses superpositions, des différences plus marquées sont présentes entre les courbes. Certaines courbes s'arrêtent avant d'atteindre une distance de $13\ \mu m$, cela est dû au fait que les fonctions F estimées empiriquement sont susceptibles d'atteindre la valeur limite 1. Les valeurs des fonctions J dans ce cas sont donc indéfinies et non interprétables.

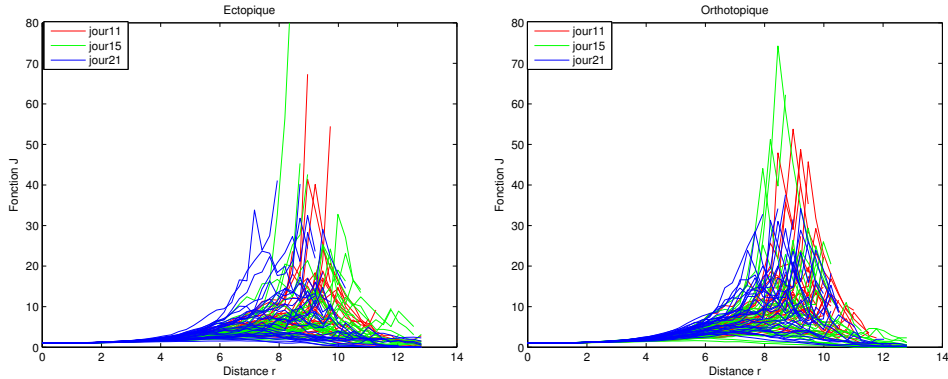


FIGURE 3.5 – Fonctions J mesurées selon les modèles et les jours après implantation. La distance r est donnée en micromètres.

Fonction \tilde{K} robuste

Les fonctions \tilde{K} robustes de l'ensemble des images de la base ont été estimées suivant l'équation 3.17 avec la correction de l'effet de bord de Ripley. Deux approches sont proposées précédemment afin de vérifier si l'écart entre la valeur de la fonction \tilde{K} robuste et celle théorique d'un processus respectant l'hypothèse de *CSR* est significatif, le test de Monte Carlo et le calcul analytique de ses quantiles. La figure 3.6 montre les fonctions issues de 100 simulations sous *CSR* avec les enveloppes de Monte Carlo correspondantes ainsi que les valeurs analytiques des quantiles critiques $q_{0.01}$ et $q_{0.99}$. Un exemple de fonction \tilde{K} robuste avec les deux enveloppes est également présenté.

Cette figure montre bien que les enveloppes obtenues analytiquement sont plus lisses et très proches de celles déterminées empiriquement. Les enveloppes rouges correspondent ici aux valeurs minimales et maximales et sont donc affectées par les cas extrêmes. Prendre une marge d'erreur tend à rapprocher ces courbes expérimentales des valeurs analytiques. Cette expérience valide donc l'utilisation du calcul analytique des quantiles. La fonction \tilde{K} robuste estimée sur l'image représentative de la figure 3.2 avec les enveloppes permet de tirer des conclusions quant à la distribution des noyaux. Elle montre une forte régularité pour une distance de $5\ \mu m$ qui correspond au diamètre moyen d'une cellule dans ces images. De même, à partir d'une distance suffisante, de 20 ou $25\ \mu m$, la distribution des noyaux est comprise entre les deux quantiles et donc respecte l'hypothèse de *CSR*.

La figure 3.7 montre les courbes ainsi obtenues pour les différents modèles et jours après implantation ainsi que leurs enveloppes (valeurs maximales et minimales) afin de souligner les superpositions et les variations. Les différentes fonctions estimées ont toutes le même comportement et se

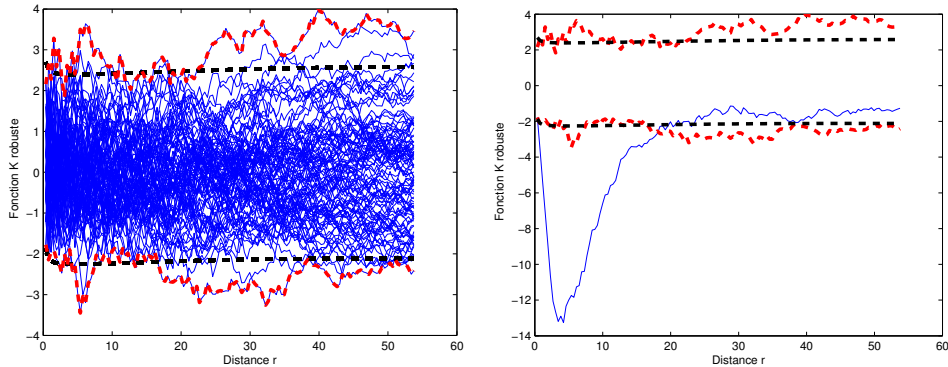


FIGURE 3.6 – Comparaison entre les enveloppes issues du test de Monte Carlo et leurs valeurs obtenues analytiquement pour la fonction \tilde{K} robuste. À droite les fonctions issues de 100 simulations d’un processus de Poisson et à gauche la fonction mesurée sur l’image représentative de la base (voir figure 3.2). En pointillés rouges sont tracées les enveloppes de la procédure de Monte Carlo et en noirs les quantiles calculés analytiquement. La distance r est donnée en micromètres.

chevauchent. Les enveloppes permettent de voir des tendances différentes suivant les jours et les modèles mais l’utilisation d’un test statistique est tout de même nécessaire afin de confirmer ou d’infirmer la présence de différences significatives entre ces fonctions.

3.4.2 Statistiques pour les données mixtes

Il a été montré qu’étudier les interactions entre les cellules et les vaisseaux sanguins peut apporter des informations pertinentes afin de mieux comprendre les mécanismes du cancer (Breward et al. [2001]). La structure des tumeurs est hétérogène au sens où les cellules cancéreuses proches des vaisseaux sanguins prolifèrent rapidement grâce à la présence d’oxygène et de nutriments et que lorsque la vitesse de prolifération dépasse celle de création de nouvelles vascularisations, les cellules meurent d’hypoxie (Weidner [1999]). Les vaisseaux sanguins présents dans une tumeur peuvent être répartis en deux catégories, ceux qui faisaient déjà partie du tissu avant l’arrivée du cancer et les nouveaux vaisseaux (néovascularisations) apparaissant dans la tumeur. Les premiers sont plus solides et rigides que les nouveaux vaisseaux et ont tendance à être régulièrement espacés. À l’opposé, les néovascularisations présentent des fuites et sont plus sinueuses. De part leur immaturité, ces vaisseaux récents ne supportent que des faibles augmentations de pression. De ce fait, une augmentation causée par la prolifération des cellules cancéreuses peut causer la compres-

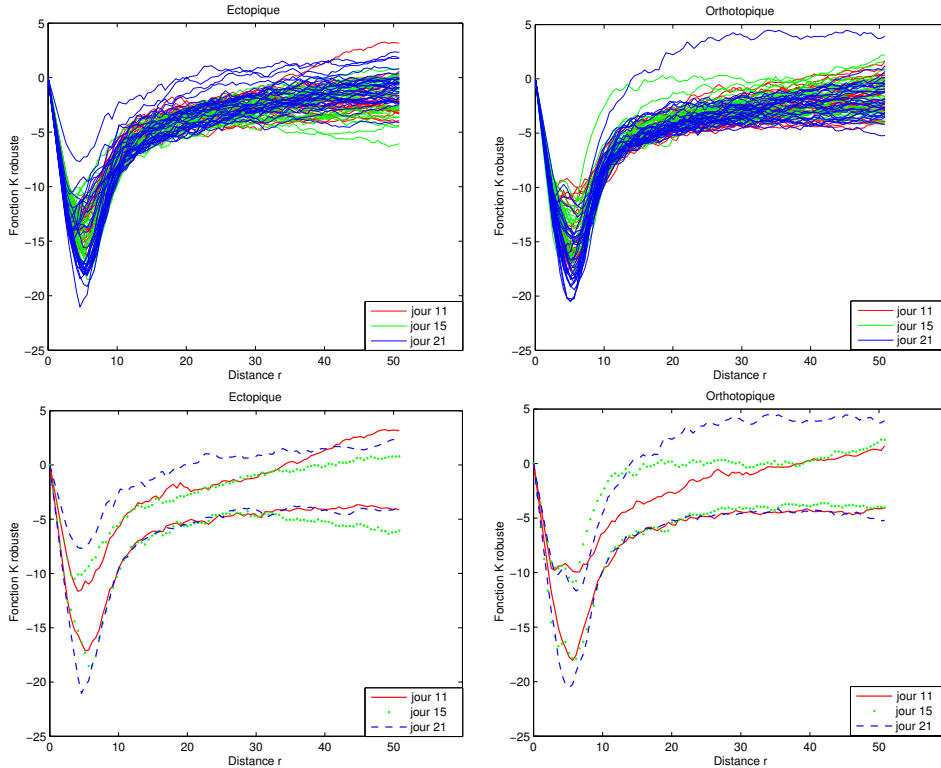


FIGURE 3.7 – Fonctions K robustes mesurées selon les modèles et les jours après implantation. La première ligne montre les fonctions tandis que la seconde contient leurs enveloppes suivant les jours. La distance r est donnée en micromètres.

sion ou l'effondrement des néovascularisations, réduisant ainsi ou arrêtant complètement l'apport en oxygène et en nutriments à son entourage.

Afin de mesurer les éventuelles interactions entre les noyaux cellulaires (données ponctuelles) et les vaisseaux sanguins (données surfaciques) dans l'ensemble de données histologiques à disposition, la fonction IPA , présentée dans la partie 3.3.2, est estimée. Afin de juger d'un écart significatif avec l'hypothèse de non interaction, les enveloppes obtenues analytiquement pour les quantiles $q_{0.01}$ et $q_{0.99}$ avec l'équation 3.35 sont utilisées. La figure 3.8 montre que ces deux enveloppes sont pertinentes par rapport à cent processus ponctuels générés de sorte à respecter l'hypothèse de CSR et étant indépendant des données surfaciques de l'image considérée. L'image ici utilisée est celle représentant un modèle ectopique dans la figure 3.9. Les données surfaciques prises en compte sont les vascularisations présentes et les processus ponctuels sont générés avec la même densité que les noyaux cellulaires.

De même que pour les enveloppes de la fonction K robuste, cette figure

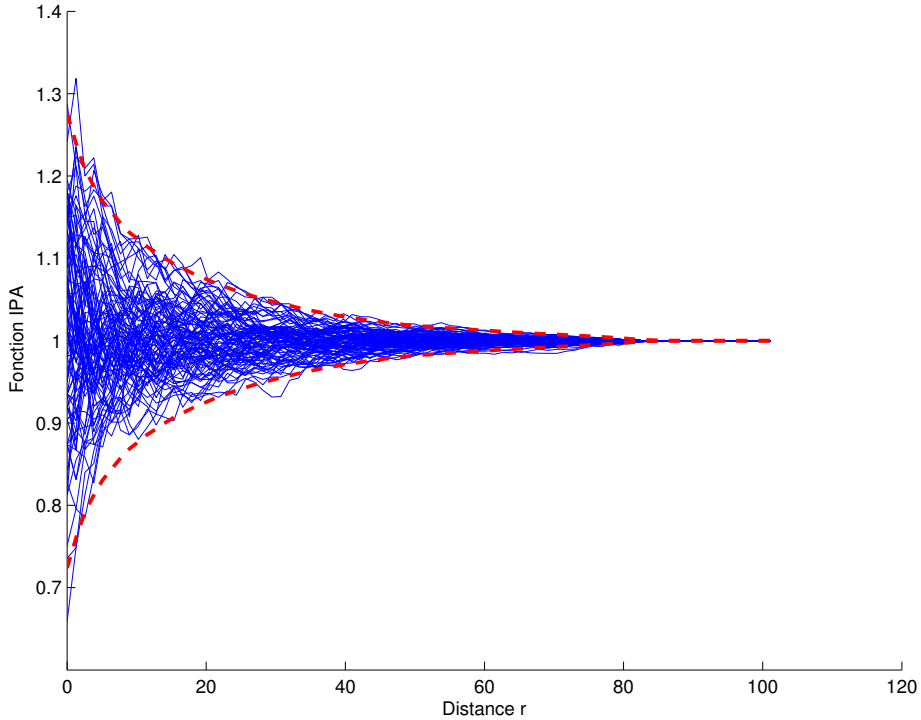


FIGURE 3.8 – Comparaison entre les enveloppes obtenues analytiquement pour la fonction *IPA* (en pointillés rouges) et 100 simulations d'un processus de Poisson indépendant des données surfaciques (en bleu). La distance r est en micromètres.

montre que les valeurs limites déterminées analytiquement correspondent bien à celles obtenues par la procédure de Monte Carlo. Cela valide donc l'utilisation des enveloppes obtenues par calcul analytique qui permet un gain de temps considérable tout en offrant des enveloppes plus lisses. De plus, ce résultat confirme expérimentalement les calculs mathématiques de la section 3.3.2.

La figure 3.9 montre les fonctions *IPA* estimées pour deux images représentatives de la base, l'une ectopique et l'autre orthotopique. Ces images ont été sélectionnées afin de mettre en évidence les deux formes de dépendances possibles entre les données surfaciques et ponctuelles, l'attraction et la répulsion. Les écarts significatifs de l'hypothèse d'indépendance ont lieu pour de courtes distances, jusqu'à environ $10 \mu m$, attestant de la présence d'une interaction entre les deux objets biologiques à cette échelle. La fonction *IPA* de l'image ectopique a des valeurs supérieures à l'écart limite, révélant ainsi une attraction et donc une tendance des noyaux à s'agglomérer autour des vaisseaux à courte distance alors que celles de l'image orthotopique sont inférieures au quantile limite montrant une ré-

pulsion entre les objets à cette même échelle. Ces résultats sont cohérents avec le fait que la structure des tumeurs est hétérogène et présente plusieurs comportements possibles entre les cellules et le réseaux sanguin. En prenant en considération les fonctions *IPA* mesurées sur l'ensemble de la base de données (voir figure 3.10), ceci confirme que des variations locales de comportement sont présentes au sein d'une même tumeur.

Sur la figure 3.10, toutes les fonctions *IPA* estimées sur la base de données sont tracées et regroupées selon le modèle et le jour après implantation. Malgré un fort chevauchement des différentes courbes des tendances peuvent être observées. Pour le modèle ectopique, les fonctions correspondant au jour 21 tendent vers un phénomène de répulsion, alors que la majorité des courbes montrent une indépendance entre les objets à quelques exceptions près attestant d'une attraction, en particulier pour le jour 11. Dans le cas du modèle orthotopique, le comportement est similaire, toutefois plus de fonctions estimées sur le jour 15 ont tendance à révéler une attraction. En dépit de ces propensions, la présence de différences significatives entre les modèles et les jours reste à confirmer ou infirmer via l'utilisation d'un test statistique comme pour les fonctions \tilde{K} robustes.

Dans le cas de l'étude des relations entre les noyaux et les vascularisations, cette fonction permet donc d'étudier la phase (prolifération ou hypoxie) dans laquelle se trouve le tissu localement. Pour une étude plus précise, elle peut être appliquée en ne considérant qu'un unique vaisseau sanguin à la fois et donc déterminer le comportement des cellules vis-à-vis de cet objet en particulier dans le cadre d'un test clinique.

3.4.3 Agrégation des fonctions issues de plusieurs images

La base de données utilisée contient dix images capturées aléatoirement dans la même lame histologique pour chaque souris. Il est donc envisageable de considérer toutes les images issues d'une même souris comme plusieurs fenêtres d'observation des mêmes données ponctuelles et surfaciques. Les formules de la partie 3.3.3 sont donc utilisées afin de regrouper les différentes fonctions en une unique description par souris.

Les figures 3.11 et 3.12 montrent respectivement les agrégations des fonctions F et G de chaque souris tracées selon les modèles et jours après implantation. Les deux fonctions montrent encore de forts chevauchements rendant complexes toute interprétation.

De même, la figure 3.13 montre les regroupements des fonctions \tilde{K} robustes selon les différents modèles et jours après implantation. Comme pour

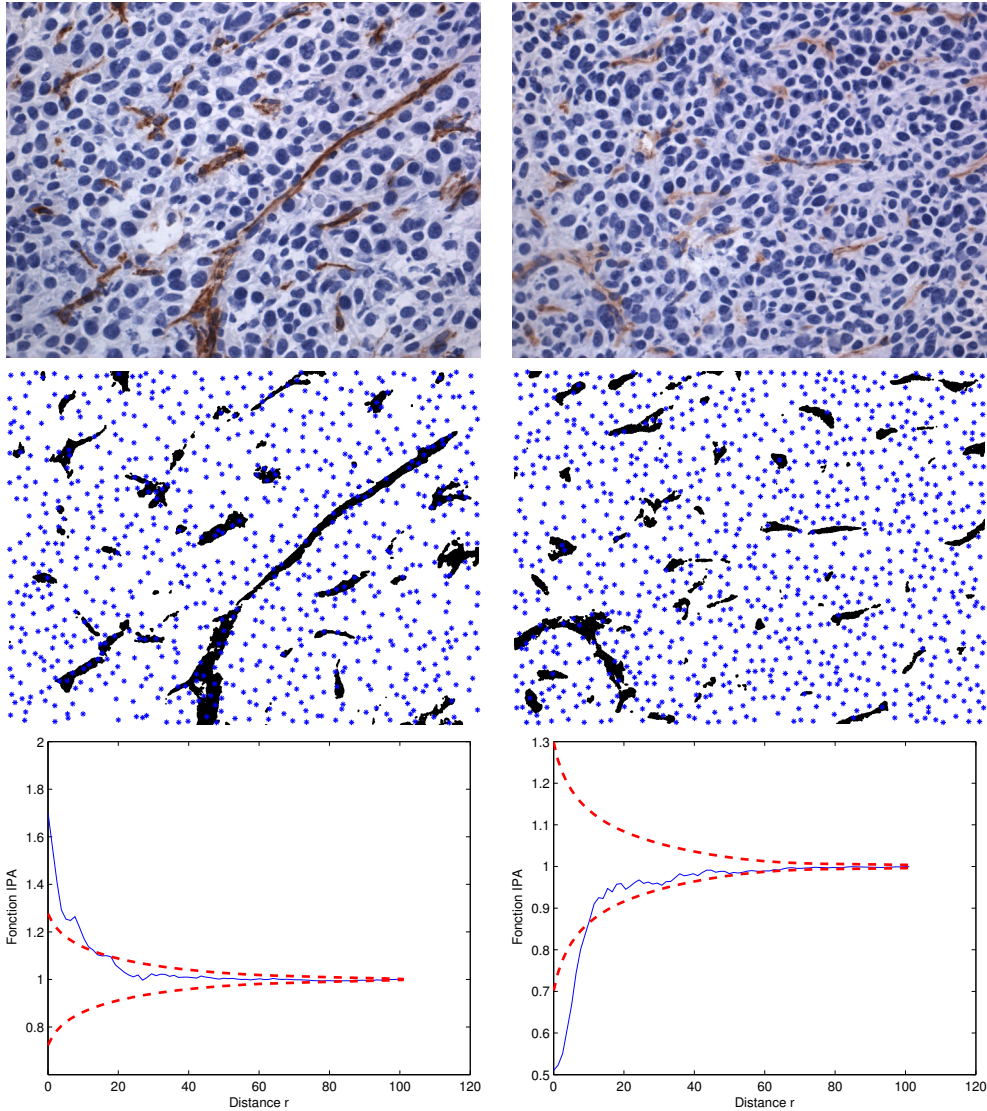


FIGURE 3.9 – Exemple de l'estimation de la fonction IPA sur deux images représentatives. À gauche, une image du modèle ectopique révélant une attraction entre les noyaux et les vascularisations, et à droite une image du modèle orthotopique illustrant une répulsion. Les images de la deuxième ligne montre l'extraction des vaisseaux sanguins en noir et les points correspondant aux cellules en bleu. La troisième ligne contient les fonctions IPA estimées avec en pointillés rouges les enveloppes critiques obtenues analytiquement. La distance r est donnée en micromètres.

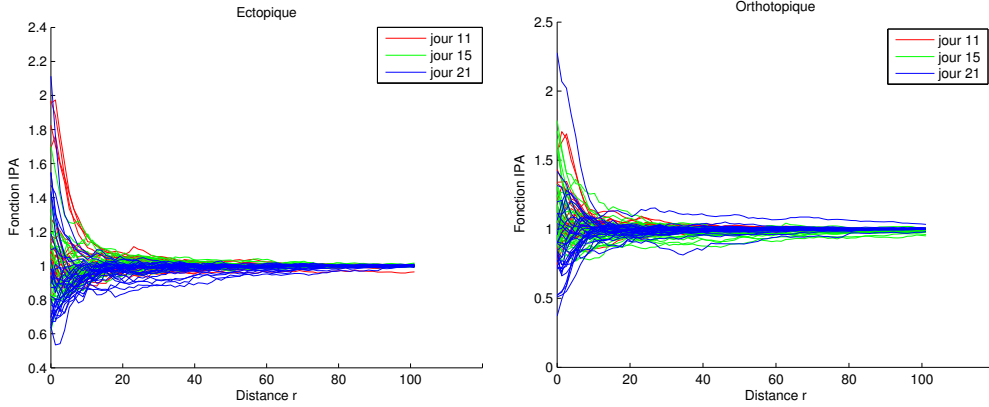


FIGURE 3.10 – Fonctions IPA mesurées selon les modèles et les jours après implantation. La distance r est donnée en micromètres.

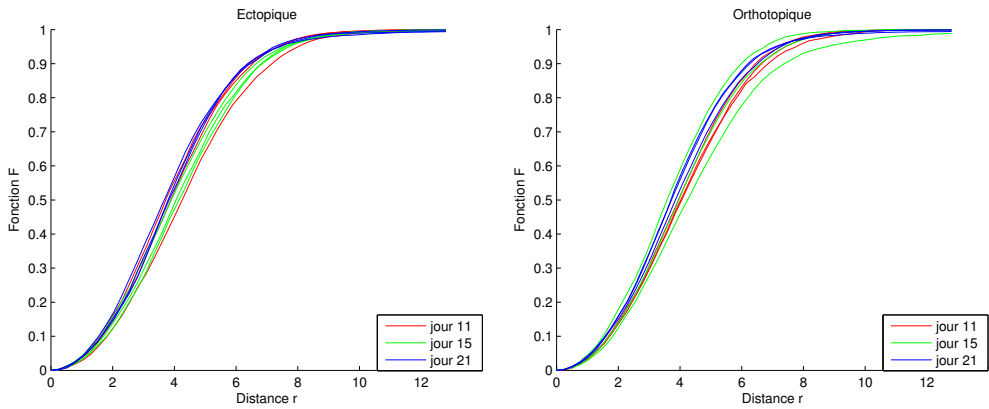


FIGURE 3.11 – Regroupement des fonctions F par souris et affichage selon les jours et les modèles. La distance r est donnée en micromètres.

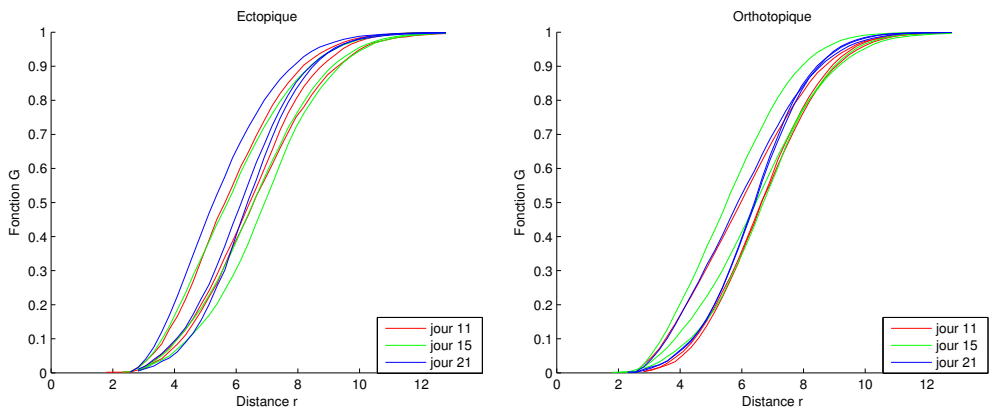


FIGURE 3.12 – Regroupement des fonctions G par souris et affichage selon les jours et les modèles. La distance r est donnée en micromètres.

les fonctions F et G , de forts chevauchements sont encore présents mais quelques tendances apparaissent, notamment de plus fortes régularités à très courtes distances pour les jours 21 des deux modèles. Cependant, avec seulement trois souris par jour et par modèle, les données ne sont pas suffisamment nombreuses pour tirer des conclusions concernant l'application biologique.

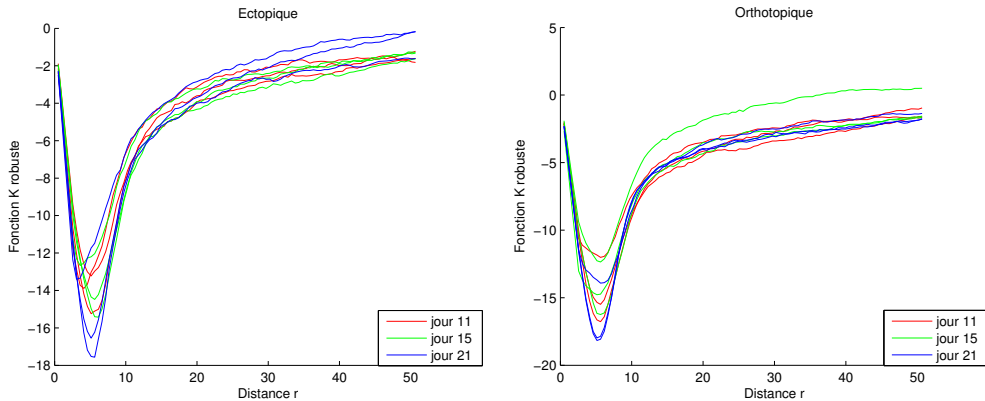


FIGURE 3.13 – Regroupement des fonctions \tilde{K} robustes par souris et affichage selon les jours et les modèles. La distance r est donnée en micromètres.

Finalement, les fonctions IPA sont regroupées via l'équation 3.41 et tracées dans la figure 3.14. Encore une fois, le manque de données ne permet pas de tirer de conclusions, cependant seuls les jours 21 des deux modèles montrent des phénomènes de répulsion, alors que les jours 11 révèlent des indépendances et que les jours 15 montrent des différences entre les deux modèles. En effet, pour le modèle ectopique, les fonctions IPA des jours 15 décrivent une indépendance entre les deux types d'objets alors que pour le modèle orthotopique, elle traduit une tendance à l'attraction.

Effectuer un regroupement des fonctions issues des différentes fenêtres d'observation peut altérer les mesures et n'est donc pas nécessairement la meilleure solution. Une fenêtre d'observation contenant un cas extrême ou aberrant peut biaiser la conclusion de l'ensemble de fenêtres. Ceci est illustré dans le cas de la fonction IPA pour les souris orthotopiques au jour 21 après implantation. En effet, dans la figure 3.10 seule une image présente un phénomène d'attraction, alors qu'une fois les différentes fenêtres d'observation regroupées, voir figure 3.14, la souris correspondante tend vers une attraction.

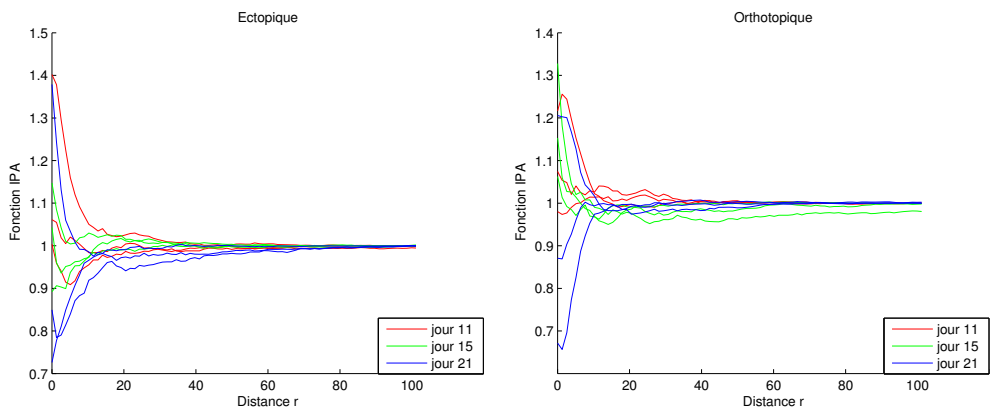


FIGURE 3.14 – Regroupement des fonctions *IPA* par souris et affichage selon les jours et les modèles. La distance r est donnée en micromètres.

3.5 Conclusion et perspectives

L'étude de la distribution spatiale des objets biologiques et leurs interactions dans les images histologiques permet d'extraire des informations importantes afin de bien comprendre les mécanismes sous-jacents des tissus, qu'ils soient malades ou non. La littérature intègre déjà de nombreuses méthodes permettant de modéliser l'organisation spatiale d'objets ponctuels, chacune ayant ses spécificités. Cependant, l'observation des interactions entre objets ponctuels et surfaciques est permise par peu de méthodes. Ce chapitre recommande l'utilisation de la fonction \tilde{K} robuste pour observer les objets modélisables par des points et présente une fonction statistique innovante afin d'étudier les interactions entre deux types de données différents ainsi que les méthodes permettant d'agréger des mesures issues de plusieurs échantillons d'une même source. Finalement, ces méthodes sont validées sur un ensemble de données histologiques dans le contexte de différencier deux modèles d'un même cancer.

Une piste de recherche à poursuivre serait de tester si l'organisation spatiale d'un ensemble d'événements peut être décrite par des modélisations paramétriques spécifiques, telles que les processus de [Neyman et Scott \[1952\]](#) ou encore de [Strauss \[1975\]](#); ce qui permettrait d'effectuer une comparaison analytique de plusieurs distributions. Une modélisation des cellules par processus de Strauss est effectuée par [Mattfeldt et al. \[2007\]](#). Utiliser une telle modélisation offre également la possibilité de rechercher des événements ou des configurations insolites. Cela inclut la détection d'événements aberrants par rapport à la construction du motif ponctuel ou encore d'événements manquants d'après la modélisation. Pour de plus amples informations sur les principes mathématiques soutenant ces techniques, il est conseillé de se référer à la section 4.6 du livre de [Illian et al. \[2008\]](#). Ces méthodes semblent prometteuses pour l'étude des tissus cancéreux, notamment dans les cas précoces, où les cellules affectées par la maladie ne respectent plus la distribution standard d'un tissu sain.

Une autre perspective consiste en la mesure de la distribution des objets surfaciques. Pour cela la fonction IAA (*Interaction Area Area*) est définie comme :

$$\forall r \in \mathbb{R}, IAA(r) = \frac{1}{N\lambda_{\text{area}}} \sum_{i=1}^N \frac{S_{\text{area} \setminus S_i} \cap S_o(r)}{S_o(r)}, \quad (3.42)$$

où N est le nombre d'objets surfaciques, S_o la surface d'observation définie comme dans la partie 3.3.2, $S_{\text{area} \setminus S_i}$ la surface des objets dans la fenêtre d'observation à l'exception de l'objet i considéré et λ_{area} est l'intensité des objets surfaciques. Cependant, cette fonction n'a pu être validée par manque de données et de vérité terrain, d'où sa description en perspectives.

Analyse de la texture dans les images biomédicales

Les caractéristiques de texture permettent de quantifier la finesse, le contraste, les variations d'intensité, les discontinuités et les contours d'une image. En conséquence, un grand nombre de méthodes se basent sur l'analyse de la texture afin d'analyser automatiquement les images médicales. Ce chapitre traite donc de ces techniques et discute de leurs efficacités pour l'aide au diagnostic. Un état de l'art des méthodes les plus couramment utilisées est présenté dans la partie 4.2, la méthode des motifs binaires locaux est ensuite étudiée en détails dans la section 4.3 conjointement à une approche permettant de maximiser la pertinence de l'information mesurée. Finalement deux applications sont considérées, ayant pour cadre les images de rétine et les images histologiques dans la partie 4.4.

Les travaux de ce chapitre ont été réalisés lors d'un stage de six mois au Laboratoire Image et Vision 4D (LIV4D) à l'École Polytechnique de Montréal.

4.1 Motivation

L'analyse de la texture est très utilisée dans les images médicales afin de détecter la présence d'une maladie ou encore l'état dans lequel celle-ci se trouve. Malgré la complexité de leur interprétation biologique, un nombre considérable de méthodes se basent sur les caractéristiques de textures pour l'aide au diagnostic en raison de leur efficacité descriptive. En particulier, le lien entre les caractéristiques de texture d'un tissu biologique et le degré de malignité d'une tumeur a été confirmé par plusieurs études ([Tabesh et al. \[2007\]](#)). Par exemple, les matrices de co-occurrences sont utilisées avec succès par [Hamilton et al. \[1997\]](#) afin de reconnaître les différents types de tissus biologiques dans les images de cancer colorectal. Les motifs binaires locaux sont également employés conjointement à une transformée en paquets d'ondelettes pour faire de même par [Qureshi et al. \[2008\]](#) sur des images de méningiomes. De plus, [Tabesh et al. \[2007\]](#) exploitent les fractales afin de diagnostiquer le cancer et déterminer son grade.

Les images histologiques intègrent un certain nombre de variations discutées dans la partie 1.4.2, en particulier par rapport à la couleur des marqueurs utilisés. Pour cela, la plupart des méthodes d'analyse de la texture opèrent sur des images en niveaux de gris. Certains travaux combinent toutefois l'information de couleur avec les caractéristiques de texture, voir notamment les articles de [Sertel et al. \[2009\]](#) et de [Hervé et al. \[2011\]](#). Néanmoins, [Mäenpää et Pietikäinen \[2004\]](#) ont montré qu'il est préférable d'extraire ces deux informations séparément plutôt que conjointement, que ce soit pour les images histologiques ou non.

La texture est également analysée afin de segmenter les images en considérant que les régions à séparer présentent des textures différentes. En particulier, [Tosun et Gunduz-Demir \[2011\]](#) extraient des caractéristiques couleur de texture à partir de graphes de matrices de longueur de plages pour segmenter des images histologiques de tissus du colon. Également, [Linder et al. \[2012\]](#) segmentent l'épithélium et le stroma en se basant sur les motifs binaires locaux, les matrices de co-occurrences et les filtres de Gabor. Une comparaison de plusieurs méthodes de caractérisation de la texture pour la segmentation sur les images histologiques a été effectuée par [González-Rufino et al. \[2013\]](#).

Toutes les méthodes classiques d'analyse de texture, présentées dans la partie 4.2, mesurent des informations différentes. Celles offrant les meilleures performances dépendent donc de l'application considérée. Ainsi, de nombreux systèmes d'analyse extraient plusieurs modèles de caractéristiques de texture et les combinent afin d'accéder aux différentes informations présentes dans l'image, voir les articles de [Swamidoss et al. \[2013\]](#) et de

Mahmoud-Ghoneim [2011] à ce sujet.

Toutes ces études et les résultats qu'elles obtiennent confirment l'efficacité de l'analyse de la texture pour l'aide au diagnostic, en particulier dans le cadre des images histologiques, motivant ainsi les travaux de ce chapitre, explorant ces différentes méthodes et les comparant dans le cadre du contexte d'étude de cette thèse.

4.2 État de l'art de l'analyse de la texture

4.2.1 Statistiques du premier ordre

Les mesures de texture du premier ordre se basent sur l'histogramme des niveaux de gris de l'image en le considérant comme leur densité de probabilité. Le principal avantage de cette approche réside dans sa simplicité de part l'utilisation de mesures classiques afin de caractériser des données (Nailon [2010]). Cependant, cette méthode ne prend pas en considération les relations spatiales ni les possibles corrélations entre les pixels. De ce fait, elle n'est pas efficace pour discriminer entre tous les types de texture. Deux images peuvent avoir les mêmes histogrammes et présenter des textures très différentes.

En général sept mesures sont faites pour décrire les propriétés d'un histogramme et donc de la texture de l'image. Il s'agit de la moyenne, la variance, la rugosité, le *skewness*, le *kurtosis*, l'énergie et l'entropie. Bien qu'insuffisantes pour décrire une texture à elles seules, ces mesures sont néanmoins très utilisées en imagerie médicale conjointement à d'autres caractéristiques de plus haut niveau.

4.2.2 Statistiques du second ordre

Les statistiques de second ordre mesurent la probabilité qu'une paire de pixels régie par une certaine relation existe dans l'image. Par exemple que deux pixels ayant la même intensité soit séparés d'une certaine distance suivant telle direction. La première méthode de vision par ordinateur, à même de capturer cette information, a été proposée par Haralick et al. [1973] sous la forme de la matrice de co-occurrences. Chaque valeur $C_{d,\theta}(i, j)$ de cette matrice correspond à la probabilité de trouver un pixel avec un niveau de gris i à une distance d suivant une direction θ d'un autre pixel avec un niveau de gris j . Généralement, vingt matrices sont mesurées pour capturer les caractéristiques de textures à cinq échelles différentes suivant les directions horizontale, verticale et diagonales.

Les matrices de co-occurrences ne sont pas utilisées telles quelles afin de décrire la texture mais servent de base de calcul pour plusieurs mesures connues sous le nom de caractéristiques de texture de Haralick. La Table 4.1 présente les cinq caractéristiques les plus couramment utilisées dans l'analyse d'images biomédicales.

Ces descripteurs quantitatifs permettent d'élaborer des vecteurs de caractéristiques de texture pour un groupe d'images et de discriminer entre

TABLE 4.1 – Caractéristiques usuelles calculées sur l’ensemble des matrices de co-occurrences $C_{d,\theta}(i, j)$. d correspond à la distance en pixels, θ à l’orientation, et i et j sont des valeurs de niveaux de gris appartenant à \mathbb{N} et dont le maximum est N . μ_x , μ_y , σ_x et σ_y sont respectivement les moyennes et les écarts-types des lois de probabilités marginales de $C_{d,\theta}(i, j)$.

Caractéristique	Équation
Énergie	$= \sum_{i=0}^N \sum_{j=0}^N C_{d,\theta}(i, j)^2 \quad (4.1)$
Entropie	$= \sum_{i=0}^N \sum_{j=0}^N C_{d,\theta}(i, j) \log_2(C_{d,\theta}(i, j)) \quad (4.2)$
Contraste	$= \sum_{i=0}^N \sum_{j=0}^N \frac{(i - j)^2 \times C_{d,\theta}(i, j)}{(N - 1)^2} \quad (4.3)$
Homogénéité	$= \sum_{i=0}^N \sum_{j=0}^N \frac{C_{d,\theta}(i, j)}{1 + (i - j)^2} \quad (4.4)$
Corrélation	$= \frac{1}{\sigma_x \sigma_y} \sum_{i=0}^N \sum_{j=0}^N (i - \mu_x)(j - \mu_y) C_{d,\theta}(i, j) \quad (4.5)$

deux groupes en se basant sur ces mesures, soit directement, soit via une combinaison des caractéristiques issues des différentes matrices. Pour surmonter les défauts en termes de temps de calcul et de stockage mémoire de la matrice de co-occurrences, [Unser \[1986\]](#) a proposé une alternative nommée histogrammes de sommes et de différences. Ainsi, seuls deux vecteurs sont nécessaires afin de mesurer les caractéristiques de texture de Haralick.

4.2.3 Statistiques d’ordre supérieur

La matrice de longueurs de plages se fonde sur l’analyse d’information statistique d’ordre supérieur ([Galloway \[1975\]](#)). Une plage correspond à un ensemble de pixels de valeurs de niveaux de gris similaires adjacents dans une certaine direction. Chaque valeur $P_\theta(i, j)$ de cette matrice contient le nombre de plages de longueur j présentes dans une image suivant la direction θ et contenant des pixels de niveaux de gris i . Une texture grossière sera alors dominée par des plages longues alors qu’une texture fine aura des

plages courtes. Cette méthode est analogue à la matrice de co-occurrences dans la mesure où quatre matrices sont calculées suivant les directions $\theta \in \{0; \frac{\pi}{4}; \frac{\pi}{2}; \frac{3\pi}{4}\}$.

Les caractéristiques statistiques dérivées de l'ensemble de ces matrices sont rappelées dans la Table 4.2. Parmi lesquelles, *Short Run Emphasis* (SRE), *Long Run Emphasis* (LRE), *Gray-Level Nonuniformity* (GLN), *Run Length Nonuniformity* (RLN) et *Run Percentage* (RP) sont les premières proposées. *Low Gray-Level Run Emphasis* (LGRE), *High Gray-Level Run Emphasis* (HGRE) ont été introduites par [Chu et al. \[1990\]](#) dans le but d'équilibrer l'importance donnée aux deux informations, longueurs de plages et niveaux de gris. Pour joindre l'information issue des deux axes de la matrice, [Dasarathy et Holder \[1991\]](#) ont proposé les caractéristiques suivantes : *Short Run Low Gray-Level Emphasis* (SRLGE), *Short Run High Gray-Level Emphasis* (SRHGE), *Long Run Low Gray-Level Emphasis* (LRLGE) et *Long Run High Gray-Level Emphasis* (LRHGE).

4.2.4 Caractéristiques des filtres de Gabor

Une autre approche afin d'analyser la texture d'une image en détectant des structures ayant une dimension et une orientation spécifique consiste à utiliser des filtres de Gabor. Ces filtres de traitement du signal ont été adaptés pour étudier les images par [Daugman \[1985\]](#) et ont la forme d'un noyau gaussien modulé par une fonction sinusoïdale plane. La fonction G d'un filtre de Gabor est définie comme :

$$\forall(x, y) \in \mathbb{N}^2, G(x, y) = \exp^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \times \cos(2\pi\omega x + \phi), \quad (4.17)$$

où σ_x et σ_y sont les largeurs du noyaux gaussien selon les directions respectives x et y , et ω et ϕ sont respectivement la fréquence et la phase de la fonction sinusoïdale.

La texture est finalement décrite en appliquant une série de filtres de Gabor sur une image intégrant toutes les orientations, les échelles et les fréquences recherchées ([Olowoyeye et al. \[2009\]](#)). Les différentes orientations sont obtenues de l'équation 4.17 en effectuant des rotations rigides des axes coordonnés. De ce fait, cette approche intègre à la fois l'analyse des fréquences spatiales et des contours locaux.

Bien que cette méthode ait montré de très bonnes performances dans de nombreuses études comparatives, elle tend à être très lourde d'un point de vue calculatoire, en particulier avec des filtres de grandes dimensions.

TABLE 4.2 – Caractéristiques usuelles calculées sur une matrice de longueur de plage $P_\theta(i, j)$. i correspond à la valeur de niveau de gris et j à la longueur de plage. M est le nombre de niveaux de gris considéré, N la longueur de plage maximale, n_r le nombre de plages et n_p le nombre de pixels dans l'image.

Caractéristique	Équation
SRE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_\theta(i, j)}{j^2} \quad (4.6)$
LRE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_\theta(i, j) \times j^2 \quad (4.7)$
GLN	$= \frac{1}{n_r} \sum_{i=1}^M \left(\sum_{j=1}^N P_\theta(i, j) \right)^2 \quad (4.8)$
RLN	$= \frac{1}{n_r} \sum_{j=1}^N \left(\sum_{i=1}^M P_\theta(i, j) \right)^2 \quad (4.9)$
RP	$= \frac{n_r}{n_p} \quad (4.10)$
LGRE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_\theta(i, j)}{i^2} \quad (4.11)$
HGRE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_\theta(i, j) \times i^2 \quad (4.12)$
SRLGE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_\theta(i, j)}{i^2 \times j^2} \quad (4.13)$
SRHGE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_\theta(i, j) \times i^2}{j^2} \quad (4.14)$
LRLGE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_\theta(i, j) \times j^2}{i^2} \quad (4.15)$
LRHGE	$= \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_\theta(i, j) \times i^2 \times j^2 \quad (4.16)$

4.2.5 Dimension fractale

La géométrie euclidienne et les modèles usuels ne permettent pas de modéliser certaines structures trop complexes. Pour rendre possible leurs descriptions, Mandelbrot [1977] a proposé la théorie des fractales. Dans ce formalisme, la dimension fractale représente l'irrégularité d'une figure géométrique, telle qu'une surface ou un volume. Ainsi, cette mesure donne une indication quant aux motifs d'intensité présents dans une image, des structures répétitives auront une faible dimension fractale alors qu'une image stochastique en aura une plus élevée selon Feder [1988]. La dimension fractale peut donc être interprétée comme une description de la texture d'une image.

Parmi les nombreuses méthodes permettant d'estimer cette dimension, la plus utilisée pour analyser la texture d'une image est celle de Minkowski–Bouligand, également appelée *box counting*. Elle consiste à superposer une grille carrée sur une image binaire S et à comptabiliser le nombre de cases N nécessaires afin de recouvrir la structure à analyser. La dimension fractale est ensuite obtenue en faisant tendre la taille δ des cases de la grille vers 0 :

$$\dim_{\text{box}}(S) = \lim_{\delta \rightarrow 0} \frac{\log N(\delta)}{\log(1/\delta)}. \quad (4.18)$$

Cette mesure est généralement effectuée sur plusieurs seuillages réguliers de l'image, donnant ainsi une description sous la forme d'un vecteur de caractéristiques. Cette information est souvent couplée avec une mesure de lacunarité (Smith Jr. et al. [1996]), une mesure complémentaire aidant à distinguer deux surfaces ayant la même dimension fractale d'après Karperien et al. [2013].

Une définition de la dimension fractale et une étude sur les méthodes permettant de l'obtenir sont fournies par Lu [1997] et Landini [1996].

4.2.6 Caractéristiques sur la transformée en ondelettes

La transformée en ondelettes est une approche réputée pour la décomposition d'un signal en une somme de fonctions basiques appelées ondelettes (Mallat [1989]). Contrairement à la transformée de Fourier, la décomposition en ondelettes conserve l'information spatiale tout en donnant accès à l'information fréquentielle. Pour traiter des images, la transformée est appliquée en utilisant deux filtres passe-bas (L) et passe-haut (H) séquentiellement sur les lignes puis sur les colonnes. Après chaque filtre, l'image est soumise à un sous échantillonnage réduisant sa résolution d'un facteur 2 suivant les lignes ou les colonnes. L'image originale est alors

transformée en quatre sous-images nommées approximation (LL), détails horizontaux (LH), détails verticaux (HL) et détails diagonaux (HH). Une décomposition à n'importe quel ordre est possible en répétant ces opérations à partir de l'image d'approximation de l'ordre précédent.

Une ondelette est une fonction de l'espace de Hilbert $L^2(\mathbb{R})$. Pour chaque ondelette ψ , un groupe d'ondelettes filles, similaires à un facteur de dilatation s et un facteur de translation d près, peut être défini suivant :

$$\forall (s, d) \in \mathbb{R}^{+*} \times \mathbb{R}, \psi_{s,d}(t) = \frac{1}{\sqrt{(s)}} \psi\left(\frac{t-d}{s}\right), \quad (4.19)$$

La décomposition en ondelettes d'une image I de $L^2(\mathbb{R}^2)$ est définie comme :

$$\forall (s, d) \in \mathbb{R}^{+*} \times \mathbb{R}, WT(s, d) = \int_{-\infty}^{\infty} I(t) \psi_{s,d}(t)^* dt. \quad (4.20)$$

Les sorties d'une transformation en ondelettes peuvent être utilisées comme caractéristiques décrivant l'image mais il est plus fréquent de calculer des statistiques des premier et second ordres ou d'autres caractéristiques de texture sur les sous-images.

4.3 Extraction de l'information de texture par motifs binaires locaux

Les motifs binaires locaux (*LBP*) ont été conçus comme une mesure de motifs invariants aux changements de niveaux de gris donnant une information quant à la texture d'une image. Cette méthode a été introduite par Harwood et al. [1993] et publiée pour la première fois par Ojala et al. [1996]. Elle a montré de bonnes performances dans plusieurs études comparatives, à la fois en termes de rapidité et de discrimination.

Les *LBP* peuvent être vus comme une approche regroupant les principes stochastiques et structurels de la texture en formant des motifs plutôt qu'en étudiant ses caractéristiques à l'échelle des pixels. La description fournie par les *LBP* prend finalement la forme de la distribution de ces motifs retenant ainsi les deux propriétés des méthodes d'analyses structurelles : les primitives de textures (motifs) et les règles de placement (distribution). De même, l'utilisation de la distribution des motifs rend la méthode de type statistique. Mais une texture n'est jamais strictement structurelle ni totalement statistique et les deux modèles se complètent naturellement. L'une des difficultés est de les mêler dans des modèles capables de s'adapter à la variabilité des textures étudiées. La description fournie par les *LBP* étant à la fois statistique et structurelle, elle permet de reconnaître une large variété de textures. De plus, les *LBP* ne sont pas affectés par les changements d'illumination monotones, permettent des analyses multi-échelles et peuvent être rendus invariants en rotation.

4.3.1 Motifs binaires locaux

Principe de base

Dans la version originale des *LBP*, le motif d'un pixel c de valeur de niveau de gris g_c en fonction des niveaux de gris $(g_p)_{\forall p \in \{0..7\}}$ de ses huit voisins est défini comme :

$$\forall c \in \mathbb{N}^2, LBP(c) = \sum_{p=0}^7 s(g_p - g_c)2^p, \quad (4.21)$$

où :

$$\forall x \in \mathbb{R}, s(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}. \quad (4.22)$$

Les motifs ont donc la forme d'un mot binaire de la taille d'un octet, l'image étant décrite par la distribution des motifs. La description par *LBP* est

donc un vecteur de $2^8 = 256$ valeurs.

Dans [Ojala et al. \[2002\]](#), les *LBP* ont été étendus afin de prendre en compte des voisinages circulaires avec un rayon et un nombre d'éléments variables. Ainsi, les points de voisinage ne correspondent pas nécessairement à des pixels uniques et leurs valeurs de niveaux de gris sont interpolées bilinéairement. Cette modification permet d'extraire de l'information de texture à une échelle spécifique et donc de cibler d'autres voisinages que les huit connexes et effectuer des analyses multi-échelles. Le motif d'un pixel c avec un rayon R et un nombre de voisins P s'écrit donc :

$$\forall P \in \mathbb{N}, \forall R \in \mathbb{R}, \forall c \in \mathbb{N}^2, LBP_{P,R}(c) = \sum_{p=0}^P s(g_p - g_c)2^p, \quad (4.23)$$

où s est défini comme dans l'équation 4.22 et les coordonnées du (x_p, y_p) du $p^{\text{ième}}$ voisin de $c = (x_c, y_c)$ sont obtenues par la formule :

$$\forall (x_c, y_c) \in \mathbb{N}^2, (x_p, y_p) = \left(x_c + R \cos \left(2\pi \frac{p}{P} \right), y_c - R \sin \left(2\pi \frac{p}{P} \right) \right). \quad (4.24)$$

La figure 4.1 montre le principe du calcul d'un tel motif pour un rayon quelconque et huit voisins.

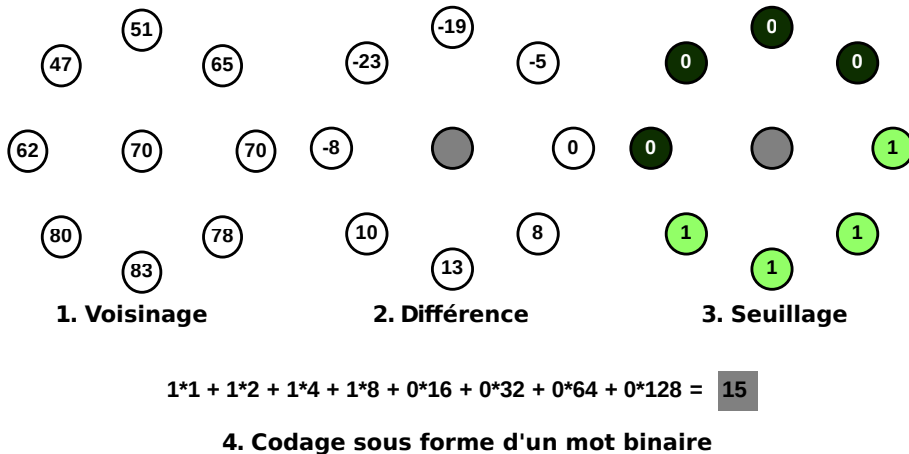


FIGURE 4.1 – Méthode de calcul d'un motif binaire local avec un voisinage circulaire de huit éléments.

Comme dit précédemment, les motifs binaires locaux peuvent être vus comme des primitives de texture. Les primitives locales extraites par les *LBP* incluent notamment les pics, les creux, les fins de lignes, les bordures, les coins et bien d'autres encore. La figure 4.2 montre quelques exemples dans le cas de $LBP_{8,R}$. Ces primitives peuvent être considérées comme des micro-textons, les éléments de texture utilisés par le système visuel humain d'après [Julesz \[1981\]](#).

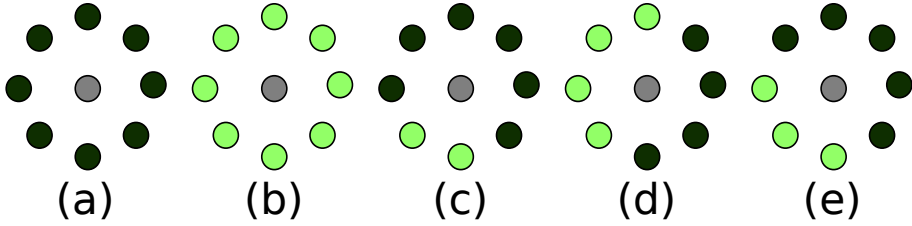


FIGURE 4.2 – Exemples de primitives de textures extraites par les *LBP*. (a) et (b) correspondent respectivement à un pic et à un creux, (c) est une fin de ligne, (d) une bordure et (e) est un coin.

Motifs invariants en rotation

De par la définition circulaire des voisinages considérés, il est possible de rendre les codes *LBP* invariants aux rotations des motifs. Cette invariance ne tient en revanche pas compte des changements dans les positions relatives de l'objet par rapport aux sources d'illumination, ni des artefacts issus de l'échantillonnage. Ce n'est donc pas une description invariante par rotation mais bien une invariance des motifs.

En pratique, cette invariance est obtenue en regroupant tous les motifs correspondants à des permutations circulaires du même mot binaire sous le label du motif de plus petite valeur. La figure 4.3 montre l'exemple de motifs étant regroupés afin d'obtenir l'invariance en rotation. Les *LBP* avec invariance en rotation sont notés $LBP_{P,R}^{ri}$.

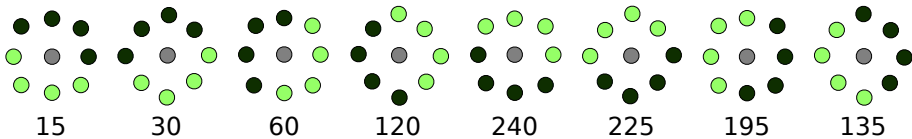


FIGURE 4.3 – Exemples de motifs regroupés en appliquant l'invariance en rotation. Tous ces *LBP* sont alors comptabilisés comme ayant le label 15.

Une autre technique pour obtenir l'invariance en rotation des *LBP* consiste à calculer la transformée de Fourier discrète des histogrammes des *LBP*. Cette méthode est proposée par Ahonen et al. [2009] et offre une réelle invariance en rotation de la description.

Motifs uniformes

Le nombre de motifs possibles, et donc la longueur du vecteur de description de la texture, augmente rapidement avec le nombre de voisins pris

en considération. Par exemple, avec un voisinage de seize points, l'histogramme contient $2^{16} = 65536$ valeurs, ce qui n'est pas exploitable dans des applications réelles. Afin de rendre possible les voisinages avec de nombreux points tout en conservant une signature de taille modeste, [Mäenpää et al. \[2000\]](#) suggèrent de ne conserver que les motifs dit *uniformes*. Les motifs uniformes sont définis comme étant ceux n'ayant qu'un nombre limité de transitions dans leurs codes binaires, tous les autres sont regroupés sous le même label. Généralement, ce nombre de transitions est fixé à deux et il a été montré dans [Ojala et al. \[2002\]](#) que les motifs uniformes deux représentent la majorité de tous les motifs présents dans une image, allant parfois jusqu'à plus de 90 %. Le nombre de transition U d'un LBP est calculé selon la formule :

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \quad (4.25)$$

Avec cette modification, la longueur de l'histogramme des LBP uniformes deux, notés $LBP_{P,R}^{u2}$, pour un voisinage de P éléments passe de 2^P valeurs à $P(P - 1) + 3$ soit 59 au lieu de 256 dans le cas de huit voisins. Ainsi, les LBP uniformes réduisent grandement la longueur du vecteur de description tout en conservant de bonnes capacités de discrimination.

Il est également possible d'extraire des motifs uniformes et invariants en rotation, ce qui donne une signature de $P + 2$ éléments dans le cas des uniformes deux, notés $LBP_{P,R}^{riu2}$. La figure 4.4 montre l'ensemble des dix motifs uniformes deux et invariants en rotation pour un voisinage de huit points.

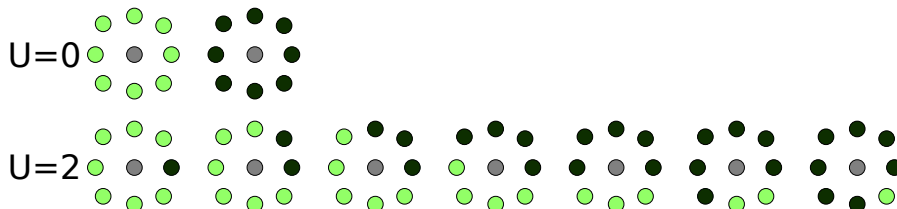


FIGURE 4.4 – Ensemble des dix motifs uniformes deux et invariants en rotation dans le cas d'un voisinage de huit points.

Modèle complété

[Guo et al. \[2010\]](#) proposent d'ajouter des informations non prises en

compte par les *LBP* classiques afin d'améliorer les performances de description de la texture, nommant cette méthode le modèle complété des motifs binaires locaux (*CLBP*). Une région locale est alors représentée par son pixel central et une transformée locale de signes et d'amplitudes des différences (LDSMT). La valeur du pixel central est codée binairesment après un seuillage global de l'image donnant la caractéristique nommée *CLBP_C*. La LDSMT décompose l'image en deux composantes complémentaires, les signes et les amplitudes des différences. Deux opérateurs sont proposés afin de les coder, respectivement nommés *CLBP_S*, qui correspond aux *LBP* classiques, et *CLBP_M*. Les trois opérateurs sont sous formes binaires et peuvent donc facilement être combinés afin de former une unique description. Les deux nouveaux opérateurs sont définis comme :

$$\forall P \in \mathbb{N}, \forall R \in \mathbb{R}, \forall s_M \in \mathbb{R}, \forall c \in \mathbb{N}^2, CLBP_{MP,R}(c) = \sum_{p=0}^{P-1} t(m_p, s_M) \times 2^p, \quad (4.26)$$

$$\forall P \in \mathbb{N}, \forall R \in \mathbb{R}, \forall s_C \in \mathbb{R}, \forall c \in \mathbb{N}^2, CLBP_{CP,R}(c) = t(g_c, s_C), \quad (4.27)$$

où m_p est l'amplitude de la différence de niveaux de gris entre le pixel central et le $p^{\text{ième}}$ voisin, s_M et s_C sont deux valeurs de seuillage devant être déterminées de façon adaptative. Guo et al. [2010] suggèrent de fixer s_M à la valeur moyenne des m_p de toute l'image et s_C au niveaux de gris moyen de l'image. De plus :

$$\forall (x, s) \in \mathbb{R}^2, t(x, s) = \begin{cases} 1 & \text{si } x \geq s \\ 0 & \text{si } x < s \end{cases}. \quad (4.28)$$

Ces trois opérateurs peuvent être combinés soit de façon jointe, soit de façon hybride. La première méthode consiste à calculer un histogramme joint tridimensionnel alors que la seconde revient à calculer l'historgramme joint bidimensionnel soit de *CLBP_S* et *CLBP_C*, soit de *CLBP_M* et *CLBP_C*, et de construire un histogramme monodimensionnel en concaténant avec la distribution restante. L'information la plus discriminante pour la texture reste le *LBP* classique, mais les deux autres opérateurs peuvent fournir des informations complémentaires. Il est également possible de définir des *CLBP* uniformes et invariants en rotation.

4.3.2 Analyse multi-échelles

Dans la littérature

Le principal défaut des *LBP* est de ne prendre en considération qu'une unique échelle. Une solution est de cumuler les informations issues de plusieurs couples de rayon de voisinage R et de nombre de voisins P sous la

forme d'un histogramme joint tel que suggéré par Ojala et al. [2002]. Cependant, cette stratégie augmente considérablement la longueur du vecteur de caractéristiques, par exemple pour trois distributions $LBP_{8,1}$, $LBP_{16,3}^{u2}$ et $LBP_{24,5}^{u2}$, la signature finale ferait $256 \times 243 \times 555 \approx 3,5 \times 10^7$ éléments. Pour cela, les distributions sont généralement simplement concaténées, négligeant de ce fait les possibles dépendances statistiques entre les distributions des différents LBP mesurés.

De plus, cette méthode risque encore de manquer des informations à cause de l'aspect creux du voisinage pour de grands rayons. Prendre un grand nombre de voisins pour compenser ce problème entraîne des signatures de très grande taille et n'est donc pas une solution. De même, les effets de crénelage ne sont pas pris en compte et la représentation est sensible au bruit puisque seules les valeurs de quelques pixels sont considérées. Une solution à ce problème, proposée par Mäenpää et Pietikäinen [2003], est de combiner les opérateurs LBP avec une banque de filtres gaussiens. Les filtres sont pris de largeurs proportionnelles aux rayons des voisinages des LBP permettant ainsi à chaque motif de capturer l'information de texture à partir de groupes de pixels et non de points.

Une alternative aux banques de filtres pour la taille des descriptions est également proposée par Mäenpää et Pietikäinen [2003] et consiste à concaténer sous forme bidimensionnelle les motifs issus des différents couples de paramètres et de déterminer les règles permettant la création de l'automate cellulaire ainsi généré. Finalement, la texture d'une image est décrite par la distribution jointe ou concaténée d'un LBP classique avec celle des occurrences des règles des automates cellulaires.

Ces deux méthodes d'analyse multi-échelles sont très lourdes en calculs et similaires du point de vue des performances. La méthode avec filtrage ne permet cependant pas de réduire la taille du vecteur de caractéristiques alors que les automates cellulaires entraînent une taille de description indépendante du nombre d'échelles prises en considération.

Approche proposée par modification de la résolution

L'extraction des LBP avec un voisinage de huit points répartis sur un cercle de rayon unitaire est rapide. Une solution pour rechercher la ou les échelles les plus discriminantes pour l'analyse de la texture dans un cadre applicatif peut consister en une réduction de la résolution des images. De cette façon, la dimension des signatures n'est pas augmentée par l'accroissement du nombre de voisins, et l'extraction de l'information n'est pas sujette aux pertes d'informations dues à l'aspect creux des grands

voisinages. Après réduction de la résolution, la nouvelle valeur d'un pixel est définie par la moyenne des pixels de haute résolution le composant, ce qui se rapproche des filtres gaussiens de [Mäenpää et Pietikäinen \[2003\]](#) mais avec moins de pixels pour décrire l'image.

Du fait de la rapidité de l'extraction des *LBP*, en particulier avec des faibles résolutions, cette approche permet de rechercher les résolutions les plus discriminantes sur les données d'apprentissage. Cette recherche impose un temps de calcul d'entraînement pouvant être long sur les grandes bases de données en échange d'un traitement des images de test très rapide. Ainsi, l'analyse multi-échelles peut être réalisée à partir d'une ou plusieurs résolutions parmi les plus pertinentes. La détection de la résolution optimale peut être réalisée selon deux méthodes :

1. en réduisant l'image à un pourcentage de sa résolution d'origine, ou
2. en imposant de nouvelles dimensions quitte à altérer les proportions de l'image.

Le choix de la méthode de recherche dépend fortement de l'application et de la texture à étudier, les modifications des dimensions de l'image pouvant donner plus ou moins d'importances à certaines directions dans la texture.

Avec cette approche, les paramètres de rayon et de taille du voisinage sont donc remplacés par la nouvelle résolution des images.

4.4 Applications

4.4.1 Paramètres et méthodes de comparaison

Méthodes de comparaison

Pour valider expérimentalement les capacités de discrimination de l'approche proposée, les méthodes classiques d'analyse de la texture sont utilisées à titre de comparaison.

Statistiques du second ordre. Les caractéristiques de texture de Haralick (GLCM) sont extraites telles que présentées dans la partie 4.2.2, c'est-à-dire selon les quatre directions principales (horizontale, verticale, et les deux diagonales) pour des distances allant de un à cinq pixels. La dynamique de l'image a été réduite à huit niveaux de gris pour réduire la taille des matrices. Finalement les cinq caractéristiques que sont l'énergie, l'entropie, le contraste, la corrélation et l'homogénéité, sont mesurées sur les vingt matrices ainsi obtenues, donnant un vecteur de cent dimensions.

Statistiques d'ordre supérieur. Quatre matrices de longueurs de plages (RLM) (voir partie 4.2.3) sont ici calculées, correspondant aux orientations horizontale, verticale et les deux diagonales sans fixer de longueur de plage maximale. La dynamique des images a été réduite à huit niveaux de gris par soucis d'homogénéité avec les caractéristiques de Haralick. Les onze caractéristiques décrites dans le tableau 4.2 sont alors mesurées, donnant finalement une description sous la forme d'un vecteur de quarante-quatre dimensions.

Caractéristiques des filtres de Gabor. Seize filtres de Gabor (voir partie 4.2.4) sont utilisés afin d'extraire l'information de texture. Quatre orientations et quatre échelles sont donc considérées et les filtres sont pris de taille 17×17 . Afin de diminuer la dimension des caractéristiques, les images sont réduites à une résolution de 960×1280 pixels et les filtres sont convolués en sous-échantillonnant les lignes et les colonnes des images d'un facteur 16. Les caractéristiques extraites sont finalement de dimension 4800, sous la forme de la concaténation de 16 vecteurs de longueur 300. Le script MATLAB utilisé a été développé et fourni par [Haghighat et al. \[2013\]](#).

Dimension fractale. L'information de texture mesurée par la dimension fractale (voir partie 4.2.5) est ici calculée selon l'algorithme de *Segmentation-based Fractal Texture Analysis* (SFTA) proposé par Costa et al. [2012]. Cette méthode se base sur un seuillage multiple de l'image suivant la méthode d'Otsu et mesure la dimension fractale par *box counting*, le niveau de gris moyen et la surface du masque, pour les différents seuils puis pour l'intersection de chaque paire de seuils mitoyens. Huit seuils sont définis, donnant donc un vecteur de caractéristiques de quarante-huit dimensions.

Caractéristiques sur une transformée en ondelettes. Une transformée en ondelettes (WT) de Daubechies (voir partie 4.2.6) est effectuée à l'ordre 5 sur les images. Les caractéristiques de premier ordre ici mesurées afin de représenter la texture sont la moyenne, l'écart-type et l'énergie comme le recommande Boucheron [2008]. Cette méthode donne donc trois mesures par image de coefficients d'ondelettes, ce qui génère un vecteur de caractéristiques de quarante-huit dimensions.

Motifs binaires locaux. Pour évaluer l'impact du redimensionnement de l'image, le modèle complété des motifs binaires locaux est également utilisé comme comparaison. Les différents opérateurs et groupes de motifs mesurés sont les mêmes que pour la méthode proposée. Également, afin de valider les performances vis-à-vis de l'analyse multi-échelles classique, la méthode avec trois voisinages la plus couramment utilisée est calculée. Les voisinages pris en compte sont donc de huit points avec un rayon unitaire, seize points avec un rayon de deux pixels et vingt-quatre points avec un rayon de trois pixels.

Méthode proposée

L'idée est ici de rechercher la ou les résolutions optimales d'image pour lesquelles les *LBP* sont les plus discriminants avec un voisinage de rayon unitaire avec huit points. Cette recherche est effectuée avec les deux méthodes proposées dans la partie 4.3.2. Dans le premier cas, la résolution varie de un à cent pourcents de la taille d'origine de l'image. Dans le second cas, l'image est forcée à une forme carrée et voit ses dimensions varier de cent à mille pixels de côté environ.

Différentes combinaisons des opérateurs du *CLBP* sont testées dans les différentes applications, de même que l'influence des différents groupes de motifs pris en compte, à savoir les motifs uniformes deux et l'invariance en rotation. Dans la littérature, la texture des images est décrite comme

étant plus visible dans les canaux bleu et vert, ce qui est confirmé expérimentalement. De ce fait, les images sont ici converties en niveaux de gris en prenant la moyenne des valeurs de bleu et de vert de chaque pixel.

Classification

L'attribution d'une étiquette pour une image se fait par une méthode de plus proche voisin avec la divergence de Hellinger comme métrique de comparaison. Dû à la faible taille des bases de données utilisées, les validations croisées sont effectuées via la méthode de *leave one out* à l'exception de l'application sur les données du *Challenge ICPR 2014 : Nuclear Atypia* où des ensembles d'apprentissage et de validation sont fournis.

4.4.2 Aide au diagnostic du glaucome sur photographies de rétine

Le glaucome et son diagnostic

Le glaucome est un groupe de maladies provoquant une neuropathie optique et est la première cause de cécité au monde. Cette maladie entraîne rarement des symptômes avant un état avancé, de ce fait, les enquêtes épidémiologiques en Amérique du Nord et en Europe ont montré qu'environ la moitié des cas ne sont pas détectés (Yu et al. [2005]). La détection précoce et la prévention représentent la seule solution permettant de traiter le glaucome et d'éviter la perte totale de la vision en raison de l'absence de traitements capables de réparer les dégâts subits par le nerf optique.

Les ophtalmologistes diagnostiquent le glaucome en se basant sur plusieurs facteurs tels que l'historique familial du patient, la pression intra-oculaire, l'épaisseur de la cornée centrale, l'apparence de l'angle de la chambre antérieure, et l'apparence, la fonction et la structure du nerf optique et la couche de fibres nerveuses. Actuellement, un diagnostic ne peut se faire à partir d'un unique résultat. Par exemple, la pression intra-oculaire est l'aspect principal de la maladie alors qu'un tiers des patients souffrant de glaucome ont une pression normale selon Kent [2008]. La plupart des tests pour détecter le glaucome sont longs et demandent un expert de la maladie en plus d'équipements spécialisés, ainsi, de nouvelles méthodes automatiques de diagnostic précises et rapides pour détecter son état précoce sont nécessaires.

D'après Kent [2008], le diagnostic automatique du glaucome intègre plusieurs obstacles. En premier lieu, il n'existe pas de consensus d'experts

médicaux quant aux symptômes précoces et le *gold standard* repose toujours sur la décision d'un expert à partir d'images du fond de l'œil. Une autre difficulté est représentée par l'inégalité des anatomies du disque optique ajoutée à la forte variabilité des dégâts causés par la maladie sur les nerfs qui augmentent encore la complexité du diagnostic. Finalement, les yeux sains et glaucomateux ont des apparences similaires, comme le montre la figure 4.5, ce qui implique qu'un diagnostic définitif est rarement prononcé.

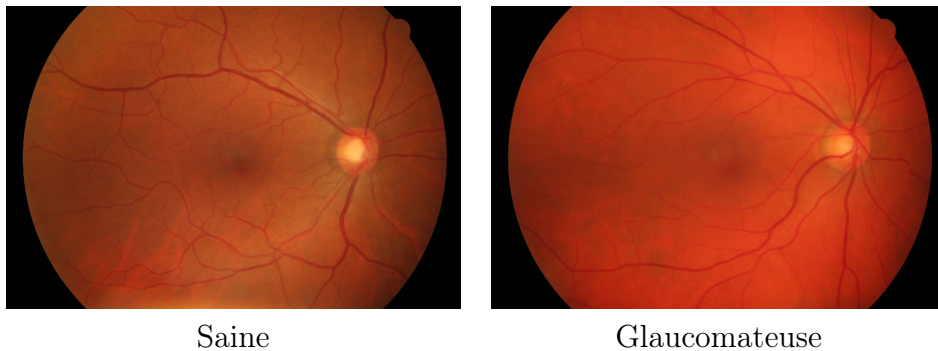


FIGURE 4.5 – Exemples d'yeux sains et glaucomateux. Il n'existe pas de différences prononcées entre les deux, d'où la complexité du diagnostic de cette maladie.

Les systèmes d'imagerie, tels que la photographie couleur du fond de l'œil, la tomographie en cohérence optique (OCT), la tomographie de la rétine d'Heidelberg (HRT) et la polarimétrie laser à balayage, sont très utilisées pour diagnostiquer les maladies de l'œil. HRT et OCT permettent de visualiser les dégâts des fibres nerveuses bien avant toute perte de vision. En revanche, ces équipements sont très coûteux. La photographie de fond de l'œil montre des informations similaires à celles qu'observent les ophtalmologistes via un ophtalmoscope. De plus, d'après l'étude de [Greaney et al. \[2002\]](#), les modalités HRT et OCT ne permettent pas de mieux distinguer un œil sain d'un souffrant de glaucome précoce ou modéré qu'une photographie du fond de l'œil. Idéalement, les données issues de ces trois modalités devraient être combinées afin d'établir un diagnostic plus fiable. Cependant, les caméras de fond d'œil peuvent être considérées comme un outil approprié afin de détecter la présence d'un risque de part leurs prix, leurs disponibilités et leurs applications courantes pour examiner l'état des yeux.

Motivation

Les méthodes de la littérature visant à détecter le glaucome (Hayashi et al. [2007]; Bock et al. [2010]; Acharya et al. [2011]; Dua et al. [2012]; Odstrcilik et al. [2012]) utilisent plusieurs étapes de pré-traitements complexes afin de supprimer toutes les variations qui ne sont pas liées à la maladie, comme par exemple, une correction de l'illumination ou une égalisation d'histogramme. De plus, les vaisseaux sanguins sont généralement supprimés par une segmentation approximative et remplacés via une technique d'*inpainting*, or, d'après Kawasaki et al. [2013] le réseaux vasculaire de l'oeil est affecté par le glaucome : les positions des vaisseaux changent à cause de la mort et de la disparition du tissu nerveux. Également, la plupart des études (Joshi et al. [2011]; Nayak et al. [2009]) sont faites sur une région d'intérêt centrée sur le disque optique, négligeant ainsi toutes les modifications susceptibles d'apparaître dans le reste de l'œil.

La méthode d'analyse de texture proposée dans ce chapitre permet de prendre en compte la totalité des données contenues dans l'image, ne cible pas de symptômes précis et ne requiert pas de pré-traitement. De plus, cette approche est simple et rapide, ce qui est une qualité nécessaire pour une méthode de dépistage préventif. Comme dit précédemment, seuls les informations issues des canaux bleu et vert sont prises en considération. La figure 4.6 souligne l'intérêt de ne pas prendre en compte le canal rouge des images.

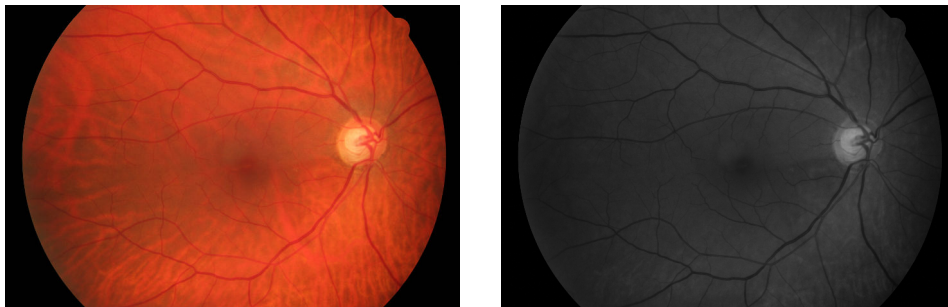


Image couleur

Moyenne des canaux bleus et verts

FIGURE 4.6 – Illustration d'une image glaucomateuse en couleur et sans le canal rouge. Cette transformation permet de mieux visualiser la structure et la texture de la rétine tout en réduisant les informations non pertinentes telles que les vaisseaux présents dans la choroïde.

Matériel utilisé

Afin de valider l'efficacité de l'analyse de la texture pour la détection du glaucome, un ensemble d'images, qui regroupe des données appartenant à la compagnie DIAGNOS inc. et des images libre d'accès fournies et décrites par Kohler et al. [2013], ont été utilisés. Les images publiques contiennent quarante-cinq images de fonds d'œil, composées pour un tiers de glaucomes, un tiers de rétinopathies diabétiques et le reste d'images saines. La société DIAGNOS inc. a fourni neuf images glaucomateuses et vingt-deux images saines. Il convient de noter que de nombreuses bases de données sont présentées dans la littérature, comme par exemple Origa^{-light} décrite par Zhang et al. [2010], mais ne sont plus accessibles publiquement.

Dans le but de s'assurer que la base ne présente pas d'images issues du même patient, seuls les yeux gauches ont été conservés. Finalement la base utilisée intègre treize images glaucomateuses, vingt saines et huit de rétinopathie diabétique. Des exemples de ces images sont montrés dans la figure 4.7.

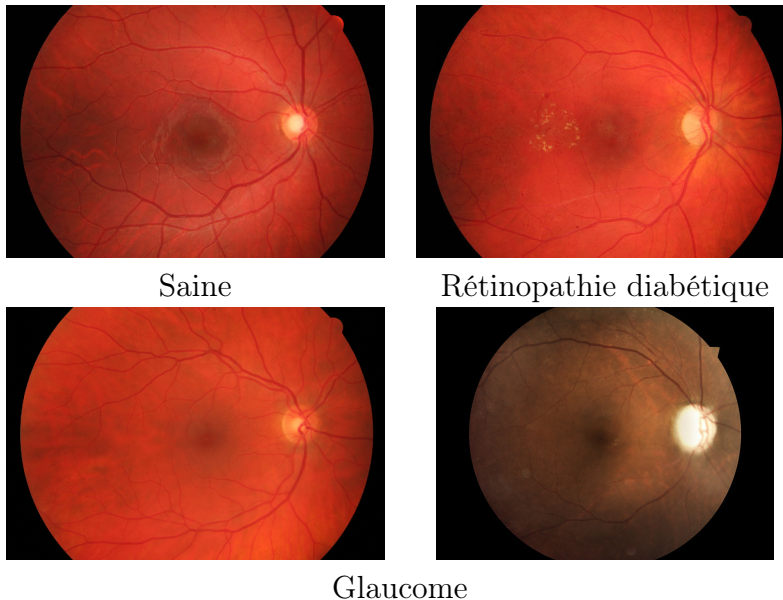


FIGURE 4.7 – Images extraites de la base de photographies de fonds d'œil.

Résultats et discussion

La figure 4.8 montre les taux de classification en fonction des deux méthodes de détection des résolutions optimales avec comme signature la

distribution des motifs uniformes deux des $CLBP_S$ concaténée avec l’histogramme joint des $CLBP_M$ et des $CLBP_C$. La courbe correspondant à la première méthode montre de nombreuses variations et un meilleur taux de reconnaissance pour une résolution réduite à environ 80 % de la taille d’origine. La seconde méthode a des variations de performances plus régulières, qui vont globalement en s’améliorant avec l’augmentation de la résolution jusqu’à atteindre un maximum de 90,2 % pour une image carrée avec un côté de mille pixels. Cependant, des reconnaissances similaires sont également obtenues pour une résolution de 225×225 pixels. En général, la seconde approche offre de meilleurs performances que la première et les deux montrent que la taille d’origine des images n’est pas la plus pertinente pour l’analyse de la texture dans ce cas particulier. Finalement, pour l’étude comparative, les signatures de $CLBP$ sont extraites sur les images réduites à 225 pixels de côté.

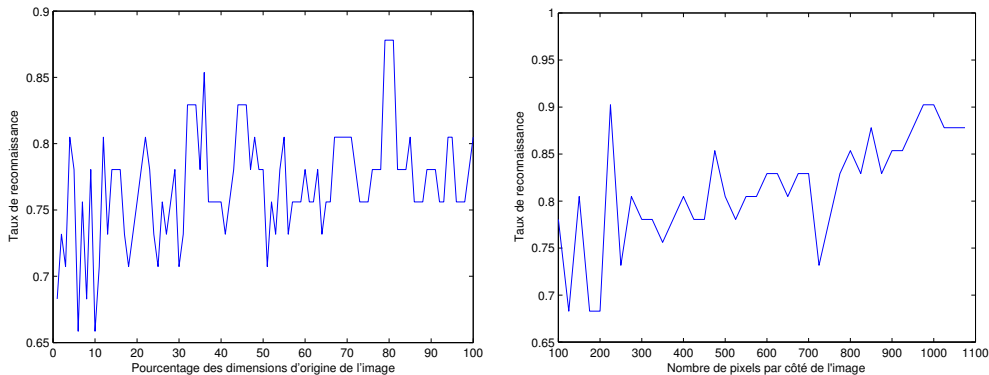


FIGURE 4.8 – Détection des résolutions optimales d’image suivant les deux méthodes proposées. À gauche en faisant varier la résolution sous la forme de pourcentages de la taille d’origine. À droite en imposant une forme carrée.

Le tableau 4.3 donne les taux de reconnaissances, les sensibilités, et les spécificités des différentes méthodes pour détecter les images atteintes de glaucome. Par soucis de comparaison, les performances de la méthode proposée avec les motifs invariants par rotation et la totalité des motifs binaires sont également fournies.

Ces résultats montrent bien que les $CLBP$ sont plus performants que les approches classiques, seule la méthode d’analyse de la dimension fractale permet d’avoir une spécificité similaire. Les LBP et les filtres de Gabor parviennent à détecter suffisamment d’images atteintes de glaucome pour avoir une sensibilité au dessus de la moyenne. La détection de l’échelle la plus discriminante permet d’améliorer les reconnaissances en augmentant grandement la sensibilité, qui est principalement recherchée dans les ap-

TABLE 4.3 – Performances des méthodes d’analyse de la texture pour la détection du glaucome. **RLBP** désigne la méthode proposée, et % représente le taux de reconnaissance.

Méthode	%	sensibilité	spécificité
RLBP _{8,1} ^{u2}	90,2 %	92,3 %	89,3 %
<i>RLBP</i> _{8,1}	78,1 %	69,2 %	82,1 %
<i>RLBP</i> _{8,1} ^{ri}	75,6 %	46,2 %	89,3 %
<i>RLBP</i> _{8,1} ^{riu2}	80,5 %	53,8 %	92,9 %
<i>CLBP</i> _{8,1}	80,5 %	69,2 %	85,7 %
<i>CLBP</i> _{8,1} ^{u2}	80,5 %	69,2 %	85,7 %
<i>CLBP</i> _{8,1} ^{ri}	75,6 %	69,2 %	78,6 %
<i>CLBP</i> _{8,1} ^{riu2}	80,5 %	69,2 %	85,7 %
<i>CLBP</i> (_{8,1} ^{u2} + _{16,2} ^{u2} + _{24,3} ^{u2})	83,0 %	69,2 %	89,3 %
GLCM	68,3 %	38,5 %	82,1 %
RLM	70,7 %	46,2 %	82,1 %
Gabor	68,3 %	69,2 %	67,9 %
SFTA	75,6 %	46,2 %	89,3 %
WT	53,7 %	23,1 %	67,9 %

plications de dépistage de maladies spécifiques, en particulier dans le cas de la présence d’une autre maladie telle que la rétinopathie diabétique. Cependant, l’échelle optimale a été sélectionnée pour les motifs uniformes deux, et n’est visiblement pas la même pour les autres types de motifs.

La description par la distribution de $CLBP_S$ concaténée avec l’histogramme joint de $CLBP_M$ et $CLBP_C$ offre les meilleures performances. A titre de comparaison, le tableau 4.4 donne les taux de reconnaissances, les sensibilités et les spécificités obtenues avec les différentes combinaisons possibles des trois opérateurs des $CLBP$ pour une résolution de 225×225 pixels. À cette échelle, les autres combinaisons d’opérateurs sont moins efficaces mais montrent néanmoins l’intérêt d’intégrer les trois informations du modèle complété des motifs binaires locaux.

Tous ces résultats préliminaires soulignent l’efficacité de la méthode proposée. Néanmoins, il est nécessaire de valider cette approche sur de plus grandes bases de données et vérifier si cette résolution demeure la plus discriminante. Également, la détection précoce du glaucome est la plus importante, pour cela une base comportant des grades de la maladie, tels que « précoce », « modéré » et « avancé », serait nécessaire.

TABLE 4.4 – Influence du choix des opérateurs de *CLBP*. *SH*, *MH* et *CH* correspondent respectivement aux distributions des opérateurs *CLBP_S*, *CLBP_M* et *CLBP_C*. La présence d’un *_* désigne la concaténation de deux histogrammes et deux notations côtes à côtes indiquent l’histogramme joint des opérateurs concernés. Ces résultats sont obtenus pour une résolution de 225×225 pixels.

Opérateurs	%	sensibilité	spécificité
<i>SH</i>	73,2%	53,8%	82,1%
<i>MH</i>	63,4%	53,8%	67,9%
<i>MCH</i>	78,0%	76,9%	78,6%
<i>SH_MCH</i>	90,2%	92,3%	89,3%
<i>SMH</i>	80,5%	84,6%	78,6%
<i>SMCH</i>	82,9%	76,9%	85,7%

4.4.3 Applications aux images histologiques du *Challenge ICPR 2014 : Nuclear Atypia*

Contexte

Le concours *ICPR 2014 : Nuclear Atypia* est organisé par Ludovic Roux et Daniel Racoceanu du laboratoire IPAL¹ de Singapour ainsi que par le professeur Frédérique Capron et Jessica Calvo, Elham Attieh, Gilles Le Naour et Anne Gloaguen de la faculté de médecine de l’UPMC et l’hôpital de la Pitié Salpêtrière à Paris.

Le cancer du sein est l’une des plus importantes causes de mortalité dans le monde. Le diagnostic de ce cancer est obtenu par l’analyse de lames histologiques issues de biopsies et marquées par de l’hématoxyline et de l’éosine. Un grade est ensuite donné à chaque image par les pathologistes en suivant le modèle développé par [Elston et Ellis \[1991\]](#). L’attribution d’un grade est réalisée selon la méthode suivante :

« In this feature both a quantitative and a qualitative judgement is made. When the nuclei are small, with little increase in size in comparison with normal breast epithelial cells, have regular outlines and uniformity of nuclear chromatin and vary little in size, 1 point is appropriate. A score of 2 points is given when the cells appear larger than normal, have open, vesicular nuclei with visible nucleoli, and there is moderate variability in both size and shape. A marked variation in size and shape, especially when large

1. CNRS UMI 2955, Université Joseph Fourier, Grenoble 1, Sorbonne Universités, UPMC Paris 06, A*STAR/I²R

and bizarre nuclei are present, scores 3 points. In this group nuclei are vesicular with prominent, often multiple nucleoli. »

La figure 4.9 illustre les trois grades d'atypie nucléaire en montrant quelques exemples de cellules.



FIGURE 4.9 – Exemples d’images illustrant l’atypie nucléaire. (a) montre des noyaux normaux alors que (b), (c) et (d) présentent respectivement des atypies de grade 1, 2 et 3.

Matériel à disposition

La base de données fournie par le concours contient 848 images capturées par deux types de microscopes différents (Aperio et Hamamatsu) avec un objectif $\times 20$. Un ensemble de 600 images sert d’ensemble d’apprentissage. Les 248 autres images servent de base de test pour le concours. Toutes les images ont été gradées par deux pathologistes seniors et la note finale est obtenue par consensus avec éventuellement l’intervention d’un troisième expert. Pour cette application, les grades ont été fournis pour chaque image et non pour chaque échantillon, ainsi deux images issues d’une même lame peuvent être de grades différents. Les images sont peu homogènes et présentent de fortes variabilités de teintes suivant les lames et les microscopes. La figure 4.10 illustre ces variations et d’autres en montrant quelques images tirées de cette base.

La base d’images utilisée dans le cadre de ce concours est cependant très mal équilibrée, comme le montre la figure 4.11. De plus, six et huit images, respectivement dans les ensembles d’apprentissage et de test, sont considérées comme ne comportant pas suffisamment d’informations pour être gradées. Cela souligne le problème de la vérité terrain. En effet, il n’est pas possible d’évaluer les performances d’une méthode automatique sur ces images et donc les approches proposées restent limitées par les capacités de diagnostic des pathologistes.

Résultats et discussion

La figure 4.12 montre l’évolution du taux de reconnaissance suivant les deux méthodes de réduction de la résolution pour cette application.

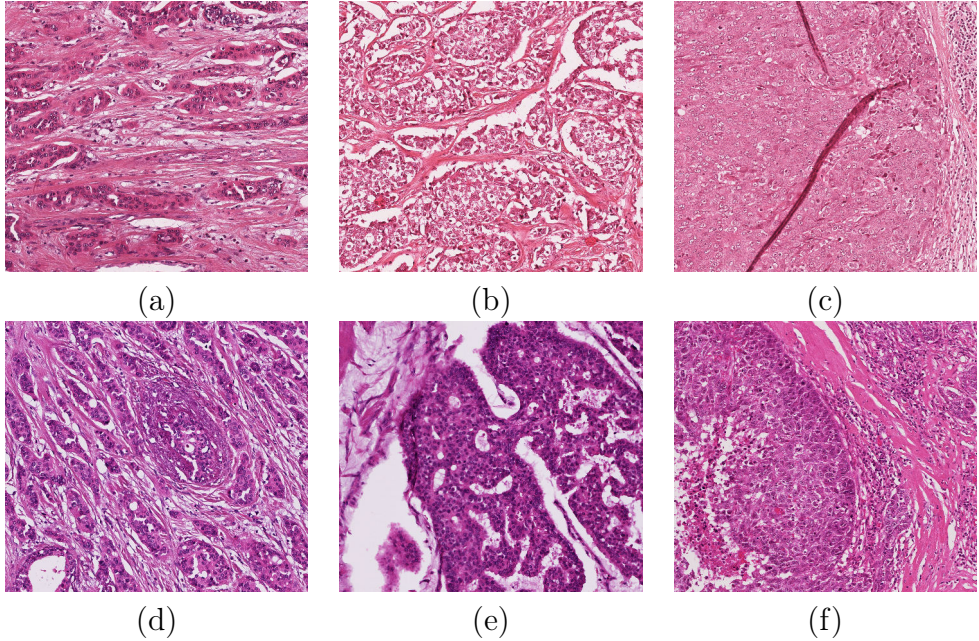


FIGURE 4.10 – Exemples d’images illustrant les variations dans la base du concours ICPR 2014. (a) et (d) sont de grades 1, (b) et (e) de grades 2, et (c) et (f) du troisième grade. Toutes les variabilités usuelles des images histologiques se retrouvent dans cette base.

Les meilleures reconnaissances sont donc de 50,8 % pour une résolution fixée à 250×250 pixels avec les motifs uniformes deux et la distribution jointe des opérateurs $CLBP_S$, $CLBP_M$ et $CLBP_C$. Des exemples sur une image représentative par grade sont donnés dans la figure 4.13 afin de montrer l’impact de ce changement de résolution.

Le tableau 4.5 donne les performances des différentes méthodes de comparaison. Toutes les méthodes offrent des résultats médiocres, en particulier les filtres de Gabor, soulignant les faibles capacités de la texture de l’image afin de décrire les grades de l’atypie nucléaire.

TABLE 4.5 – Comparaison des performances de la méthode proposée avec les approches classiques de l’analyse de la texture. **RLBP** désigne la méthode proposée, et $multLBP$ les $CLBP(u_{8,1}^2 + u_{16,2}^2 + u_{24,3}^2)$.

RLBP	GLCM	RLM	Gabor	SFTA	WT	$multLBP$
50,8 %	49,2 %	44,4 %	25,8 %	43,5 %	46,4 %	43,5 %

Les lames histologiques présentent des performances individuelles très variables. Comme le souligne en effet le tableau 4.6, les reconnaissances dé-

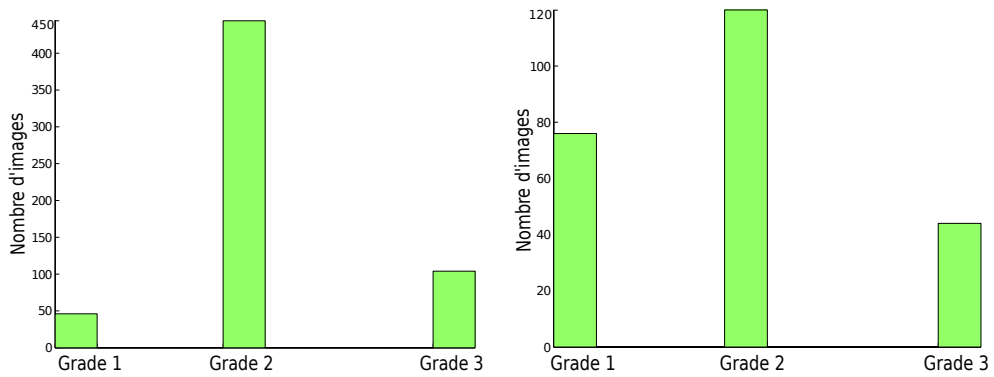


FIGURE 4.11 – Distribution des grades dans les données d’apprentissage et de test du concours ICPR 2014. L’histogramme de gauche montre la répartition des données d’apprentissage et celui de droite des données de test.

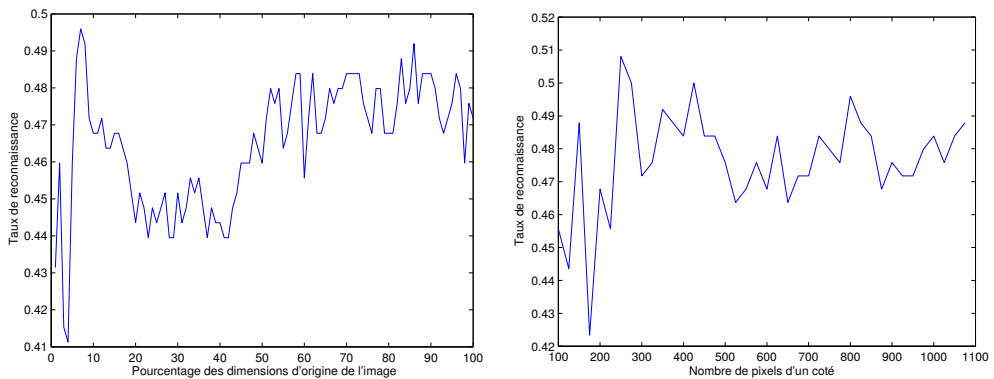


FIGURE 4.12 – Taux de reconnaissance en fonction de la résolution de l’image.

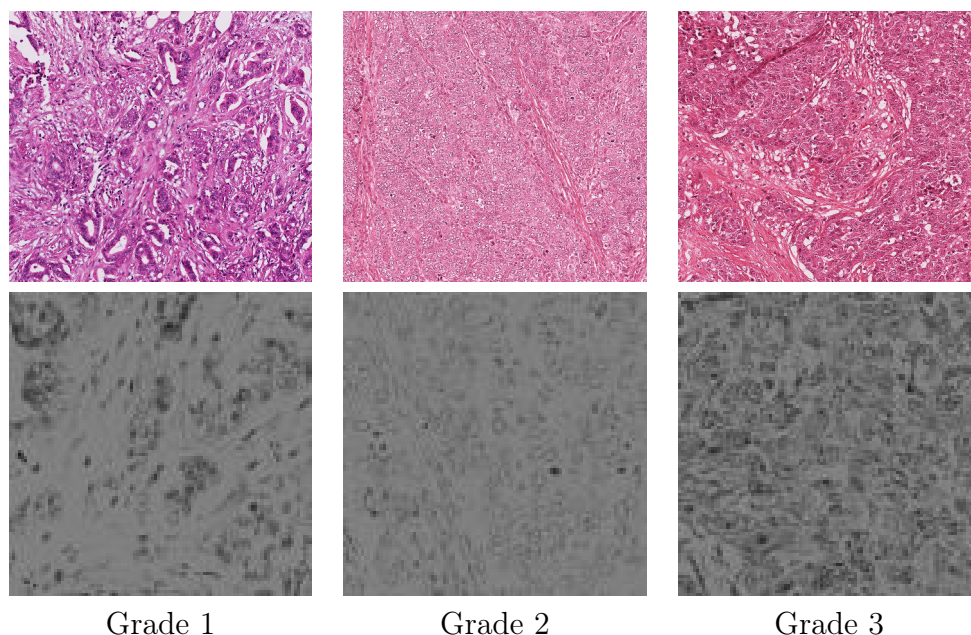


FIGURE 4.13 – Illustration de l’impact du changement de résolution. Les images de la première ligne sont celles d’origine. Celles de la seconde ligne sont des détails des images dépourvues de canal rouge et redimensionnées. Les noyaux sains apparaissent comme de simples points noirs alors que ceux atteints d’atypie présentent des formes d’autant plus nombreuses quand le grade est important.

pendent à la fois des lames et du microscope utilisé. Ces résultats montrent un manque de robustesse aux variations intra-classes de la méthode proposée et insistent sur la complexité de cette application. En particulier, une des raisons pour les différences entre les microscopes peut être liée à la résolution des images produites, le scanner Hamamatsu produisant des images avec des pixels rectangulaires et non carrés. De même, comme dit précédemment, les grades sont attribués par images et non par lames, ceci peut donc expliquer pourquoi les lames les mieux classées sont celles présentant une moindre variété d’étiquettes. A titre de comparaison, les meilleures performances lors du concours ont été de 142 points, contre 117 avec cette approche lui octroyant ainsi la deuxième place sur six.

Les grades sont attribués selon des modifications de l’apparence des noyaux cellulaires, et donc de leur texture et de leur morphologie. Or une image peut ne contenir qu’un nombre réduit de noyaux altérés, faisant que leur impact sur la distribution des *LBP* risque d’être faible. De plus, ces images révèlent de fortes variations intra-grades, liées à l’emplacement de la biopsie tel que discuté dans le chapitre 1. Une solution, pour réduire

TABLE 4.6 – Taux de reconnaissance de chaque lame sur l’application ICPR 2014.

A06	A08	A09	A13	A16
100 %	57,1 %	45,8 %	43,8 %	45,0 %
H06	H08	H09	H13	H16
50,0 %	75,0 %	37,5 %	18,8 %	42,5 %

ces variations et de donner plus d’importances aux noyaux atypiques, peut être de découper l’image en plusieurs imasettes et de supprimer celles ne contenant pas de noyaux. Cela implique néanmoins d’agrèger les descriptions de texture issues de ces différentes imasettes sans annuler l’effet des noyaux atypiques.

4.4.4 Application aux données histologiques de modèles animaux de cancer colorectal

Motivation

Cette application utilise les données présentées dans la partie 1.5. Comme le soulignent [Tabesh et al. \[2007\]](#) et [Jafari-Khouzani et Soltanian-Zadeh \[2003\]](#), le lien entre les caractéristiques de texture du tissu et le degré de malignité d’un cancer est avéré. L’intérêt ici est donc de vérifier la présence de variations dans la texture dans ce cadre applicatif afin de permettre de discriminer les différents jours après implantations ainsi que les modèles.

Trois études sont donc réalisées, la classification selon les jours après implantation (11, 15 et 21), selon le modèle d’implantation des tumeurs (ectopique et orthotopique) et enfin selon les jours et les modèles (six classes).

Résultats et discussion

Les courbes correspondant aux deux approches proposées afin de sélectionner les dimensions les plus pertinentes sont données dans la figure 4.14. Pour cette application, altérer les proportions de l’image ne permet pas d’extraire plus d’informations utiles. En revanche, réduire les images à une faible résolution permet de mieux discriminer pour les trois études. La résolution optimale pour mesurer la texture par les motifs binaires locaux est cependant différente pour chaque étude. Également, ces courbes montrent que les images originales offrent des performances égales voire inférieures

au hasard, contrairement aux faibles résolutions, ce qui souligne la difficulté de ces trois études.

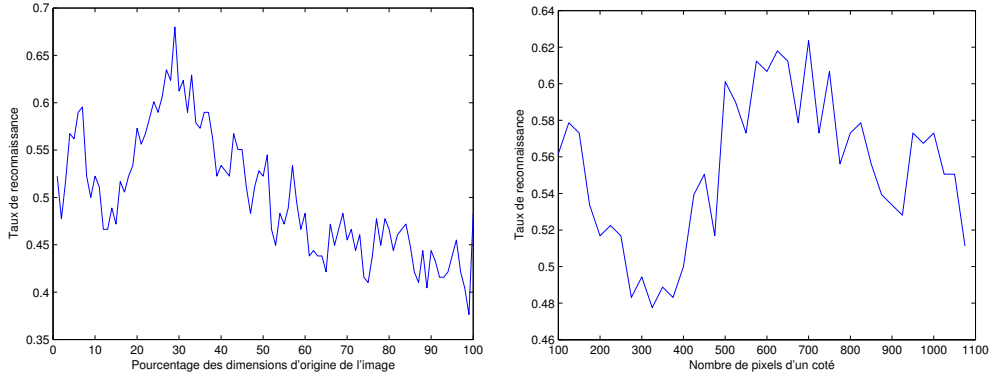
Le tableau 4.7 montre les performances de l’approche proposée comparées à celles des méthodes classiques d’analyse de la texture pour les trois études. Les validations croisées sont réalisées image par image en *leave one out* en supprimant de l’ensemble d’apprentissage les images appartenant à la même souris que l’image testée. Ceci afin d’éviter tous biais issus de la similarité entre images provenant d’un même échantillon.

TABLE 4.7 – Comparaison des taux de reconnaissances de la méthode proposée avec les approches classiques de l’analyse de la texture. **RLBP** désigne la méthode proposée.

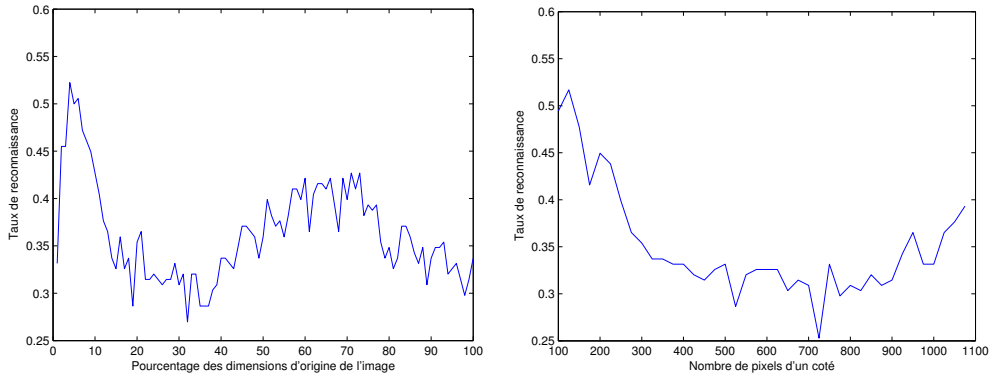
Méthode	Modèles	Jours	Les deux
RLBP	68,0 %	52,3 %	24,7 %
$CLBP(u_{8,1}^2 + u_{16,2}^2 + u_{24,3}^2)$	44,5 %	23,6 %	10,0 %
GLCM	45,5 %	23,6 %	6,2 %
RLM	37,6 %	20,2 %	8,4 %
SFTA	43,3 %	30,9 %	14,0 %
WT	42,7 %	27,0 %	7,3 %

Les méthodes couramment employées dans la littérature offrent toutes des performances très similaires, à l’exception de la matrice de longueurs de plages qui est moins performante. Cependant, seule la méthode proposée parvient à des reconnaissances supérieures au hasard, qui correspond dans ce cas à 50 % pour les modèles, 33,3 % pour les jours et 16,7 % pour les deux. Pour replacer ces résultats en perspective, ces méthodes sont généralement appliquées dans des contextes de détection de maladie ou d’évaluation de la sévérité d’une même maladie. Or, dans cette application, les différentes classes correspondent à une même maladie affectant des cellules issues d’un même organe mais évoluant dans des environnements différents. De ce fait, les images sont très similaires visuellement, ce qui rend cette application très difficile.

Modèles : ectopique et orthotopique



Jours : 11, 15 et 21



Les deux : 6 classes

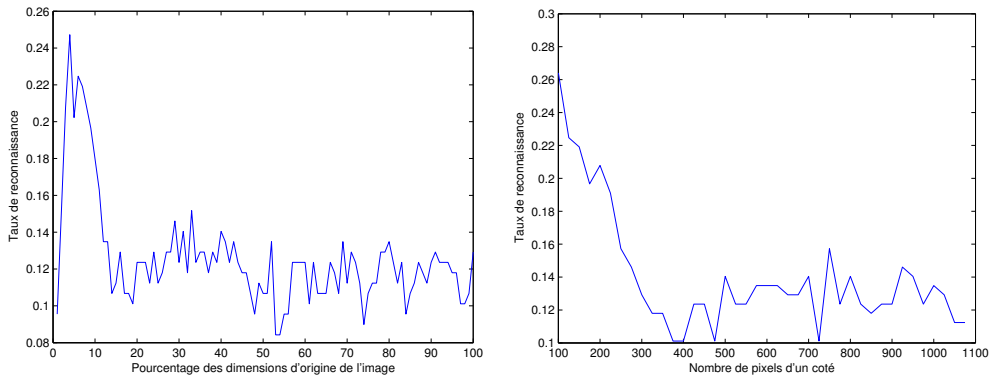


FIGURE 4.14 – Taux de reconnaissance en fonction de la résolution de l'image pour les trois études avec les deux modalités de modification de résolution. Les deux approches montrent que les basses résolutions sont les plus discriminantes quelque soit l'étude considérée.

4.5 Conclusion et perspectives

L'information de la texture extraite à l'échelle des pixels offre de bonnes performances dans le domaine de l'analyse d'images histopathologiques malgré qu'elle ne soit pas aisément interprétable biologiquement. Ce manque d'interprétabilité est également révélateur de l'intérêt de ces méthodes puisqu'elles capturent de façon quantitative une information peu prise en compte par les pathologistes.

De nos jours, les images sont acquises avec des résolutions de plus en plus élevées. Ce surplus d'informations risque cependant de noyer des données pertinentes en fonction des études et des objectifs considérés. Pour répondre à ce problème, les méthodes de la littérature se concentrent sur des approches multi-échelles afin de capturer le maximum d'informations. Ces techniques sont au final très lourdes en calcul et imposent souvent une étape de sélection de caractéristiques. Ce chapitre introduit une autre approche afin de remédier à ce problème en analysant les données à disposition afin de localiser la ou les échelles les plus porteuses d'informations de texture pour une application donnée. Une variante à cette approche est également discutée et consiste à déformer les images pour faire ressortir de l'information pertinente. Les expérimentations réalisées sur trois applications différentes montrent l'intérêt de sélectionner une échelle optimale plutôt que de capturer la texture à partir des images d'origine.

Actuellement, seule la résolution optimale a été prise en compte. Une première évolution de cette approche est donc de cumuler les informations issues de plusieurs échelles pertinentes. Cela impose néanmoins de tester la complémentarité des informations de texture issues des différentes échelles considérées via un test statistique. Une deuxième piste à poursuivre serait de valider le comportement de cette méthode sur de grandes bases de données médicales et d'étudier ses relations avec d'autres méthodes de classification plus adaptées à de larges ensembles d'images. Finalement, il serait intéressant de compléter l'information de texture extraite par les motifs binaires locaux par une mesure plus globale telle que la dimension fractale afin d'améliorer la robustesse de la méthode.

Prise de décision et interprétation des mesures

L'analyse d'images histologiques se fait par l'extraction de nombreuses caractéristiques qui ne sont pas nécessairement interprétables biologiquement. En revanche, la dernière étape de toute analyse est de parvenir à décrire les images à partir de ces caractéristiques. Pour cela, il est nécessaire d'avoir des outils permettant d'interpréter les mesures, de repérer leurs capacités de discrimination ainsi que leurs différences significatives. Ce dernier chapitre traite donc des statistiques mises en œuvre, tant descriptives que déductives, que ce soit pour analyser les données à disposition (5.2) ou pour sélectionner les caractéristiques les plus discriminantes afin d'attribuer une étiquette aux échantillons (5.3). Finalement, les méthodes ici présentées sont illustrées et appliquées sur les données histologiques à disposition.

5.1 Motivation

L'analyse d'images histologiques s'achève par l'interprétation et l'étude de la pertinence des différentes caractéristiques extraites à partir des images vis-à-vis du contexte médical correspondant. Dans la littérature, plusieurs méthodes de tests statistiques sont ainsi appliquées avec des objectifs différents. Notamment, [Miedema et al. \[2012\]](#) collectent de nombreuses caractéristiques usuelles puis les analysent d'un point de vue statistique via le test d'hypothèse non paramétrique DiProPerm qui convient particulièrement aux données de hautes dimensions. Leur objectif est de déterminer les caractéristiques les plus discriminantes afin de comparer les lésions mélanocytaires.

Dans un autre contexte, [Tuczek et al. \[1996\]](#) étudient la pertinence de l'analyse d'images afin de retrouver le grade de carcinomes des seins par des études de corrélations entre la vérité terrain et les caractéristiques extraites. Par ces tests, ils montrent quelles caractéristiques sont associées avec les différents critères permettant d'établir le grade global.

[Wild et al. \[2000\]](#) emploient des analyses de la variance répétées pour étudier les différences dans le dénombrement manuel des micro-vaisseaux dans le cas des images de contrôle et de celles de traitement. Également, ils analysent les associations entre les mesures automatiques et manuelles visuellement par l'observation d'un *scatterplot* et mathématiquement par le calcul du coefficient de corrélation de Pearson.

Le coefficient de variation est utilisé par [Taverna et al. \[2009\]](#) pour évaluer les variations de la dimension fractale de la surface représentée par les vascularisations entre tissus cancéreux et sains. Cette étude a permis de conclure que cette information est un indice géométrique utile pour le diagnostic.

Une autre façon de sélectionner un sous-ensemble de caractéristiques et d'en valider les performances consiste à employer des outils ne permettant que peu d'interprétations biologiques mais attribuant des étiquettes de façon supervisée aux données. Quelques exemples de ces méthodes de classification et de sélection sont fournis dans le tableau 5.1 conjointement aux caractéristiques utilisées.

TABLE 5.1 – Exemples de méthodes de classification et de sélection de caractéristiques à partir d’images histologiques. Kppv désigne les K plus proches voisins, GLCM les matrice de co-occurrences en niveaux de gris, PCA l’analyse en composante principale, LDA l’analyse discriminante linéaire, SVM sont les supports vecteurs machine, ANOVA et MANOVA respectivement l’analyse de la variance et sa version multivariables, mRMR la sélection par *minimum Redundancy Maximum Relevance*, et RBF et PNN respectivement les réseaux *Radial Basis Function* et de neurones probabilistes.

Organe	Mesures	Sélection	Classification	Références
Utérus	Triangulation de Delaunay	Test de Kruskal-Wallis et <i>stepwise discriminant analysis</i>	Kppv	Keenan et al. [2000]
Follicule	GLCM couleur et organisation des noyaux	PCA et LDA	Classifieur bayésien	Sertel et al. [2009]
	<i>Cell-graphs</i>	PCA	SVM, classifieur bayésien et Kppv	Oztan et al. [2012]
Prostate	GLCM et LBP	ANOVA	RBF	Oliveira et al. [2014]
Rein	Forme des noyaux	mRMR	SVM	Kothari et al. [2013b]
Sein	Texture et architecture des noyaux	mRMR	Multi-FOV et forêt d’arbres décisionnels	Basavanhally et al. [2013]
	Texture	MANOVA	Kppv, PNN, SVM	Loukas et al. [2013]

5.2 Statistiques pour l'interprétation des données

Les statistiques permettent de résumer en quelques valeurs des ensembles de données, telles que des caractéristiques extraites des images, afin de permettre une meilleure interprétation ou d'en tirer des conclusions. Les statistiques descriptives sont des procédures servant à synthétiser, organiser, représenter ou décrire des données alors que les statistiques déductives permettent d'obtenir des inférences sur des populations à partir d'échantillons grâce à des tests de signification.

5.2.1 Statistiques descriptives

Cette première famille regroupe de nombreux outils usuels soit numériques, soit graphiques, permettant de représenter un ensemble de données. La meilleure combinaison de statistiques pour un groupe n'est pas connue d'avance, il est nécessaire d'avoir des connaissances concernant la distribution des données. Par exemple, des populations sont très souvent représentées par leur tendance centrale et une mesure de dispersion. Dans le cas où les données sont symétriques, la moyenne et l'écart-type remplissent ce rôle. En revanche, si la distribution des données est dissymétrique il est plus judicieux d'utiliser la médiane et l'écart interquartiles. En effet, la moyenne est sensible à la présence de valeurs aberrantes contrairement à la médiane.

De nombreuses procédures en biologie et en médecine requièrent de mesurer la densité des objets biologiques présents dans les images histologiques. Comme dans la partie 2.4.1, deux catégories d'objets sont considérées, la densité de chacune est définie comme :

Objets ponctuels : la densité est le nombre de points par unité de surface.

Objets surfaciques : la densité peut ici être définie de deux façons. La première est similaire à celle des objets ponctuels en calculant le nombre d'objets par unité de surface. La seconde consiste à mesurer le rapport entre la surface occupée par les objets et la surface totale des images.

Bien que la densité puisse être considérée comme une statistique descriptive, il est possible de décrire un ensemble de mesures de densités issues d'une base de données pour obtenir des conclusions plus approfondies.

Les outils graphiques permettent une rapide visualisation des données

et sont donc très efficaces pour décrire leurs propriétés. Il est donc recommandé de les employer pour finalement sélectionner les statistiques numériques les plus appropriées. Les plus courants sont présentés ici :

Histogramme : un histogramme permet d'avoir une idée quant à la distribution et à la dispersion d'un ensemble de données. En particulier, cela permet de visualiser rapidement la présence et la quantité de données aberrantes ainsi que de vérifier la symétrie de la distribution.

Boxplot : ce graphe se base sur la synthèse en cinq valeurs que sont le minimum, le quartile inférieur, la médiane, le quartile supérieur et le maximum. Souvent, les extrema sont remplacés par les quantiles à 2,5 % et 97,5 % et les points aberrants sont indiqués par des points individuels. Ces graphiques permettent également de repérer rapidement des distributions asymétriques, via la position de la médiane dans la boîte, et sont efficaces pour comparer des données groupées.

Nuage de points (*scatterplot*) : les nuages de points permettent de faire ressortir les relations entre deux variables numériques en traçant les valeurs d'une caractéristique en fonction de celles d'une autre. Ces graphes sont généralement la première étape dans l'analyse statistique de la corrélation entre variables et des analyses de régression qui suivent.

5.2.2 Statistiques déductives

Les statistiques descriptives sont utilisées afin de présenter et de synthétiser des résultats. Cela peut être la base d'une prise de décisions et de conclusions, pourtant, il est probable que ces résultats soient issus d'une coïncidence. Si les expériences sont reproduites après une nouvelle collecte de données, des conclusions différentes peuvent être obtenues. Les statistiques déductives permettent de savoir si les résultats sont dus au hasard ou sont représentatifs d'effets sous-jacents *réels*.

Notion de test statistique

La base de tout test statistique est la définition d'une hypothèse nulle (H_0) qui peut être rejetée en lien avec une hypothèse alternative (H_1) sous certaines conditions. L'hypothèse nulle est généralement définie sous une forme conservatrice d'absence de différences entre échantillons ou encore d'absence de corrélations, etc. alors que l'hypothèse alternative correspond à l'opposé de l'hypothèse nulle. Le test statistique et les hypothèses nulle

et alternative à vérifier doivent être sélectionnés en fonction des données à étudier, en particulier selon leur distribution, connue ou supposée.

La signification d'un test statistique est définie par sa valeur p , qui correspond à la probabilité d'avoir un résultat identique ou plus extrême que celui obtenu avec l'échantillon à étudier dans le cas où l'hypothèse nulle est vraie. Le degré de signification du test est défini comme le seuil de probabilité α décidé avant de réaliser le test. Si la valeur p du test est inférieure à ce seuil α , le test rejette l'hypothèse nulle. Dans le cas contraire, il n'y a pas suffisamment d'informations afin de rejeter l'hypothèse nulle, néanmoins, cela ne signifie pas que cette hypothèse est acceptée. Dans la plupart des tests, α est choisi entre 0,01 et 0,05.

Deux types d'erreur peuvent être commis lors de la réalisation d'un test statistique. Le degré de signification α peut être interprété comme la probabilité de rejeter l'hypothèse nulle alors qu'elle est en réalité vraie, ce qui est défini comme une erreur de type 1. Une erreur de type 2 correspond à un faux positif, et donc à une acceptation erronée de l'hypothèse nulle. Cette seconde erreur est généralement due à une trop petite taille d'échantillons. Le tableau 5.2 résume les deux cas d'erreurs statistiques et leurs relations avec les hypothèses et la puissance du test. La puissance d'un test est définie comme la probabilité de rejeter correctement une hypothèse nulle fautive (Greene [2000]). Cette puissance est liée à trois facteurs : le nombre d'échantillons, la distribution des données et le degré de signification désiré. Elle est généralement déterminée pour deux raisons :

1. avant collecte des données en se basant sur des informations a priori afin de décider le nombre d'échantillons nécessaire pour l'étude,
2. après analyse des données, généralement quand les résultats ne sont pas significatifs. Dans ce cas, la puissance de l'étude permet de vérifier si l'absence de signification est réellement liée aux données ou est due à un nombre trop faible d'échantillons.

TABLE 5.2 – Erreurs de types 1 et 2 ainsi que leurs relations avec le degré de signification α et la puissance du test $1 - \beta$.

	H_0 est vraie	H_1 est vraie
H_0 est acceptée	Décision correcte	Erreur de type 2 : β
H_0 est rejetée	Erreur de type 1 : α	Décision correcte

Généralement, la puissance d'un test est principalement affectée par le nombre d'échantillons. En effet, le degré de signification est souvent fixé entre 0.01 et 0.05 et ne peut donc pas être beaucoup modifié. Également, la distribution des données ne peut être altérée que dans de rares applications

et finalement obtenir plus d'échantillons est coûteux et chronophage. En revanche, si trop d'observations sont effectuées, ou si un test possède une trop grande puissance avec beaucoup d'échantillons, même un effet trivial peut être détecté par erreur comme un effet significatif. Ainsi, tout peut être prouvé virtuellement au mépris des effets véritables (Park [2010]). A l'opposé, si trop peu d'observations ont lieu, un test d'hypothèse sera plus faible et moins convaincant. Par conséquent, il sera moins probable de repérer un effet significatif existant. Cependant, il n'existe pas d'idée claire de ce que représentent « *trop* » et « *trop peu* » puisque cela dépend des modèles et des spécificités. Le livre de Cohen [1988] présente de plus amples informations sur la puissance d'un test, en particulier les méthodes et tables de valeurs permettant de la déterminer.

Corrélation

Deux caractéristiques ou plus issues des images histologiques peuvent ne pas être distribuées indépendamment les unes des autres mais manifester un certain degré d'association ou de corrélation. Ainsi, l'analyse de ces liens peut aider à comprendre et interpréter les mécanismes sous-jacents aux effets observés.

Le test de Pearson est couramment utilisé afin d'obtenir un degré de l'association entre deux caractéristiques. Soit deux caractéristiques A et B capturées sur n échantillons et de moyennes respectives \bar{A} et \bar{B} , le coefficient de corrélation est estimé selon :

$$r_{A,B} = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}}. \quad (5.1)$$

Ce coefficient est une mesure du degré de corrélation linéaire entre deux caractéristiques et ses valeurs sont comprises dans l'intervalle $[-1, 1]$. Des valeurs nulles indiquent une absence d'association, alors que -1 et 1 réfèrent respectivement à des corrélations linéaires totales négative et positive.

Toutefois, ce test est limité aux données distribuées selon une loi normale et ne permet la détection que d'associations linéaires. Le test non paramétrique de Spearman vise à détecter des relations monotones entre deux caractéristiques sans imposer une distribution gaussienne des données. Le coefficient correspondant est obtenu selon :

$$r_{A,B} = \frac{\sum_{i=1}^n A_i^2 + \sum_{i=1}^n B_i^2 - \sum_{i=1}^n (A_i - B_i)^2}{2\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (5.2)$$

Afin de déterminer la signification statistique de ces deux coefficients, l'hypothèse nulle H_0 : « *il n'y a aucune association entre les deux caractéristiques* » doit être testée comparativement à l'hypothèse alternative H_1 : « *il existe des associations entre les deux caractéristiques* ».

De nombreux autres tests de corrélation existent dans la littérature, pour de plus amples informations, il est possible de se référer au livre de [Cohen et al. \[2013\]](#) ou au chapitre 2 de celui de [Marques et Joaquim \[2007\]](#).

Analyse de la variance (ANOVA)

Test paramétrique. L'analyse de la variance (ANOVA) est utilisée afin de déterminer si une variable quantitative présente des valeurs significativement différentes selon plusieurs facteurs qualitatifs. Chaque facteur peut prendre plusieurs valeurs, nommées niveaux. Un groupe est défini comme l'ensemble de tous les échantillons correspondant à un unique niveau par facteur. L'ANOVA compare les variations entre groupes avec celles présentes à l'intérieur de chaque groupe afin d'évaluer s'il existe des différences entre les moyennes des différents groupes.

Pour cela, l'ANOVA teste l'hypothèse nulle H_0 : « *les différents groupes ont des moyennes égales* » contre l'alternative H_1 : « *au moins deux groupes ont des moyennes différentes* ». Si les variations entre les moyennes des groupes sont importantes par rapport aux variations à l'intérieur des groupes, alors des différences significatives sont détectées. Ce qui amène au rejet de H_0 et à l'acceptation de H_1 .

La première étape dans l'analyse de la variance est de définir un modèle théorique basé sur le problème à étudier. Il est souvent possible de rédiger plusieurs modèles pour la même problématique, dépendamment des informations à inclure dans l'étude. Le modèle général est ainsi écrit :

$$y_{ijk\dots} = \Lambda + f(i, j, k, \dots) + \epsilon_{ijk\dots}, \quad (5.3)$$

où i , j et k sont les niveaux des facteurs, $y_{ijk\dots}$ est la variable quantitative, Λ est un effet global, également appelé effet commun des facteurs, $f(i, j, k, \dots)$ est une relation entre les facteurs et $\epsilon_{ijk\dots} = N(0, \sigma^2)$ est l'erreur de mesure suivant une distribution normale de moyenne nulle. Il existe plusieurs types d'ANOVA en fonction du nombre de facteurs à considérer. Ces ANOVA factorielles permettent de tester deux effets différents :

- L'effet principal du facteur : il s'agit d'une mesure permettant de montrer des différences cohérentes entre les niveaux d'un même facteur. Les hypothèses H_0 et H_1 pour cet effet d'un facteur présentant n niveaux, où μ représente la moyenne, sont définies comme :

$$\begin{cases} H_0 & : \quad \forall (i, j) \in \{1..n\}^2, \mu_i = \mu_j \\ H_1 & : \quad \exists (i, j) \in \{1..n\}^2, \mu_i \neq \mu_j \end{cases}, \quad (5.4)$$

— L'effet d'interaction : il existe quand les différences au sein d'un facteur dépendent des niveaux d'autres facteurs. Les hypothèses H_0 et H_1 pour l'interaction γ entre deux facteurs présentant respectivement n et m niveaux sont :

$$\begin{cases} H_0 & : \quad \forall i \in \{1..n\}, j \in \{1..m\}, \gamma_{ij} = 0 \\ H_1 & : \quad \exists (i, j), i \in \{1..n\}, j \in \{1..m\}, \gamma_{ij} \neq 0 \end{cases}. \quad (5.5)$$

Les modèles nommés à *effets mixtes* contiennent un effet aléatoire supplémentaire et sont souvent appropriés pour représenter des données dépendantes. C'est en particulier le cas si les données sont prises de façon hiérarchique ou sont collectées à partir d'individus similaires. Les modèles mixtes étant un sujet à part entière, le livre de [Raudenbush et Bryk \[2002\]](#) permet d'en avoir une meilleure compréhension.

La réalisation d'un test paramétrique d'ANOVA impose de respecter plusieurs hypothèses vis-à-vis des données à étudier. Il existe quatre principales hypothèses :

1. les données suivent une distribution normale,
2. les erreurs sont indépendantes,
3. l'absence de données aberrantes,
4. l'homoscédasticité, c'est-à-dire que les variances des différents groupes sont homogènes.

Test non-paramétrique. Une problématique fondamentale concerne le respect de ces hypothèses, en particulier celles de normalité et d'homoscédasticité. En effet, le non respect de la normalité des données entraîne un faible degré de signification ainsi qu'une faible puissance de l'étude ([Wilcox \[2014\]](#); [Tavakol et Wilcox \[2013\]](#)). Or ces hypothèses sont rarement respectées dans la pratique.

La normalité des variables est généralement vérifiée via des procédures statistiques tels que les coefficients de dissymétrie et d'aplatissement ou encore le test de Kolmogorov-Smirnov ou via des méthodes graphiques. Afin de valider l'hypothèse d'homoscédasticité, le test de Levene est normalement employé. Pourtant, de nombreuses publications ont montré que ces tests ne sont pas satisfaisants, principalement parce qu'ils n'ont pas assez de puissance afin de repérer les situations pour lesquelles le non-respect des hypothèses est une préoccupation réelle ([Tavakol et Wilcox \[2013\]](#)). Même de simples écarts de l'hypothèse de normalité peuvent engendrer des erreurs lors de la réalisation du test.

Les tests non-paramétriques parviennent à surmonter les problèmes liés à des distributions non normales ou à l'hétéroscédasticité. Il existe plusieurs techniques sur lesquelles se basent ces tests non-paramétriques. Une possibilité est de comparer les groupes en utilisant la médiane ou la moyenne tronquée au lieu de la moyenne arithmétique. Également des méthodes statistiques robustes plus récentes sont proposées dans la littérature et introduites dans le livre de Wilcox [2012].

Le diagramme de la figure 5.1 montre quelques exemples de tests paramétriques et leurs équivalents non-paramétriques pour étudier des données continues (du Prel et al. [2010]).

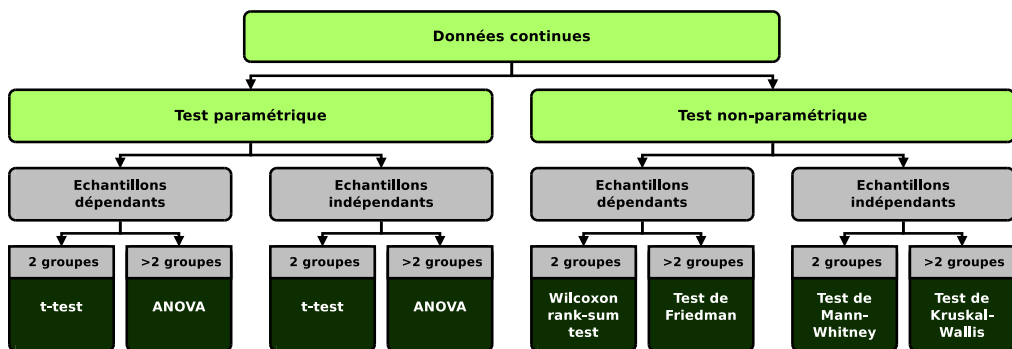


FIGURE 5.1 – Diagramme pour sélectionner un test statistique pour la comparaison de groupes dans le cas de données continues.

Tests post hoc. Des tests post hoc sont généralement effectués à la suite d'un rejet de l'hypothèse nulle par une ANOVA. Dans le cas d'un facteur avec seulement deux niveaux, rejeter l'hypothèse nulle indique qu'il existe des différences significatives entre les deux populations. En revanche, si le facteur présente plus de deux niveaux, alors le rejet permet de conclure qu'au moins une population est différente des autres sans préciser la ou lesquelles. L'utilisation de test post hoc est nécessaire afin de déterminer où se situent les différences.

De nombreux tests de comparaisons multiples (post hoc) sont présentés dans la littérature (Jaccard et al. [1984]), le plus efficace dépendant des spécificités des situations. Le test le plus basique et le plus couramment employé est la correction de Bonferonni. Cependant le *false discovery rate*, introduit par Benjamini et Hochberg [1995] puis amélioré par Benjamini et Yekutieli [2001], est recommandé car plus récent et plus puissant statistiquement.

Cas des données fonctionnelles. Un défi statistique récent dans le domaine de l'analyse de données fonctionnelles est la comparaison entre courbes ou ensembles de courbes. De tels problèmes sont pris en compte dans le cadre de l'analyse de la variance fonctionnelle (FANOVA). De nombreux travaux ont été réalisés afin d'étendre l'ANOVA classique pour la FANOVA (Lin [2000]; Angelini et al. [2003]). En revanche, peu d'attention a été portée pour la déduction ou le test d'hypothèses fonctionnelles selon Abramovich et Angelini [2006].

Une approche simple afin de tester des données via des modèles FANOVA consiste à réaliser une série de tests d'ANOVA scalaire pour comparer un ensemble de courbes à plusieurs abscisses. Cette méthode entraîne un problème de multiplicité non négligeable en raison du nombre important de tests simultanés. Ignorer cet effet de multiplicité provoque des erreurs de type 1. La correction la plus couramment employée est encore une fois celle de Bonferroni qui est cependant connue pour avoir une puissance très faible. Une autre approche est de traiter les données fonctionnelles comme des vecteurs multivariés et d'appliquer des techniques d'ANOVA multivariées traditionnelles accompagnées parfois de procédures de réduction de la dimension (Chen [1994]). Cependant la malédiction de la dimension (Bellman [1957]) rend ces approches problématiques.

De nombreuses autres méthodes existent dans la littérature mais la plupart d'entre elles se basent sur une hypothèse de distribution normale des données. Cuesta-Albertos et Febrero-Bande [2010] proposent une méthode efficace, flexible et permettant d'utiliser des modèles complexes d'ANOVA sans imposer d'hypothèses de normalité ni d'homoscédasticité. L'idée principale de cette méthode est d'utiliser des projections aléatoires afin de transformer les données fonctionnelles en données scalaires puis de résoudre ce simple problème d'ANOVA et de déduire des conclusions pour les données fonctionnelles en regroupant les informations issues de nombreuses projections en corrigeant les valeurs p par la méthode de *False Discovery Rate*. Une technique d'ANOVA, conseillée par les auteurs, pour utiliser conjointement avec cette approche est celle proposée par Brunner et al. [1997].

Plus d'informations sur l'analyse de données fonctionnelles avec les méthodes d'ANOVA peuvent être trouvées dans le livre de Zhang [2013].

5.3 Sélection de caractéristiques et classification des données

L'extraction de caractéristiques a finalement pour but de permettre de décrire les données ou encore de les classifier. Cependant, toutes les mesures effectuées ne sont pas nécessairement pertinentes et incorporent souvent des redondances. Ces informations peuvent être préjudiciables pour l'interprétation, ou la classification des données. De ce fait, des outils de sélection des caractéristiques les plus judicieuses sont proposés dans la littérature. C'est un champ de recherche de plus en plus important à mesure que les techniques automatiques d'analyses quantitatives d'images deviennent de plus en plus utilisées.

Les méthodes de sélection de caractéristiques analysent les mesures afin d'en sélectionner les sous-ensembles les plus pertinents vis-à-vis de l'étude en cours. Outre réduire la dimension de l'espace de caractéristiques, ces méthodes permettent de diminuer la spécialisation d'un modèle, de le rendre plus rapide et de mieux interpréter les procédés ayant permis la génération des données (Guyon et Elisseeff [2003]). Les techniques de sélection peuvent être regroupées en trois familles, les méthodes dites *filter*, *wrapper* et *embedded* et sont étudiées dans l'article de Saeys et al. [2007].

Les méthodes *filter* déterminent la pertinence des caractéristiques indépendamment de l'algorithme de classification et suppriment celles ayant un score trop faible. Ces approches se basent donc sur une mesure de la signification des caractéristiques. Les plus communes d'entre elles incluent l'information mutuelle, les mesures de corrélation, des mesures de distance ou encore l'utilisation de tests statistiques tels que l'ANOVA.

Les techniques *wrapper* intègrent le modèle de classification dans la sélection, un groupe de caractéristiques est alors évalué en entraînant et testant le classifieur. La troisième famille de méthodes, les *embedded*, recherchent le sous-ensemble optimal lors de la construction du classifieur. Comme les techniques, *wrapper*, elles sont donc spécifiques à un algorithme de classification mais requièrent en plus une diversité suffisante dans les données d'apprentissage. Plus d'informations sur les méthodes de sélection de caractéristiques peuvent être trouvées dans le manuscrit de thèse de Chouaib [2011].

À partir des caractéristiques sélectionnées, il est possible de mettre en œuvre des algorithmes de classification afin d'identifier les maladies ou leurs états en comparant les mesures issues d'un échantillon avec un ensemble de caractéristiques pré-calculées et correspondant à des états connus. De nombreuses techniques sont présentes dans la littérature allant d'un simple

plus proches voisins à des méthodes complexes mettant en œuvre plusieurs algorithmes. Le choix de la méthode dépend évidemment de la nature des données et des caractéristiques extraites. Une analyse exhaustive des méthodes de classification peut être obtenue dans l'ouvrage de [Duda et al. \[2012\]](#).

5.4 Applications

Les applications de ce chapitre se concentrent sur les données histologiques présentées dans la partie 1.5. Plusieurs caractéristiques ont été extraites à partir de ces images. Tout d’abord, les objets biologiques ont été détectés puis extraits dans le chapitre 2. Ensuite, les relations spatiales entre noyaux cellulaires puis les interactions entre noyaux et vascularisations sont mesurées dans le chapitre 3, la texture est également extraite comme présenté dans le chapitre 4. Finalement, cette partie étudie les capacités de description et de discrimination de ces caractéristiques afin de différencier les modèles animaux d’implantations de tumeurs.

5.4.1 Densité

Une fois les objets biologiques présents dans les images détectés et extraits, il est possible de déterminer leurs densités en fonction de leurs natures (voir partie 5.2.1). Ici, les noyaux peuvent être considérés comme des objets ponctuels tandis que les vascularisations varient de simples objets bien définis à de vastes réseaux de structures imbriquées les unes dans les autres (voir figure 5.2), ce qui rend leur quantification complexe. Pour cela, les vaisseaux sanguins sont considérés comme des objets surfaciques.

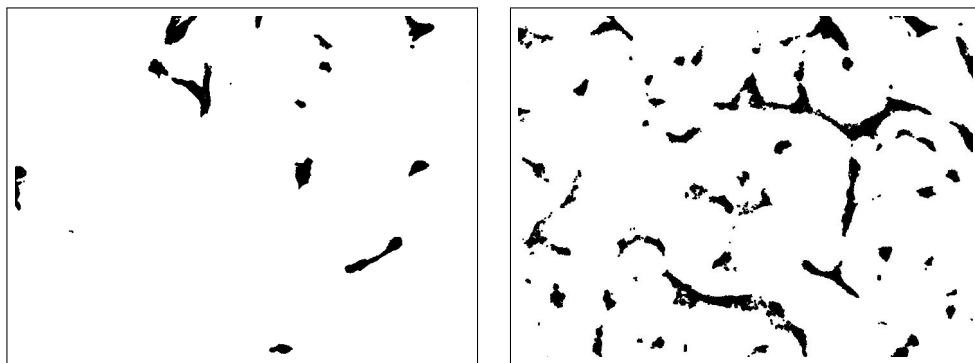


FIGURE 5.2 – Illustration de la variabilité de la structure des vascularisations.

Une description visuelle par *boxplots* des densités de ces deux objets selon les jours après implantation et les modèles des tumeurs est montrée dans la figure 5.3. Le modèle ectopique montre une diminution de la densité des vascularisations combinée avec une hausse de la densité des noyaux entre les jours 15 et 21, ce qui suggère que la croissance tumorale est plus importante que la formation de nouveaux vaisseaux sanguins. La densité

de vascularisation semble plus stable dans le modèle orthotopique et suit globalement la même évolution que la densité des cellules.

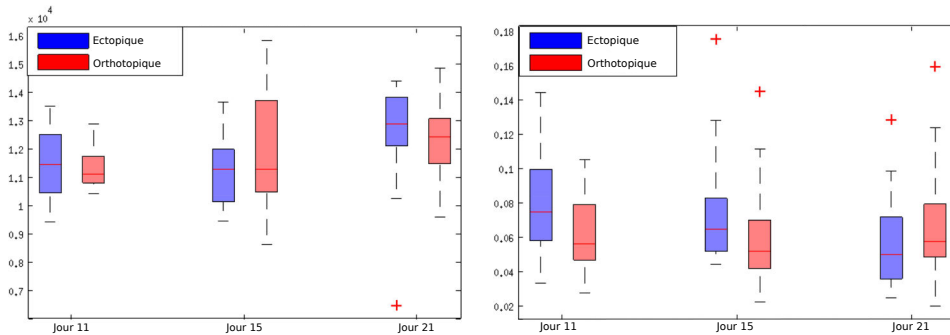


FIGURE 5.3 – *Boxplots* des densités des noyaux (à gauche) et des vascularisations (à droite) pour chaque modèle et chaque jour après implantation.

5.4.2 Corrélation

D’après [Armstrong \[2003\]](#), analyser les possibles associations entre plusieurs variables peut être profitable du point de vue de l’étude des maladies. Ainsi, afin d’observer la présence de corrélations entre les densités des noyaux et des vascularisations, les deux variables sont tracées dans un *scatterplot* dans la figure 5.4.

Visuellement, une forte corrélation négative est visible pour le 11^{ième} jour du modèle ectopique alors qu’il n’y a pas d’association remarquable pour le même jour avec le modèle orthotopique. Afin d’obtenir une estimation quantitative précise du degré de corrélation présent dans les *scatterplots*, le coefficient de corrélation de Spearman r est calculé puis sa signification est évaluée tel que présenté dans la partie 5.2.2. Le tableau 5.3 donne les coefficients de corrélation accompagnés des valeurs p correspondantes. Les associations entre densités ne sont pas les mêmes suivant les modèles. Dans le cas du modèle ectopique, une corrélation négative significative est présente lors du 11^{ième} jour après implantation alors que le modèle orthotopique montre une corrélation significative positive uniquement lors du 21^{ième} jour. Ces résultats sont cohérents avec les observations réalisées à partir des *boxplots*.

5.4.3 Application de l’ANOVA

Afin de déterminer la présence de différences significatives entre les modèles et les jours après implantation selon les différentes caractéristiques,

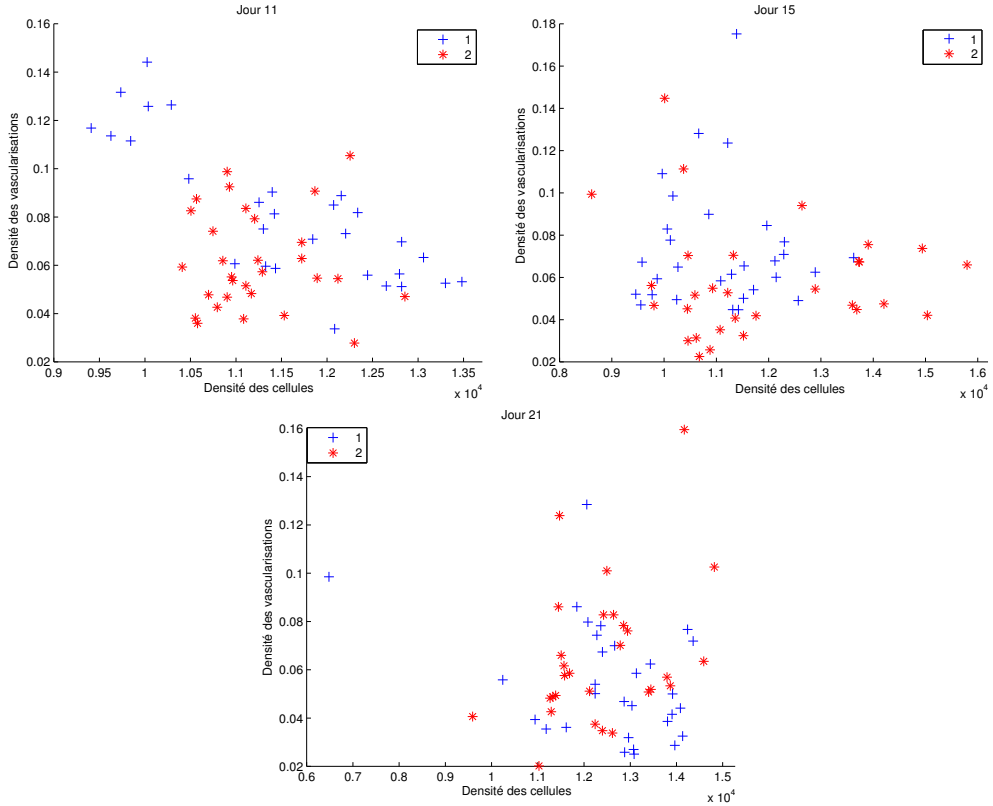


FIGURE 5.4 – *Scatterplots* des densités des cellules et des vascularisations pour les deux modèles d’implantation (1 : ectopique et 2 : orthotopique) pour chaque jour.

TABLE 5.3 – Coefficients de corrélation de Spearman (r) entre les densités des noyaux et des vaisseaux ainsi que les valeurs p correspondantes. Les valeurs p notées en rouge sont celles inférieures à $\alpha = 0,05$ et indiquent le rejet de l’hypothèse nulle.

Modèle	Ectopique			Orthotopique			
	Jour	11 ^{ième}	15 ^{ième}	21 ^{ième}	11 ^{ième}	15 ^{ième}	21 ^{ième}
r		-0,7937	0,0158	-0,2875	0,0102	0,0065	0,4241
valeur p		3×10^{-7}	0,9345	0,1235	0,9572	0,9730	0,0227

un test de l’hypothèse H_0 : « il n’y a pas de différences significatives entre les groupes » et de son alternative H_1 : « il existe des différences significatives entre au moins deux groupes » est réalisé par l’analyse de la variance selon deux facteurs :

- Jour après implantation : avec trois niveaux 11, 15 et 21.

— Modèle d’implantation : avec deux niveaux ectopique et orthotopique.

Le degré de signification de ces tests est fixé à $\alpha = 0,05$.

Le modèle théorique de l’ANOVA à effet mixte utilisée peut être écrit de la façon suivante :

$$y_{ijk} = \Lambda + D_i + M_j + \gamma_{ij} + b_{ijk} + \epsilon_{ijk}, \quad (5.6)$$

où y est la caractéristique à étudier, $i = 1, 2, 3$ et $j = 1, 2$ indiquent les niveaux respectivement des facteurs jours et modèles, $k = 1, \dots, 180$ est l’indice de la mesure et correspond donc au numéro de l’image, Λ est l’effet global, ϵ_{ijk} est l’erreur de mesure, et γ_{ij} représente l’interaction entre les deux facteurs. Puisque plusieurs mesures y_{ij} sont effectuées sur chaque souris, les dépendances entre ces mesures doivent être prises en compte ainsi que l’effet aléatoire introduit par les souris. Pour cela, un terme supplémentaire b_{ijk} , qui permet de capturer les variabilités entre souris et de les traiter comme un effet aléatoire, est ajouté au modèle.

Les résultats des tests statistiques, respectivement pour les densités des cellules et des vascularisations, les fonctions K robustes, et les fonctions IPA, sont donnés dans le tableau 5.4. L’ANOVA mise en œuvre pour les données fonctionnelles est celle de [Cuesta-Albertos et Febrero-Bande \[2010\]](#) précédemment décrite dans la partie 5.2.2. Les problèmes d’ANOVA classiques ainsi obtenus, ainsi que les ANOVA pour les données non fonctionnelles, sont résolus avec la méthode robuste proposée par [Brunner et al. \[1997\]](#). Puisque les tests sont réalisés à partir de données issues d’un sous-ensemble réduit d’une population, il n’est pas possible d’affirmer avec certitude si l’hypothèse nulle est vraie ou fausse. Le test statistique permet simplement de déterminer si cette hypothèse est probable ou non. De ce fait, les hypothèses peuvent être rejetées ou non rejetées mais en aucun cas acceptées. Un rejet, valeur p inférieure à α , indique donc que les variations entre les groupes sont plus importantes que celles intra groupes, ce qui inclut les variations entre échantillons et entre souris.

Ces résultats montrent que les informations issues des noyaux cellulaires et des vascularisations, qu’elles soient du premier ou du second ordre, sont complémentaires afin de distinguer les modèles et les jours après implantation. Toutes les caractéristiques révèlent des interactions entre les jours et les modèles, les variations dans les valeurs des mesures entre les jours dépendent donc des niveaux des modèles et inversement. La fonction IPA mesure les interactions spatiales entre ces deux populations et est donc la plus pertinente dans ce contexte. Le test d’ANOVA permet uniquement de déterminer la présence de différences significatives mais ne précise pas où se situent ces différences. Afin d’avoir une analyse plus complète, des tests

TABLE 5.4 – Valeurs p obtenues lors des tests de H_0 pour les différentes caractéristiques suivant les deux facteurs et leur interaction. Les valeurs p notées en rouge sont celles inférieures à $\alpha = 0,05$ et indiquent le rejet de l’hypothèse nulle.

	Modèles	Jours	Interaction
Densité des noyaux	$6,73 \times 10^{-1}$	$9,28 \times 10^{-6}$	$4,57 \times 10^{-2}$
Densité des vaisseaux	$3,4 \times 10^{-2}$	$6,66 \times 10^{-2}$	$4,9 \times 10^{-3}$
Fonction K robuste	$8,69 \times 10^{-2}$	$1,3 \times 10^{-4}$	$4,1 \times 10^{-2}$
Fonction IPA	$2,94 \times 10^{-2}$	$1,89 \times 10^{-3}$	$1,42 \times 10^{-3}$

d’ANOVA deux à deux ont été effectués avec une correction des valeurs p réalisées par la méthode de *false discovery rate*. Le tableau 5.5 montre les résultats de tous ces tests pour les différentes caractéristiques étudiées.

Ces tests montrent que les différentes caractéristiques permettent de distinguer des aspects différents de l’étude. Les résultats de la densité des vascularisations soulignent les faibles variations dans le cas du modèle orthotopique, ce qui est cohérent avec les conclusions tirées de l’analyse descriptive précédente. Les mesures du premier et du second ordre sur les noyaux cellulaires ne suffisent pas pour discriminer les deux modèles et permettent de séparer les mêmes jours. Finalement, la mesure la plus discriminante pour cette étude est la fonction d’interaction spatiale entre cellules et vascularisations proposée dans cette thèse. Cet aspect n’est pas développé dans la littérature or ces tests statistiques confirment l’intérêt de cette mesure nouvelle pour l’étude des cancers à partir d’images histologiques et valident les conclusions de [Armstrong \[2007\]](#) soulignant que les interactions entre deux populations sont souvent plus pertinentes que leurs propriétés intrinsèques.

5.4.4 Sélection de caractéristiques et classification

La sélection de caractéristiques permet d’étudier la pertinence de chacune de leurs dimensions contrairement aux tests statistiques qui analysent chaque mesure comme un tout. Pour explorer les performances de ces caractéristiques dans le cadre d’une prise de décision automatique d’appartenance d’un échantillon à un modèle, plusieurs algorithmes de classifications sont comparés ainsi que l’intégration d’une méthode de sélection de caractéristiques. Les performances ainsi obtenues sont montrées dans le tableau 5.6. Les validations croisées ont été effectuées en *leave one out* en prenant soin de ne pas prendre en compte les données issues de la même

TABLE 5.5 – Valeurs p corrigées obtenues lors des test deux à deux de H_0 pour les différentes caractéristiques. Les tests entre les jours du modèle ectopique sont représentés sur fond gris clair alors que ceux du modèle orthotopique sont sur fond gris sombre. La diagonale donne les résultats des tests entre les deux modèles pour chaque jour. Les valeurs p notées en rouge sont celles inférieures à $\alpha = 0,05$ et indiquent le rejet de l’hypothèse nulle.

Densité des noyaux			
Jour	11	15	21
11	$2,37 \times 10^{-1}$	$1,35 \times 10^{-1}$	$2,7 \times 10^{-4}$
15	$2,1 \times 10^{-1}$	$1,15 \times 10^{-1}$	$2,37 \times 10^{-1}$
21	$1,39 \times 10^{-2}$	$2,7 \times 10^{-4}$	$9,85 \times 10^{-1}$
Densité des vaisseaux			
Jour	11	15	21
11	$1,35 \times 10^{-2}$	$6,57 \times 10^{-1}$	$6,57 \times 10^{-1}$
15	$4,81 \times 10^{-1}$	$8,59 \times 10^{-2}$	$4,93 \times 10^{-1}$
21	$2,79 \times 10^{-3}$	$3,33 \times 10^{-2}$	$8,99 \times 10^{-1}$
Fonction K robuste			
Jour	11	15	21
11	$7,98 \times 10^{-2}$	$3,74 \times 10^{-1}$	$2,47 \times 10^{-2}$
15	$7,98 \times 10^{-2}$	$5,18 \times 10^{-1}$	$8,24 \times 10^{-2}$
21	$1,73 \times 10^{-2}$	$4,5 \times 10^{-3}$	$9,98 \times 10^{-1}$
Fonction IPA			
Jour	11	15	21
11	$5,96 \times 10^{-1}$	$7,25 \times 10^{-3}$	$6,17 \times 10^{-2}$
15	$4,2 \times 10^{-2}$	$7,2 \times 10^{-3}$	$1,54 \times 10^{-2}$
21	$4,21 \times 10^{-2}$	$5,98 \times 10^{-2}$	$9,93 \times 10^{-1}$

souris que l’image de test pour ne pas introduire de biais dans l’étude.

Ces performances confirment les résultats des tests statistiques, à savoir que malgré la présence locale de différences significatives, ces caractéristiques ne permettent pas à elles seules d’attribuer le modèle correspondant à chaque échantillon. Les performances sont très proches du hasard à l’exception de la texture avec la méthode du plus proche voisin et de la concaténation des fonctions K et IPA avec réduction de la dimension et classification par analyse discriminante linéaire (LDA). L’algorithme uti-

TABLE 5.6 – Taux de reconnaissance des modèles suivant les caractéristiques employées. Les caractéristiques de texture ici utilisées sont celles proposées dans le chapitre 4. Wrapper PPV désigne un algorithme de plus proche voisin utilisé avec une sélection de caractéristique de type *sequential forward*.

Caractéristique	PPV	Wrapper PPV	LDA	SVM
Densités	47,8 %	/	/	44,4 %
Fonction <i>K</i> robuste	54,9 %	46,1 %	48,9 %	51,7 %
Fonction <i>IPA</i>	41,6 %	54,5 %	52,8 %	47,2 %
Texture	68,0 %	46,1 %	51,7 %	53,9 %
<i>K</i> et <i>IPA</i>	54,5 %	54,5 %	59,6 %	53,4 %
Densités et texture	42,1 %	42,1 %	51,7 %	55,6 %
Densités, <i>K</i> et <i>IPA</i>	48,9 %	48,9 %	49,4 %	49,4 %
<i>K</i> , <i>IPA</i> et texture	54,5 %	54,5 %	46,6 %	53,4 %

lisé pour la sélection des caractéristiques n’améliore pas sensiblement les performances, en particulier pour l’analyse de la texture, ce qui s’explique par la sensibilité des méthodes *wrapper* aux minima locaux. Tout cela souligne la complexité de ce problème qui consiste à différencier des images visuellement très similaires montrant le même tissu affecté par la même maladie mais évoluant dans un milieu différent. Également, cela montre que les densités, les relations spatiales et l’interaction permettent de détecter la présence de différences et de les interpréter biologiquement dans un contexte clinique mais ne permettent pas de reconnaissance automatique. Pour cette tâche, la caractéristique qui s’avère être la plus prometteuse reste l’information de texture qui ne peut être interprétée biologiquement.

5.5 Conclusion et perspectives

L'intérêt de l'extraction de caractéristiques à partir d'images histologiques réside soit dans la capacité qu'elles offrent afin de comprendre la pathologie présente, soit dans l'aide au diagnostic. D'un point de vue clinique, cela permet donc d'étudier et d'interpréter les mécanismes sous-jacents des maladies et leurs impacts sur les objets biologiques et leurs relations éventuelles. Cette étude est donc exploratoire et n'intègre donc que rarement une vérité terrain, ce qui en fait un domaine plein de défis. Les tests statistiques sont donc une étape indispensable dans le cadre clinique afin d'étudier la signification de chaque mesure dans un contexte médical spécifique. Le second aspect, plus applicatif, a pour but d'assister les pathologistes dans leur prise de décision en se basant sur des mesures quantitatives et des décisions reproductibles.

Dans ce chapitre, la signification des caractéristiques extraites sur les images histologiques présentées dans la partie 1.5 est étudiée. Une analyse quantitative des densités des deux populations d'objets biologiques est présentée ainsi que leur corrélation, soulignant que l'association de ces deux densités est différente selon le modèle d'implantation de la tumeur. Une analyse de la variance souligne l'importance de combiner les informations issues des deux populations d'objets, que cette mesure soit du premier ou du second ordre, dans un contexte d'étude clinique exploratoire des deux modèles d'implantation de tumeurs. En revanche, l'utilisation de l'information de texture est plus pertinente pour l'aide au diagnostic dans ce cadre d'étude.

Les suites à donner à l'aspect clinique concernent l'analyse de données fonctionnelles. Ainsi, une piste serait d'étudier plus avant ces données via les méthodes spécialisées décrites dans le livre de [Ramsay et al. \[2009\]](#). Ces outils permettent une représentation plus adaptée des mesures afin de mieux percevoir certaines caractéristiques et sources de variations pour mener des analyses exploratoires telles qu'une analyse en composante principale fonctionnelle. Également, les méthodes de statistiques déductives sont en développement perpétuel. De ce fait, des méthodes modernes plus robustes de test et de correction des valeurs p méritent d'être mises en œuvre. Du point de vue de l'aspect applicatif de l'aide au diagnostic, il serait avantageux de tester le comportement de différentes méthodes de sélection de caractéristiques ainsi que de différents algorithmes de classification, plus adaptés aux particularités de ce problème et de ces données. Finalement, une validation approfondie sur plus de données serait intéressante afin de confirmer les conclusions actuelles.

Conclusion générale et perspectives

L'objectif global de ce manuscrit de thèse est d'explorer les défis associés au développement d'algorithmes robustes pour les différentes tâches liées à l'évaluation quantitative des images histopathologiques. L'analyse quantitative de données histologiques est un domaine de recherche émergent qui gagne en importance avec les avancées technologiques de l'imagerie numérique des tissus. Ce champ de recherche récent présente des défis importants et des problématiques uniques. En effet, différentes images de tissus associées avec différentes maladies manifestent des caractéristiques particulières, à la fois en terme de morphologie des tissus et de critères décisifs. Par conséquent, il n'existe pas de système d'analyse unique pour toutes les problématiques histopathologiques. En revanche, plusieurs étapes d'analyse d'images restent communes à la plupart des applications. L'analyse d'images histopathologiques peut avoir trois objectifs : l'aide au diagnostic, l'évaluation de la sévérité d'une maladie ainsi que l'étude et l'interprétation des mécanismes sous-jacents des maladies et leurs impacts sur les objets biologiques. Dans le cadre de l'étude et de la compréhension des différences entre les modèles animaux d'implantation de tumeurs et leurs mécanismes de croissance, les travaux de cette thèse visent à concevoir des algorithmes afin d'aborder les problèmes génériques de l'analyse d'images histopathologiques tels que la détection des objets biologiques, l'étude de leur organisation spatiale et de leurs interactions, ainsi que de la texture du tissu.

Le chapitre 2 traite de la première étape d'analyse histologique qui est la séparation des marqueurs et la détection des objets biologiques d'intérêt. En premier lieu, une technique simple et rapide est proposée pour la séparation des marqueurs les plus utilisés dans la littérature qui sont le DAB et l'éosine contre marqués par l'hématoxyline. Cette approche offre des résultats comparables à la déconvolution des couleurs avec l'avantage d'être moins bruités. Une méthode de détection des objets ponctuels se basant sur la carte topographique est ensuite proposée puis validée expérimenta-

lement. À l'avenir, la conception d'une technique à même de déterminer automatiquement le pas de quantification de la carte topographique suivant la qualité des images serait intéressante afin d'économiser du temps de calculs et de la mémoire. De plus, ces différentes méthodes gagneraient à être validées sur de plus grands ensembles de données acquises par différents microscopes et scanners et annotées par différents pathologistes.

Une fois les emplacements des objets biologiques connus, le chapitre 3 traite de l'analyse quantitative de l'organisation spatiale des objets ponctuels, représentés par les noyaux cellulaires dans ce cadre applicatif, ainsi que leurs interactions avec les objets surfaciques que sont les vascularisations. Ces mesures particulières permettent de bien comprendre les mécanismes sous-jacents des tissus. Une étude comparative des différentes méthodes statistiques pour étudier les relations spatiales entre points est tout d'abord réalisée. Ensuite, une mesure statistique originale inspirée par la fonction K de Ripley est introduite afin de capturer les dépendances entre les noyaux et le réseau vasculaire. De plus, un calcul analytique des valeurs limites de cette fonction est effectué afin de déterminer si leurs distributions spatiales suivent une hypothèse d'indépendance ou présentent des interactions est présenté. De nombreuses pistes de recherches restent ouvertes dans ce domaine, notamment la modélisation paramétrique des distributions des objets ainsi que l'étude de l'interaction entre les différents objets surfaciques.

De nombreuses études ont confirmé que la présence d'une maladie, en particulier le cancer, dans un tissu altère sa texture. De ce fait, le chapitre 4 discute des différentes méthodes permettant de capturer cette information. Une approche générique basée sur la modification de la résolution des images et le modèle complété des motifs binaires locaux a été proposée et validée sur plusieurs jeux de données. Ces travaux ont été effectués lors d'un stage de six mois et demi à l'École Polytechnique de Montréal. Une technique afin de déterminer a priori la résolution optimale pour différencier les maladies reste à définir ainsi que le regroupement des informations issues des différentes échelles pertinentes.

Finalement, le chapitre 5 discute des méthodes permettant d'interpréter les mesures et d'étudier leurs capacités discriminatives ainsi que leurs différences significatives dans le cadre applicatif des modèles animaux de cancer. Dans ce chapitre la signification de chaque caractéristique mesurée sur les images est étudiée par rapport au contexte d'application. Le domaine des statistiques déductives est en continuelle évolution avec la proposition de méthodes toujours plus robustes et précises. Il convient donc de confirmer les résultats obtenus par l'utilisation des futures techniques d'analyse. Également, il serait intéressant d'appliquer les outils spécifiquement adap-

tés aux données fonctionnelles afin de mieux explorer ces mesures et leurs variations.

Les travaux de recherche présentés dans ce manuscrit ne peuvent couvrir qu'une infime partie des problématiques dans l'analyse d'images histopathologiques en raison de la myriade de maladies qui restent encore à étudier. De plus, l'un des obstacles les plus importants pour la propagation de l'utilisation des méthodes automatiques d'analyse histopathologique est l'absence de standards et de formats de données consentis à grande échelle. Cette avancée permettrait de soutenir le développement de détecteurs d'objets biologiques, d'algorithmes pour l'estimation de la qualité et du marquage ainsi que de la prise de décision. Il serait très profitable que la communauté scientifique s'entende sur de simples standards et modalités d'échange des données ainsi que sur l'établissement et le partage des vérités terrains. Également, de nombreuses recherches portent sur les marqueurs dans le but de faciliter l'automatisation des analyses, que ce soit au niveau du marqueur lui-même, de son application sur l'échantillon, de la segmentation, ou encore de l'acquisition de l'image. À long terme, toutes les perspectives évoquées précédemment se doivent donc de prendre en compte l'avancée rapide du domaine et s'y adapter.

Bibliographie

- Abramovich, F., Angelini, C., 2006. Testing in mixed-effects FANOVA models. *Journal of Statistical Planning and Inference* 136 (12), 4326–4348.
Cité page 131
- Acar, E., Plopper, G. E., Yener, B., 2012. Coupled analysis of in vitro and histology tissue samples to quantify structure-function relationship. *PLoS ONE* 7 (3), e32227.
Cité page 52
- Acharya, U., Dua, S., Du, X., Vinitha Sree, S., Chua, C. K., 2011. Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Transactions on Information Technology in Biomedicine* 15 (3), 449–455.
Cité page 107
- Ahonen, T., Matas, J., He, C., Pietikäinen, M., 2009. Rotation invariant image description with local binary pattern histogram Fourier features. In : *Image Analysis*. Springer, pp. 61–70.
Cité page 98
- Al-Kadi, O. S., 2010. Texture measures combination for improved meningioma classification of histopathological images. Elsevier, *Pattern Recognition* 43 (6), 2043–2053.
Cité page 15
- Al-Kofahi, Y., Lassoued, W., Grama, K., Nath, S. K., Zhu, J., Oueslati, R., Feldman, M., Lee, W. M. F., Roysam, B., 2011. Cell-based quantification of molecular biomarkers in histopathology specimens. *Histopathology* 59 (1), 40–54.
Cité page 11
- Al-Kofahi, Y., Lassoued, W., Lee, W., Roysam, B., 2010. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering* 57 (4), 841–852.
Cité page 32
- Alsheh Ali, M., Seguin, J., Fischer, A., Mignet, N., Wendling, L., Hurtut, T., 2013. Comparison of the spatial organization in colorectal tumors

- using second-order statistics and functional ANOVA. In : Proc. ISPA. IEEE, pp. 257–261. *Cité page 39*
- Angelini, C., De Canditiis, D., Leblanc, F., 2003. Wavelet regression estimation in nonparametric mixed effect models. *Journal of Multivariate Analysis* 85 (2), 267–291. *Cité page 131*
- Armstrong, R. A., 2003. Measuring the degree of spatial correlation between histological features in thin sections of brain tissue. *Neuropathology* 23 (4), 245–253. *Cité page 135*
- Armstrong, R. A., 2007. Quantitative microscopic analysis of histological sections of brain tissue. *Modern Research and Educational Topics in Microscopy*, 442–452. *Cité page 138*
- Baddeley, A. J., Møller, J., Waagepetersen, R., 2000. Non and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* 54 (3), 329–350. *Cité page 61*
- Balsat, C., Blacher, S., Signolle, N., Beliard, A., Munaut, C., Goffin, F., Noel, A., Foidart, J., Kridelka, F., 2011. Whole slide quantification of stromal lymphatic vessel distribution and peritumoral lymphatic vessel density in early invasive cervical cancer : A method description. *ISRN oObstetrics and Gynecology* 354861. *Cité page 53*
- Barot, S., Gignoux, J., Menaut, J., 1999. Demography of a savanna palm tree : Predictions from comprehensive spatial pattern analyses. *Ecology* 80 (6), 1987–2005. *Cité page 62*
- Basavanhally, A., Ganesan, S., Feldman, M., Shih, N., Mies, C., Tomaszewski, J., Madabhushi, A., 2013. Multi-field-of-view framework for distinguishing tumor grade in ER breast cancer from entire histopathology slides. *IEEE Transactions on Biomedical Engineering* 60 (8), 2089–2099. *Cité page 123*
- Bedford, T., Van den Berg, J., 1997. A remark on the Van Lieshout and Baddeley J-function for point processes. *Advances in Applied Probability*, 19–25. *Cité page 57*
- Bellman, R. E., 1957. *Dynamic Programming*. Rand Corporation Research Study. Princeton University Press. *Cité page 131*
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300. *Cité page 130*

- Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188. *Cité page 130*
- Bernsen, H. J. J. A., Rijken, P. F. J. W., Hagemeyer, N. E. M., Van Der Kogel, A. J., 1999. A quantitative analysis of vascularization and perfusion of human glioma xenografts at different implantation sites. *Microvascular Research* 57 (3), 244–257. *Cité page 53*
- Berry, A. C., 1941. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society* 49 (1), 122–136. *Cité page 70*
- Besag, J., Diggle, P. J., 1977. Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 327–333. *Cité page 64*
- Besag, J. E., 1977. Comments on Ripley’s paper. *Journal of the Royal Statistical Society B39*, 193–195. *Cité pages 59 et 64*
- Bilgin, C., Bullough, P., Plopper, G. E., Yener, B., 2010. ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Mining and Knowledge Discovery* 20 (3), 416–438. *Cité page 52*
- Bilgin, C., Demir, C., Nagi, C., Yener, B., 2007. Cell-graph mining for breast tissue modeling and classification. In : *Proc. EMBS. IEEE*, pp. 5311–5314. *Cité page 52*
- Bock, R., Meier, J., Nyúl, L. G., Hornegger, J., Michelson, G., 2010. Glaucoma risk index :Automated glaucoma detection from color fundus images. *Medical Image Analysis* 14 (3), 471–481. *Cité page 107*
- Boucheron, L. E., 2008. Object- and Spatial-level Quantitative Analysis of Multispectral Histopathology Images for Detection and Characterization of Cancer. Thèse de Doctorat, University of California at Santa Barbara, Santa Barbara, CA, USA. *Cité pages 16, 26, 52 et 104*
- Boykov, Y., Funka-Lea, G., 2006. Graph cuts and efficient ND image segmentation. *International Journal of Computer Vision* 70 (2), 109–131. *Cité page 31*
- Breward, C. J. W., Byrne, H. M., Lewis, C. E., 2001. Modelling the interactions between tumour cells and a blood vessel in a microenvironment within a vascular tumour. *European Journal of Applied Mathematics* 12 (05), 529–556. *Cité page 77*

- Brey, E. M., Lalani, Z., Johnston, C., Wong, M., McIntire, L. V., Duke, P. J., Patrick, C. W., 2003. Automated selection of DAB-labeled tissue for immunohistochemical quantification. *Journal of Histochemistry & Cytochemistry* 51 (5), 575–584. *Cité pages 27 et 28*
- Brühlhart, M., Traeger, R., 2005. An account of geographic concentration patterns in Europe. *Regional Science and Urban Economics* 35 (6), 597–624. *Cité page 58*
- Brunner, E., Dette, H., Munk, A., 1997. Box-type approximations in non-parametric factorial designs. *Journal of the American Statistical Association* 92 (440), 1494–1502. *Cité pages 131 et 137*
- Caicedo, J. C., González, F. A., Romero, E., 2011. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. *Journal of Biomedical Informatics* 44 (4), 519–528. *Cité page 17*
- Caselles, V., Morel, J. M., 1999. Topographic maps and local contrast changes in natural images. *International Journal of Computer Vision* 33, 5–27. *Cité page 39*
- Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., Moss, R. H., 2007. A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics* 31 (6), 362–373. *Cité page 16*
- Chan, T. F., Vese, L. A., 2001. Active contours without edges. *IEEE Transactions on Image Processing* 10 (2), 266–277. *Cité page 31*
- Chen, C., Wang, W., Ozolek, J. A., Rohde, G. K., 2013. A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching. *Cytometry Part A* 83 (5), 495–507. *Cité page 31*
- Chen, J. C., 1994. Testing for no effect in nonparametric regression via spline smoothing techniques. *Annals of the Institute of Statistical Mathematics* 46 (2), 251–265. *Cité page 131*
- Chouaib, H., 2011. Sélection de caractéristiques : méthodes et applications. Thèse de Doctorat, Université Paris Descartes, Paris, France. *Cité page 132*
- Chu, A., Sehgal, C. M., Greenleaf, J. F., 1990. Use of gray value distribution of run lengths for texture analysis. *Elsevier, Pattern Recognition Letters* 11 (6), 415–419. *Cité page 92*

- Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Routledge Academic. *Cité page 127*
- Cohen, J., Cohen, P., West, S. G., Aiken, L. S., 2013. Applied multiple regression/correlation analysis for the behavioral sciences. Routledge Academic. *Cité page 128*
- Corbett, T. H., Griswold, D. P., Roberts, B. J., Peckham, J. C., Schabel, F. M., 1975. Tumor induction relationships in development of transplantable cancers of the colon in mice for chemotherapy assays, with a note on carcinogen structure. *Cancer Research* 35 (9), 2434–2439. *Cité page 20*
- Cornish, E. A., Fisher, R. A., 1938. Moments and cumulants in the specification of distributions. *Revue de l'Institut International de Statistique*, 307–320. *Cité page 67*
- Cosatto, E., Miller, M., Graf, H. P., Meyer, J. S., 2008. Grading nuclear pleomorphism on histological micrographs. In : Proc. ICPR. IEEE, pp. 1–4. *Cité page 32*
- Costa, A. F., Humpire-Mamani, G., Traina, A. J. M., 2012. An efficient algorithm for fractal analysis of textures. In : Proc. SIBGRAPI. IEEE, pp. 39–46. *Cité page 104*
- Cruz-Roa, A., Caicedo, J. C., González, F. A., 2011. Visual pattern mining in histology image collections using bag of features. *Artificial Intelligence in Medicine* 52 (2), 91–106. *Cité page 17*
- Cuesta-Albertos, J. A., Febrero-Bande, M., 2010. A simple multiway ANOVA for functional data. *TEST : An Official Journal of the Spanish Society of Statistics and Operations Research* 19 (3), 537–557. *Cité pages 131 et 137*
- Dalle, J., Li, H., Huang, C., Leow, W. K., Racoceanu, D., Putti, T. C., 2009. Nuclear pleomorphism scoring by selective cell nuclei detection. In : Proc. WACV. IEEE. *Cité page 32*
- Dasarathy, B. V., Holder, E. B., 1991. Image characterizations based on joint gray level-run length distributions. Elsevier, *Pattern Recognition Letters* 12 (8), 497–502. *Cité page 92*
- Daugman, J. G., 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A* 2 (7), 1160–1169. *Cité page 92*

- Demir, C., Gultekin, S. H., Yener, B., 2005. Learning the topological properties of brain tumors. *IEEE Transactions on Computational Biology and Bioinformatics* 2 (3), 262–270. *Cité page 52*
- Di Cataldo, S., Ficarra, E., Acquaviva, A., Macii, E., 2010. Automated segmentation of tissue images for computerized IHC analysis. *Elsevier, Computer methods and programs in biomedicine* 100 (1), 1–15. *Cité page 24*
- Di Cataldo, S., Ficarra, E., Macii, E., 2012. Computer-aided techniques for chromogenic immunohistochemistry : Status and directions. *Elsevier, Computers in Biology and Medicine* 42 (10), 1012–1025. *Cité page 9*
- Diggle, P. J., 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, 87–101. *Cité page 61*
- Diggle, P. J., 1983. *Statistical analysis of spatial point patterns*. Academic Press. *Cité pages 64 et 65*
- Diggle, P. J., Chetwynd, A. G., 1991. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 1155–1163. *Cité page 61*
- Diggle, P. J., Gómez-Rubio, V., Brown, P. E., Chetwynd, A. G., Gooding, S., 2007. Second-order analysis of inhomogeneous spatial point processes using case–control data. *Biometrics* 63 (2), 550–557. *Cité page 61*
- Diniz, C., 2010. Microscopic image analysis using computer-assisted methodology to quantify immunostained receptors. *Microscopy : Science, Technology, Applications and Education* 2 (4), 1516–1525. *Cité page 12*
- Doyle, S., Madabhushi, A., Feldman, M., Tomaszewski, J., 2006. A boosting cascade for automated detection of prostate cancer from digitized histology. In : *MICCAI*. Springer, pp. 504–511. *Cité page 30*
- du Prel, J. B., Röhrig, B., Hommel, G., Blettner, M., 2010. Choosing statistical tests : Part 12 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International* 107 (19), 343. *Cité page 130*
- Dua, S., Acharya, U., Chowriappa, P., Sree, S., 2012. Wavelet-based energy features for glaucomatous image classification. *IEEE Transactions on Information Technology in Biomedicine* 16 (1), 80–87. *Cité page 107*
- Duda, R. O., Hart, P. E., Stork, D. G., 2012. *Pattern classification*. John Wiley & Sons. *Cité page 133*

- Dzyubachyk, O., van Cappellen, W. A., Essers, J., Niessen, W. J., Meijering, E., 2010. Advanced level-set-based cell tracking in time-lapse fluorescence microscopy. *IEEE Transactions on Medical Imaging* 29 (3), 852–867. *Cité page 31*
- Elston, C. W., Ellis, I. O., 1991. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer : Experience from a large study with long-term follow-up. *Histopathology* 19 (5), 403–410. *Cité page 111*
- Feder, J., 1988. *Fractals*. Springer. *Cité page 94*
- Fleischer, F., Beil, M., Kazda, M., Schmidt, V., 2006. Analysis of spatial point patterns in microscopic and macroscopic biological image data. In : *Case Studies in Spatial Point Process Modeling*. Springer, pp. 235–260. *Cité page 53*
- Fuchs, T. J., Lange, T., Wild, P. J., Moch, H., Buhmann, J. M., 2008. Weakly supervised cell nuclei detection and segmentation on tissue microarrays of renal clear cell carcinoma. In : *Pattern Recognition*. Springer, pp. 173–182. *Cité page 30*
- Galloway, M. M., 1975. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* 4 (2), 172–179. *Cité page 91*
- Garnier, M., Alsheh Ali, M., Seguin, J., Mignet, N., Hurtut, T., Wendling, L., 2014. Grading cancer from liver histology images using inter and intra region spatial relations. In : *Image Analysis and Recognition*. Vol. 8815. Springer, pp. 247–254. *Cité page 47*
- Gavrilovic, M., Azar, J. C., Lindblad, J., Wahlby, C., Bengtsson, E., Busch, C., Carlbom, I. B., 2013. Blind color decomposition of histological images. *IEEE Transactions on Medical Imaging* 32 (6), 983–994. *Cité page 26*
- Gevers, T., Smeulders, A. W. M., 1999. Color-based object recognition. Elsevier, *Pattern Recognition* 32 (3), 453–464. *Cité page 33*
- González-Rufino, E., Carrión, P., Cernadas, E., Fernández-Delgado, M., Domínguez-Petit, R., 2013. Exhaustive comparison of colour texture features and classification methods to discriminate cells categories in histological images of fish ovary. Elsevier, *Pattern Recognition* 46 (9), 2391–2407. *Cité page 88*

- Goreaud, F., Pélissier, R., 2003. Avoiding misinterpretation of biotic interactions with the intertype K12-function : Population independence vs. random labelling hypotheses. *Journal of Vegetation Science* 14 (5), 681–692. *Cité page 60*
- Greaney, M. J., Hoffman, D. C., Garway-Heath, D. F., Nakla, M., Coleman, A. L., Caprioli, J., 2002. Comparison of optic nerve imaging methods to distinguish normal eyes from those with glaucoma. *Investigative Ophthalmology & Visual Science* 43 (1), 140–145. *Cité page 106*
- Greene, W. H., 2000. *Econometric analysis* 4th edition. Macmillan. *Cité page 126*
- Gunduz, C., Yener, B., Gultekin, S. H., 2004. The cell graphs of cancer. *Bioinformatics* 20 (suppl 1), 145–151. *Cité page 52*
- Guo, Z., Zhang, L., Zhang, D., 2010. A Completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing* 19, 1657–1663. *Cité pages 99 et 100*
- Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., Yener, B., 2009. Histopathological image analysis : a review. *IEEE Reviews in Biomedical Engineering* 2, 147–171. *Cité pages 26 et 52*
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3, 1157–1182. *Cité page 132*
- Haghighat, M., Zonouz, S., Abdel-Mottaleb, M., 2013. Identification using encrypted biometrics. In : *Computer Analysis of Images and Patterns*. Springer, pp. 440–448. *Cité page 103*
- Hamilton, P. W., Bartels, P. H., Thompson, D., Anderson, N. H., Montironi, R., Sloan, J. M., 1997. Automated location of dysplastic fields in colorectal histology using image texture analysis. *The Journal of pathology* 182 (1), 68–75. *Cité page 88*
- Haralick, R. M., Shanmugam, K., Dinstein, I., 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics SMC-3* (6), 610–621. *Cité page 90*
- Harwood, D., Ojala, T., Pietikäinen, M., Kelman, S., Davis, L., 1993. Texture classification by center-symmetric auto-correlation, using kullback discrimination of distributions. Tech. rep., Computer Vision Laboratory, Center for Automation Research (University of Maryland, College Park, Maryland). CAR-TR-678. *Cité page 96*

- Hayashi, Y., Nakagawa, T., Hatanaka, Y., Aoyama, A., Kakogawa, M., Hara, T., Fujita, H., Yamamoto, T., 2007. Detection of retinal nerve fiber layer defects in retinal fundus images using Gabor filtering. In : Medical Imaging. International Society for Optics and Photonics, pp. 65142Z–65142Z. *Cité page 107*
- He, L., Long, L. R., Antani, S., Thoma, G. R., 2010. Computer assisted diagnosis in histopathology. Sequence and Genome Analysis : Methods and Applications, 271–287. *Cité page 5*
- He, L., Long, L. R., Antani, S., Thoma, G. R., 2012. Histology image analysis for carcinoma detection and grading. Computer Methods and Programs in Biomedicine 107 (3), 538–556. *Cité page 26*
- Heldmuth, L. O., 2012. In vivo monitoring of elastic changes during cancer development and therapeutic treatment. Thèse de Doctorat, Université Paris Diderot. *Cité page 19*
- Hervé, N., Servais, A., Thervet, E., Olivo-Marin, J., Meas-Yedid, V., 2011. Statistical color texture descriptors for histological images analysis. In : Proc. ISBI : from nano to macro. IEEE, pp. 724–727. *Cité page 88*
- Ho, K. S., Poon, P. C., Owen, S. C., Shoichet, M. S., 2012. Blood vessel hyperpermeability and pathophysiology in human tumour xenograft models of breast cancer : A comparison of ectopic and orthotopic tumours. BMC Cancer 12 (1), 1–10. *Cité page 19*
- Holmes, T. J., Cheng, P. C., 2006. Basic principles of imaging. Multi-Modality Microscopy. Singapore : World Scientific Publishing Co., Pte., Ltd, 1–74. *Cité page 5*
- Illian, J., Penttinen, P. A., Stoyan, H., Stoyan, P. D., 2008. Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice. John Wiley & Sons. *Cité pages 58, 72 et 85*
- Irshad, H., Veillard, A., Roux, L., Racoceanu, D., 2014. Methods for nuclei detection, segmentation, and Classification in digital histopathology : A review. Current status and future potential. IEEE Reviews in Biomedical Engineering 7, 97–114. *Cité page 31*
- Jaccard, J., Becker, M. A., Wood, G., 1984. Pairwise multiple comparison procedures : A review. Psychological Bulletin 96 (3), 589. *Cité page 130*
- Jafari-Khouzani, K., Soltanian-Zadeh, H., 2003. Multiwavelet grading of pathological images of prostate. IEEE Transactions on Biomedical Engineering 50 (6), 697–704. *Cité page 116*

- Jain, R. K., Mehta, R., Dimitrov, R., Larsson, L. G., Musto, P. M., Hodges, K. B., Ulbright, T. M., Hattab, E. M., Agaram, N., Idrees, M. T., et al., 2011. Atypical ductal hyperplasia : Interobserver and intraobserver variability. *Modern Pathology* 24 (7), 917–923. *Cité page 11*
- Jiao, Y., Berman, H., Kiehl, T. R., Torquato, S., 2011. Spatial organization and correlations of cell nuclei in brain tumors. *PLoS ONE* 6 (11), e27323. *Cité pages 30 et 53*
- Joshi, G., Sivaswamy, J., Krishnadas, S. R., 2011. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE Transactions on Medical Imaging* 30 (6), 1192–1205. *Cité page 107*
- Julesz, B., 1981. Textons, the elements of texture perception, and their interactions. *Nature* 290 (5802), 91–97. *Cité page 97*
- Jung, C., Kim, C., 2010. Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization. *IEEE Transactions on Biomedical Engineering* 57 (10), 2600–2604. *Cité page 32*
- Karperien, A., Jelinek, H. F., Milosevic, N. T., 2013. Reviewing lacunarity analysis and classification of microglia in neuroscience. *Fractals and complexity*. Wroclaw, Poland : Firma Format, 888–893. *Cité page 94*
- Kårsnäs, A., 2014. Image Analysis Methods and Tools for Digital Histopathology Applications Relevant to Breast Cancer Diagnosis. Thèse de Doctorat, Uppsala University, Uppsala, Sweden. *Cité page 9*
- Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes : Active contour models. *International Journal of Computer Vision* 1 (4), 321–331. *Cité page 31*
- Kawasaki, R., Wang, J. J., Rochtchina, E., Lee, A. J., Wong, T. Y., Mitchell, P., 2013. Retinal vessel caliber is associated with the 10-year incidence of glaucoma : The Blue Mountains Eye Study. *Ophthalmology* 120 (1), 84–90. *Cité page 107*
- Kayser, K., Görtler, J., Metze, K., Goldmann, T., Vollmer, E., Mireskandari, M. and Kosjerina, Z., Kayser, G., 2008. How to measure image quality in tissue-based diagnosis (diagnostic surgical pathology). *Diagnostic Pathology* 3 (1), 1–7. *Cité page 15*
- Keenan, S. J., Diamond, J., McCluggage, G. W., Bharucha, H., Thompson, D., Bartels, P. H., Hamilton, P. W., 2000. An automated machine vision

- system for the histological grading of cervical intraepithelial neoplasia (CIN). *The Journal of Pathology* 192 (3), 351–362. *Cité page 123*
- Kent, C., 2008. Automated glaucoma diagnosis : Any time soon ? Review of *Ophthalmology* 15 (6). *Cité page 105*
- Khan, A., Rajpoot, N., Treanor, D., Magee, D., 2014. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering* 61 (6), 1729–1738. *Cité page 15*
- King, T. W., Brey, E. M., Youssef, A. A., Johnston, C., Patrick Jr, C. W., 2002. Quantification of vascular density using a semiautomated technique for immunostained specimens. *Analytical and Quantitative Cytology and Histology* 24 (1), 39–48. *Cité page 28*
- Kohler, T., Budai, A., Kraus, M., Odstrcilik, J., Michelson, G., Hornegger, J., 2013. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In : *Proc. CBMS. IEEE*, pp. 95–100. *Cité page 108*
- Korzynska, A., Roszkowiak, L., Lopez, C., Bosch, R., Witkowski, L., Lejeune, M., et al., 2013. Validation of various adaptive threshold methods of segmentation applied to follicular lymphoma digital images stained with 3, 3'-Diaminobenzidine & Haematoxylin. *Diagnostic Pathology* 8 (1), 48. *Cité page 30*
- Kothari, S., Phan, J. H., Stokes, T. H., Wang, M. D., 2013a. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association* 20 (6), 1099–1108. *Cité page 17*
- Kothari, S., Phan, J. H., Young, A. N., Wang, M. D., 2013b. Histological image classification using biologically interpretable shape-based features. *BMC Medical Imaging* 13 (1), 9. *Cité pages 16, 17, 47 et 123*
- Krishnan, M. M. R., Pal, M., Paul, R. R., Chakraborty, C., Chatterjee, J., Ray, A. K., 2012. Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis. *Journal of Medical Systems* 36 (3), 1745–1756. *Cité page 17*
- Lagache, T., Lang, G., Sauvonnet, N., Olivo-Marin, J., 2013a. Analysis of the spatial organization of molecules with robust statistics. *PLoS ONE* 8 (12), e80914. *Cité pages 66, 67 et 68*

- Lagache, T., Meas-Yedid, V., Olivo-Marin, J., 2013b. A statistical analysis of spatial colocalization using Ripley's K function. In : Proc. ISBI. IEEE, pp. 896–901. *Cité page 53*
- Landini, G., 1996. Applications of fractal geometry in pathology. Fractal Geometry in Biological Systems : An Analytical Approach, 205–246. *Cité page 94*
- Lang, G., Marcon, E., 2013. Testing randomness of spatial point patterns with the Ripley statistic. ESAIM : Probability and Statistics 17, 767–788. *Cité page 67*
- Lin, Y., 2000. Tensor product space ANOVA models. Annals of Statistics, 734–755. *Cité page 131*
- Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J., 2012. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic Pathology 7, 22. *Cité pages 16 et 88*
- Lotwick, H. W., Silverman, B. W., 1982. Methods for analysing spatial processes of several types of points. Journal of the Royal Statistical Society. Series B (Methodological), 406–413. *Cité page 59*
- Loukas, C., Kostopoulos, S., Tanoglidi, A., Glotsos, D., Sfikas, C., Cavouras, D., 2013. Breast cancer characterization based on image classification of tissue sections visualized under low magnification. Computational and mathematical methods in medicine 2013. *Cité page 123*
- Loukas, C. G., Linney, A., 2004. A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy : Proliferation, hypoxia and vasculature. Computer Methods and Programs in Biomedicine 74 (3), 183–199. *Cité page 30*
- Lu, N., 1997. Fractal imaging. Morgan Kaufmann Publishers Inc. *Cité page 94*
- Luck, B. L., Carlson, K. D., Bovik, A. C., Richards-Kortum, R. R., 2005. An image model and segmentation algorithm for reflectance confocal images of in vivo cervical tissue. IEEE Transactions on Image Processing 14 (9), 1265–1276. *Cité page 31*
- Lyon, H., De Leenheer, A., Horobin, R., Lambert, W., Schulte, E., Van Liedekerke, B., Wittekind, D., 1994. Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes,

- stains and chromogenic reagents. *The Histochemical Journal* 26 (7), 533–544. *Cité page 24*
- Ma, W., Lozanoff, S., 1999. A full color system for quantitative assessment of histochemical and immunohistochemical staining patterns. *Biotechnic & Histochemistry* 74 (1), 1–9. *Cité page 28*
- Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., Thomas, N. E., 2009. A method for normalizing histology slides for quantitative analysis. In : *Proc. ISBI. IEEE*, pp. 1107–1110. *Cité page 27*
- Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M., 2000. Robust texture classification by subsets of local binary patterns. In : *Proc. ICPR. Vol. 3. IEEE*, pp. 935–938. *Cité page 99*
- Mäenpää, T., Pietikäinen, M., 2003. Multi-scale binary patterns for texture analysis. In : *Image Analysis. Springer*, pp. 885–892. *Cité pages 101 et 102*
- Mäenpää, T., Pietikäinen, M., 2004. Classification with color and texture : Jointly or separately ? *Elsevier, Pattern Recognition* 37 (8), 1629–1640. *Cité page 88*
- Mahmoud-Ghoneim, D., 2011. Optimizing automated characterization of liver fibrosis histological images by investigating color spaces at different resolutions. *Theoretical Biology and Medical Modelling* 8 (1), 25. *Cité pages 16 et 89*
- Mallat, S. G., 1989. A theory for multiresolution signal decomposition : The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 674–693. *Cité page 94*
- Malpica, N., Ortiz de Solorzano, C., Vaquero, J. J., Santos, A., Vallcorba, I., Garcia-Sagredo, J. M., Pozo, F. d., 1997. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry Part A* 28 (4), 289–297. *Cité page 31*
- Mandelbrot, B. B., 1977. *Fractals. Freeman.* *Cité page 94*
- Marcon, E., Puech, F., 2010. Measures of the geographic concentration of industries : Improving distance-based methods. *Journal of Economic Geography* 10 (5), 745–762. *Cité page 61*
- Marques, S., Joaquim, P., 2007. *Applied Statistics Using SPSS, Statistica, Matlab and R. Springer.* *Cité page 128*

- Mattfeldt, T., Eckel, S., Fleischer, F., Schmidt, V., 2006. Statistical analysis of reduced pair correlation functions of capillaries in the prostate gland. *Journal of Microscopy* 223 (2), 107–119. *Cité page 18*
- Mattfeldt, T., Eckel, S., Fleischer, F., Schmidt, V., 2007. Statistical modelling of the geometry of planar sections of prostatic capillaries on the basis of stationary Strauss hard-core processes. *Journal of Microscopy* 228 (3), 272–281. *Cité page 85*
- Mattfeldt, T., Eckel, S., Fleischer, F., Schmidt, V., 2009. Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *Journal of Microscopy* 235 (1), 106–118. *Cité page 53*
- Miedema, J., Marron, J. S., Niethammer, M., Borland, D., Woosley, J., Coposky, J., Wei, S., Reisner, H., Thomas, N. E., 2012. Image and statistical analysis of melanocytic histology. *Histopathology* 61 (3), 436–444. *Cité page 122*
- Monasse, P., Guichard, F., May 2000. Fast computation of a contrast-invariant image representation. *IEEE Transactions on Image Processing* 9 (5), 860–872. *Cité page 40*
- Nailon, W. H., 2010. Texture analysis methods for medical image characterisation. *Biomedical Imaging InTech*, 75–100. *Cité page 90*
- Nayak, J., Acharya, R., Bhat, P. S., Shetty, N., Lim, T., 2009. Automated diagnosis of glaucoma using digital fundus images. *Journal of Medical Systems* 33 (5), 337–346. *Cité page 107*
- Newberg, J., Murphy, R. F., 2008. A framework for the automated analysis of subcellular patterns in human protein atlas images. *Journal of Proteome Research* 7 (6), 2300–2308. *Cité page 27*
- Neyman, J., Scott, E. L., 1952. A theory of the spatial distribution of galaxies. *The Astrophysical Journal* 116, 144. *Cité page 85*
- Odstrcilik, J., Kolar, R., Jan, J., Gazarek, J., Kuna, Z., Vodakova, M., 2012. Analysis of retinal nerve fiber layer via Markov random fields in color fundus images. In : *Proc. IWSSIP. IEEE*, pp. 504–507. *Cité page 107*
- Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. *Elsevier, Pattern Recognition* 29 (1), 51–59. *Cité page 96*

- Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (7), 971–987. *Cité pages 97, 99 et 101*
- Oliveira, D. L. L., Nascimento, M. Z., Neves, L. A., Batista, V. R., Godoy, M. F., Jacomini, R. S., Duarte, Y. A. S., Arruda, P. F. F., Neto, D. S., 2014. Automatic classification of prostate stromal tissue in histological images using Haralick descriptors and Local Binary Patterns. In : *Journal of Physics : Conference Series*. Vol. 490. IOP Publishing, p. 012151. *Cité page 123*
- Olowoyeye, A., Tuceryan, M., Fang, S., 2009. Medical volume segmentation using bank of Gabor filters. In : *Proceedings of the 2009 ACM Symposium on Applied Computing*. ACM, pp. 826–829. *Cité page 92*
- Oztan, B., Kong, H., Gürçan, M. N., Yener, B., 2012. Follicular lymphoma grading using cell-graphs and multi-scale feature analysis. In : *Medical Imaging*. International Society for Optics and Photonics, p. 831516. *Cité pages 18, 52 et 123*
- Packeisen, J., Korsching, E., Herbst, H., Boecker, W., Buerger, H., 2003. Demystified... tissue microarray technology. *Molecular Pathology* 56 (4), 198. *Cité page 11*
- Park, H. M., 2010. Hypothesis testing and statistical power of a test. The Trustees of Indiana University. *Cité page 127*
- Peng, Y., Jiang, Y., Eisengart, L., Healy, M. A., Straus, F. H., Yang, X. J., 2011. Segmentation of prostatic glands in histology images. In : *Proc. ISBI : From Nano to Macro*. IEEE, pp. 2091–2094. *Cité page 31*
- Pinidiyaarachchi, A., Wählby, C., 2005. Seeded watersheds for combined segmentation and tracking of cells. In : *Image Analysis and Processing*. Springer, pp. 336–343. *Cité page 31*
- Qi, X., Xing, F., Foran, D. J., Yang, L., 2012. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. *IEEE Transactions on Biomedical Engineering* 59 (3), 754–765. *Cité page 32*
- Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R., Gurcan, M., 2008. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In : *Proc. MICCAI*. Springer, pp. 196–204. *Cité pages 17 et 88*

- Rabinovich, A., Laris, C. A., Inc, Q., Agarwal, S., Price, J. H., Belongie, S., 2003. Unsupervised color decomposition of histologically stained tissue samples. In : Advances in Neural Information Processing Systems. pp. 667–674. *Cité page 27*
- Ramos-Vara, J. A., 2005. Technical aspects of immunohistochemistry. Veterinary Pathology Online 42 (4), 405–426. *Cité pages 8 et 9*
- Ramsay, J. O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer. *Cité page 141*
- Raudenbush, S. W., Bryk, A. S., 2002. Hierarchical linear models : Applications and data analysis methods. Sage. *Cité page 129*
- Ripley, B. D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability, 255–266. *Cité page 59*
- Ripley, B. D., 1977. Modelling spatial patterns. Journal of the Royal Statistical Society. Series B (Methodological), 172–212. *Cité pages 63 et 64*
- Ripley, B. D., 1991. Statistical inference for spatial processes. Cambridge University Press. *Cité page 67*
- Rizzardi, A. E., Johnson, A. T., Vogel, R. I., Pambuccian, S. E., Henriksen, J., Skubitz, A. P., Metzger, G. J., Schmechel, S. C., 2012. Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. Diagnostic Pathology 7, 42. *Cité page 11*
- Rolls, G., 2011. An Introduction to Specimen Preparation. Leica Biosystems, Wetzlar, Germany.
URL <http://www.leicabiosystems.com/pathologyleaders/an-introduction-to-specimen-preparation/> *Cité page 5*
- Ruifrok, A. C., 1997. Quantification of immunohistochemical staining by color translation and automated thresholding. Analytical and quantitative cytology and histology 19 (2), 107–113. *Cité page 28*
- Ruifrok, A. C., Johnston, D. A., 2001. Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23 (4), 291–299. *Cité pages 26, 27 et 34*
- Ruifrok, A. C., Katz, R. L., Johnston, D. A., 2003. Comparison of quantification of histochemical staining by hue-saturation-intensity (HSI) transformation and color-deconvolution. Applied Immunohistochemistry & Molecular Morphology 11 (1), 85–91. *Cité page 30*

- Ruttenberg, B. E., Luna, G., Lewis, G. P., Fisher, S. K., Singh, A. K., 2013. Quantifying spatial relationships from whole retinal images. *Bioinformatics* 29 (7), 940–946. *Cité page 68*
- Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23 (19), 2507–2517. *Cité page 132*
- Salama, M. E., Lange, H., Tripp, S. R., Kohan, J., Landis, N. D., Krueger, J. S., Potts, S. J., 2013. AngioMap is a novel image analysis algorithm for assessment of plasma cell distribution within bone marrow vascular niche. *Applied Immunohistochemistry & Molecular Morphology*. *Cité page 68*
- Sertel, O., Kong, J., Catalyurek, U. V., Lozanski, G., Saltz, J. H., Gurcan, M. N., 2009. Histopathological image analysis using model-based intermediate representations and color texture : Follicular lymphoma grading. *Journal of Signal Processing Systems* 55 (1-3), 169–183. *Cité pages 16, 88 et 123*
- Setiadi, A. F., Ray, N. C., Kohrt, H. E., Kapelner, A., Carcamo-Cavazos, V., Levic, E. B., Yadegarynia, S., van der Loos, C. M., Schwartz, E. J., Holmes, S., Lee P, P., 2010. Quantitative, architectural analysis of immune cell subsets in tumor-draining lymph nodes from breast cancer patients and healthy lymph nodes. *PLoS ONE* 5 (8). *Cité pages 30 et 53*
- Sezgin, M., Sankur, B., 2004. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging* 13 (1), 146–168. *Cité page 30*
- Signolle, N., Revenu, M., Plancoulaine, B., Herlin, P., 2010. Wavelet-based multiscale texture segmentation : Application to stromal compartment characterization on virtual slides. *Signal Processing* 90 (8), 2412–2422. *Cité page 17*
- Smith Jr., T. G., Lange, G. D., Marks, W. B., 1996. Fractal methods and results in cellular morphology-dimensions, lacunarity and multifractals. *Journal of Neuroscience Methods* 69 (2), 123–136. *Cité page 94*
- Stenkvist, B., Bengtsson, E., Eriksson, O., Jarkrans, T., Nordin, B., Westman-Naeser, S., 1983. Histopathological systems of breast cancer classification : Reproducibility and clinical significance. *Journal of Clinical Pathology* 36 (4), 392–398. *Cité page 11*
- Strauss, D. J., 1975. A model for clustering. *Biometrika* 62 (2), 467–475. *Cité page 85*

- Swamidoss, I. N., Kårsnäs, A., Uhlmann, V., Ponnusamy, P., Kampf, C., Simonsson, M., Wählby, C., Strand, R., 2013. Automated classification of immunostaining patterns in breast tissue from the human protein Atlas. *Journal of Pathology Informatics* 4 (Suppl). *Cité page 88*
- Tabesh, A., Teverovskiy, M., Pang, H., Kumar, V. P., Verbel, D., Kotsianti, A., Saidi, O., 2007. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Transactions on Medical Imaging* 26 (10), 1366–1378. *Cité pages 18, 88 et 116*
- Tadrous, P. J., 2010. Digital stain separation for histological images. *Journal of Microscopy* 240 (2), 164–172. *Cité pages 7, 25 et 26*
- Tavakol, M., Wilcox, R. R., 2013. Medical education research : The application of robust statistical methods. *International Journal of Medical Education* 4, 93–95. *Cité page 129*
- Taverna, G., Colombo, P., Grizzi, F., Franceschini, B., Ceva-Grimaldi, G., Seveso, M., Giusti, G., Piccinelli, A., Graziotti, P., 2009. Fractal analysis of two-dimensional vascularity in primary prostate cancer and surrounding non-tumoral parenchyma. *Pathology-Research and Practice* 205 (7), 438–444. *Cité page 122*
- Tosun, A. B., Gunduz-Demir, C., 2011. Graph run-length matrices for histopathological image segmentation. *IEEE Transactions on Medical Imaging* 30 (3), 721–732. *Cité page 88*
- Tuczek, H. V., Fritz, P., Schwarzmann, P., Wu, X., Mähner, G., 1996. Breast carcinoma. Correlations between visual diagnostic criteria for histologic grading and features of image analysis. *Analytical and Quantitative Cytology and Histology* 18 (6), 481–493. *Cité page 122*
- Unser, M., 1986. Sum and difference histograms for texture classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (1), 118–125. *Cité page 91*
- van Lieshout, M. N. M., Baddeley, A. J., 1996. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica* 50 (3), 344–361. *Cité page 57*
- Veta, M., van Diest, P. J., Kornegoor, R., Huisman, A., Viergever, M. A., Pluim, J. P., 2013. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. *PLoS ONE* 8 (7), e70221. *Cité page 32*

- Waller, L. A., Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopoulou, I. G., Kennedy, W. R., Wendelschafer-Crabb, G., 2011. Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine*, 2827–2841. *Cité page 53*
- Wang, D., Shi, L., Wang, Y. J., Man, G. C., Heng, P. A., Griffith, J. F., Ahuja, A. T., 2011. Color quantification for evaluation of stained tissues. *Cytometry Part A* 79 (4), 311–316. *Cité page 15*
- Weidner, N., 1999. Tumour vascularity and proliferation : Clear evidence of a close relationship. *The Journal of Pathology* 189 (3), 297–299. *Cité page 77*
- Weyn, B., Tjalma, W. A. A., Vermeulen, P., Van Daele, A., Van Marck, E., Jacob, W., 2004. Determination of tumour prognosis based on angiogenesis-related vascular patterns measured by fractal and syntactic structure analysis. *Clinical Oncology* 16 (4), 307–316. *Cité page 53*
- Wilcox, R. R., 2012. Introduction to robust estimation and hypothesis testing. Academic Press. *Cité page 130*
- Wilcox, R. R., 2014. Modern robust statistical methods can provide substantially higher power and a deeper understanding of data. *Annals of Phytomedicine* 3 (1), 25–30. *Cité page 129*
- Wild, R., Ramakrishnan, S., Sedgewick, J., Griffioen, A. W., 2000. Quantitative assessment of angiogenesis and tumor vessel architecture by computer-assisted digital image analysis : Effects of VEGF–toxin conjugate on tumor microvessel density. *Microvascular Research* 59 (3), 368–376. *Cité page 122*
- Winsor, L., 1994. Tissue processing. *Laboratory Histopathology*. New York : Churchill Livingstone, 4.2–1–4.2–42. *Cité pages 6 et 7*
- Xia, G. S., Delon, J., Gousseau, Y., 2010. Shape-based Invariant Texture Indexing. *International Journal of Computer Vision* 88 (3), 382–403. *Cité page 40*
- Yamada, I., Rogerson, P., 2003. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35 (2), 97–109. *Cité page 62*
- Yang, X., Li, H., Zhou, X., 2006. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems I : Regular Papers* 53 (11), 2405–2414. *Cité page 31*

- Yu, J., Abidi, S. S. R., Artes, P., McIntyre, A., Heywood, M., 2005. Diagnostic support for glaucoma using retinal images : A hybrid image analysis and data mining approach. *Studies in Health Technology and Informatics* 116, 187. *Cité page 105*
- Zeng, Z., Strange, H., Han, C., Zwigelaar, R., 2013. Unsupervised cell nuclei segmentation based on morphology and adaptive active contour modelling. In : *Image Analysis and Recognition*. Vol. 7950. Springer Berlin Heidelberg, pp. 605–612. *Cité page 31*
- Zhang, J. T., 2013. *Analysis of variance for functional data*. CRC Press. *Cité page 131*
- Zhang, Z., Yin, F. S., Liu, J., Wong, W. K., Tan, N. M., Lee, B. H., Cheng, J., Wong, T. Y., 2010. ORIGA-light : An online retinal fundus image database for glaucoma analysis and research. In : *Proc. EMBC. IEEE*, pp. 3065–3068. *Cité page 108*