AIX-MARSEILLE UNIVERSITÉ

FACULTÉ DES SCIENCES DE LUMINY
163, Avenue de Luminy
13288 MARSEILLE cedex 9

LABORATOIRE LSIS
Avenue Escadrille Normandie-Niemen
13397 MARSEILLE cedex 20

THÈSE DE DOCTORAT ÈS SCIENCES

DISCIPLINE INFORMATIQUE
SPÉCIALITÉ IMAGERIE NUMÉRIQUE

---

présentée par Guillaume THIBAULT

en vue d'obtenir le grade de docteur d'*Aix-Marseille Université*

Indices de formes et de textures : de la 2D vers la 3D
Application au classement de noyaux de cellules

---

soutenue le jeudi 18 juin 2009 devant le jury composé de

Christine GRAFFIGNE	Professeur	Paris-Descartes	Présidente
Chafiaa HAMITOUCHE	Professeur	ENST de Bretagne	Rapporteuse
Stéphane LALLICH	Professeur	Lyon II	Rapporteur
Pierre CAU	Professeur	Aix-Marseille	Examinateur
Bernard FERTIL	Directeur de Recherche	LSIS	Examinateur
Jean SEQUEIRA	Professeur	Aix-Marseille	Directeur de thèse
Jean-Luc MARI	Maître de Conférences	Aix-Marseille	Directeur de thèse

Thèse préparée au sein du Laboratoire LSIS – Equipe I&M

A ma DouDou,
A mon frère (joyeux anniversaire) et mes sœurs,
A tout ceux qui ont rendu ce travail possible.

REMERCIEMENTS

A tout ceux qui ont rendu ce travail possible, ils ont ma reconnaissance et se reconnaîtront...

Avec toute ma sincérité.

Merci

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

TABLE DES ALGORITHMES

INTRODUCTION

La reconnaissance de forme (*RDF*) est une partie majeure de l'intelligence artificielle qui vise à automatiser le discernement de situations typiques au niveau de la perception. Née dans les années cinquante avec les premiers systèmes d'acquisitions optiques dans l'industrie et l'armement, la reconnaissance de forme est devenue un enjeu capital pour de très nombreuses applications liées à la vision par ordinateur : la reconnaissance des caractères manuscrits (numérisation des livres, lecture automatique des lettres postales et des chèques bancaires, etc.), la sécurité avec la vidéo surveillance (détection d'incendies) et la reconnaissance faciale, l'imagerie médicale (échographie, scanner, imagerie par résonance magnétique, classification de chromosomes, etc.), la télédétection, etc. et dans ce qui nous intéresse tout au long de ce manuscrit : l'aide au diagnostic.

L'homme est le système de reconnaissance le plus parfait. La diversité des tâches de reconnaissance que nous pouvons effectuer sur des formes très variées est et restera considérable. Il nous est par exemple possible de distinguer sans effort deux types de plantes différentes, retrouver la clef souhaitée au milieu d'un trousseau, distinguer un visage d'homme de celui d'une femme, etc. On souhaiterait que les machines[1] puissent faire le plus grand nombre possible de tâches aussi bien qu'un homme, voire mieux dans certains cas ou sinon avec un taux d'erreur acceptable si cela est compensé par le gain obtenu (temps, fiabilité, etc.). Par exemple, dans le domaine des *fouilles de données* (*data mining*), une machine peut traiter un très grand nombre d'individus avec fiabilité alors qu'il serait impossible à un homme d'effectuer la même tâche dans un temps raisonnable. Les progrès scientifiques et techniques permettent d'imiter certaines de ces capacités à l'aide d'ordinateurs, mais il n'existe pas encore de méthode apportant le même résultat qu'un être humain et surtout pas dans plusieurs domaines.
Ainsi, le problème que cherche à résoudre la reconnaissance de forme est d'associer une étiquette à une donnée qui peut se présenter sous forme d'une image ou d'un signal. Des données différentes peuvent recevoir la même étiquette et sont alors des exemplaires de la classe identifiée par l'étiquette. La reconnaissance de forme a donc pour objectif de concevoir et d'analyser des algorithmes automatiques ou semi-automatiques permettant aux machines d'observer leur environnement et de réaliser une détection, une classification ou un classement[2] des formes présentes. Par exemple, la très célèbre et incontournable transformée de Hough [Hough 1962; Duda and Hart 1972] permet de détecter et de classer des droites, des courbes, des cercles ou des ellipses présentes dans une image. La reconnaissance de forme planes se fait au moyen d'un des trois principes de comparaison suivants :

- comparaison directe des formes (*shape matching*, [Siddiqi et al. 1999; Zhang and Lu 2004]).
- comparaison d'une représentation des formes. Il faut alors s'assurer qu'il y a une équivalence entre l'espace des formes et celui des représentations.
- comparaison de critères (caractéristiques) attribués aux formes.

[1] Les ordinateurs

[2] Dans le chapitre 2, les différences entre classement et classification sont expliquées

Au cours de ce manuscrit, nous nous intéressons au troisième principe afin de déterminer, créer et extraire des caractéristiques.
Mais il n'existe pas de méthode universelle de reconnaissance de forme. La plupart des méthodes sont définies en fonction du problème que l'on souhaite résoudre et connaissent des difficultés de réutilisation ou de généralisation. Cela peut s'expliquer par trois raisons :

- la première est la grande variété des disciplines dans lesquelles la reconnaissance de forme est employée.
- la deuxième est la diversité des applications, comme cela a pu être cité ci-dessus.
- la troisième concerne la variété des moyens techniques disponibles. Il existe toute une gamme de systèmes d'acquisition plus ou moins perfectionnés qui conditionnent la qualité de l'image. On verra un exemple de ce problème concernant des microscopes dans le chapitre 1.

C'est donc le domaine d'application qui dicte les algorithmes ou les stratégies à employer. Ainsi il apparaît que le choix d'une méthode est conditionné par le contexte d'application dans lequel on se place ; il est par conséquent primordial de définir avec précision ce contexte.

Dans le cadre de ce travail, nous nous intéressons au classement de noyaux de cellules afin d'aider une équipe d'experts composée de biologistes et de généticiens de l'hôpital *La Timone* à Marseille, qui travaille sur la maladie de la Progéria.
Le premier chapitre commence donc par présenter le contexte de ce travail et ainsi expliquer les motivations et les enjeux. Pour cela, une description de la maladie de la Progéria est présentée, ainsi que les différents points auxquels s'intéressent les experts dans leurs travaux. La présentation de ces points dans leur contexte, permet d'évaluer leur importance et ainsi définir les priorités dans notre travail. Ces priorités sont ensuite étudiées par ordre décroissant dans les chapitre suivants afin d'essayer d'apporter des solutions pour chacune d'elles et ainsi répondre au problème posé.

CHAPITRE 1

CONTEXTE ET MOTIVATIONS : LA PROGÉRIA

1.1 Contexte

La *Progéria*, également connue sous le nom de syndrome de *Hutchinson-Gilford*[1] est une maladie génétique orpheline. Un enfant sur huit millions naît avec cette maladie. Actuellement un peu plus d'une centaine de cas est recensée à travers le monde. C'est une maladie extrêmement grave qui entraîne plusieurs atteintes importantes et douloureuses chez ces enfants qui à ce jour ne sont pas soignés, malgré un tout premier essai clinique en cours. Cette maladie est une laminopathie, c'est-à-dire la conséquence d'un dysfonctionnement dans les protéines appelées *lamines*. Les lamines sont présentes dans la majeure partie des cellules différenciées de notre corps où elles jouent un rôle important que nous décrivons plus loin.
Au niveau clinique, la Progéria est caractérisée par une sénescence prématurée et accélérée (Progeria vient du grec *gerôn*, "vieillard", cf. figure 1.1). Les principaux symptômes visibles de la Progéria sont les suivants :

- Un retard de croissance caractérisé par un arrêt brutal de la courbe de croissance vers un an.
- Une alopécie[2].
- Une exophtalmie[3].
- Une très petite mâchoire par rapport à un encéphale disproportionné.
- Une acro-ostéolyse[4].
- Anomalies squelettiques.

Au delà de leur aspect physique très caractéristique, les enfants atteints par cette maladie ont également des pathologies habituellement retrouvées chez des personnes âgées : graves problèmes cardio-vasculaires causés par une athérosclérose[5] (le premier infarctus du myocarde survient vers cinq ans), diabète, ostéoporose, raideur articulaire, peau fine et glabre[6], etc. Leur espérance de vie est de $13,5$ ans (le triste record est de vingt-six ans) ; ils pèsent une trentaine de kilos et ne dépassent pas $1,15$ mètre.

[1] Noms des deux médecins qui ont décrit cette maladie il y a cent ans en Angleterre : www.progeriaresearch.org
[2] Une perte massive de cheveux.
[3] Désigne la saillie ou la propulsion du globe oculaire hors de l'orbite.
[4] Phalanges manquantes : http://www.vulgaris-medical.com/encyclopedie/acro-osteolyse-198.html.
[5] http://www.doctissimo.fr/html/sante/encyclopedie/sa_787_atherosclerose.htm
http://www.medecine-et-sante.com/maladiesexplications/atherosclerose.html
[6] Sans poil

Figure 1.1. Deux photos d'enfants atteints de la Progéria où l'on observe les symptômes.

En 2003, une avancée majeure de la recherche sur cette maladie [Sandre-Giovannoli et al. 2003; Eriksson et al. 2003] a été réalisée en parallèle par deux laboratoires internationaux. Une équipe marseillaise dirigée par le professeur Nicolas Levy (généticien) au *CHU de la Timone* ainsi qu'une équipe américaine du *National Institut of Health* (NIH). Leurs recherches ont mis en exergue la cause de cette maladie : une mutation dans le gène *LMNA* sur le chromosome 1. En 2003 ce gène était connu pour être responsable de plusieurs maladies spécifiques présentant une atteinte musculaire. La *lamine A* produite par le gène LMNA en conditions physiologiques est présente à la périphérie et à l'intérieur du noyau des cellules. Par assemblage avec d'autres lamines, cette protéine participe au maintien de la structure du noyau et de son enveloppe (figure 1.2). Le gène muté qui est responsable de la maladie, produit une protéine anormale qui ne remplit plus son rôle de maintien structural. Il a été observé de nombreuses anomalies sur des cellules en culture, notamment une dégénérescence plus rapide de la cellule et des anomalies structurales des noyaux. De plus ces cellules présentent également des problèmes de division (mitose). Le lien direct entre toutes ces anomalies observées en culture et les maladies que développent ces patients ne sont pas encore clairement compris. Mais un modèle est proposé : dans un organisme malade, ces cellules se renouvellent mal et entraînent une altération de nombreux tissus aboutissant à ce que l'on appelle le vieillissement prématuré.

La figure 1.2 montre deux noyaux de cellule dont l'un est sain et l'autre pathologique (présentant un défaut de lamine A, donc atteint par la maladie). En examinant les noyaux de cette figure, plusieurs questions se posent :

1. Comment parvient-on à visualiser les noyaux ? Quel matériel est utilisé et dans quelles circonstances ?
2. Comment faire pour différencier les noyaux ? Autrement dit : quels sont les critères visuels pour classer correctement les noyaux dans les catégories "sain" ou "pathologique" ?

Les sections suivantes répondent à ces deux questions, posent les problématiques et montrent l'intérêt de notre collaboration avec l'équipe du professeur Levy, ainsi que la participation du Pr. Cau (biologiste cellulaire au sein de l'équipe).

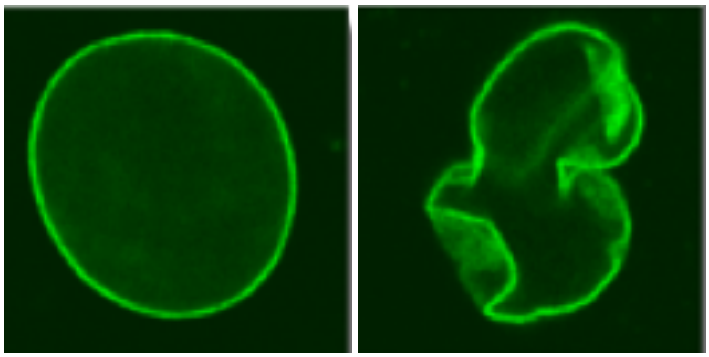


Figure 1.2. Deux exemples de noyaux de cellules : à gauche un noyau sain et à droite un noyau pathologique ayant un défaut de lamine A.

1.2 Evaluation d'un protocole de soins

Un des enjeux majeurs dans cette pathologie est l'utilisation de drogues en premier lieu sur des cellules cultivées *in vitro*[7] afin d'évaluer leur efficacité. Les experts observent si la drogue permet une amélioration des noyaux :
Pour évaluer l'efficacité d'un protocole, les experts procèdent en quatre étapes :

1. Prélèvement d'un échantillon par l'une des deux techniques suivantes :
   - Prise de sang et extraction des cellules *lymphoblastoïdes*.
   - *Biopsie*[8] de peau, afin d'extraire les *fibroblates*. Les noyaux utilisés dans la suite de ce documents sont issus de cette technique.
2. Mise en culture des cellules.
3. Traitement des cellules par une ou plusieurs molécules sur une partie des cultures tandis que l'autre partie n'est pas *traitées* afin de servir de témoin.
4. Acquisition des images des noyaux de cellule.
5. Calcul du pourcentage de noyaux pathologiques dans les deux parties des cultures (traitées et non traitées).

Pour calculer le pourcentage de noyaux pathologiques, les experts diagnostiquent l'état de chaque noyau de l'échantillon à l'aide des critères cités dans la section 1.5, puis calculent le pourcentage de noyaux pathologiques. Afin d'avoir un pourcentage qui décrive correctement l'échantillon, il faut classer plusieurs centaines de noyaux. Ce calcul de pourcentage nécessite parfois plusieurs jours de travail à une personne. Un patient sain possède environ 8% de noyaux pathologiques, mais ce pourcentage peut atteindre 80% chez des patients atteints car c'est une conséquence de la maladie.
Mais il est fréquent que deux experts soient en désaccord sur le diagnostic de certains noyaux. Ces désaccords semblent inévitables car de nombreux critères dépendent de l'appréciation de l'expert.

[7] *in vitro* signifie que cela se passe hors de l'organisme vivant, ici hors du patient.
[8] La *biopsie* est un petit prélèvement non invasif.

1.3 Acquisition des images de noyaux à l'aide du microscope à fluorescence

Afin d'observer les noyaux, il faut prélever des cellules chez le patient puis les mettre en culture. Après quelques jours de culture, les cellules sont déposées sur une lame de verre, marquées par des fluorochromes afin d'être observées au microscope à fluorescence et photographiées. Actuellement ce microscope est l'outil d'acquisition le plus utilisé par les experts (les biologistes et les généticiens) ; il permet d'acquérir rapidement des images de très bonne qualité qui permettent une analyse visuelle 2D d'un grand nombre de noyaux. En revanche, il ne fournit aucune information de profondeur (information 3D qui représente l'épaisseur du noyau).
On utilise des *anticorps* et des *fluorochromes* (des marqueurs fluorescents, cf. annexe C.1) afin de faire apparaître les éléments recherchés. Pour cela, les experts utilisent tout d'abords des anticorps qui réagissent à la présence soit de protéines spécifiques soit du matériel génétique (ADN), puis trois marqueurs qui s'observent sur des longueurs d'ondes de lumière qui leurs sont propres :

- DAPI (longueur d'onde bleue), qui est utilisé ici pour marquer l'ADN, avec un marquage plus intense pour les régions contenant de l'ADN inactif.
- FITC (longueur d'onde verte) est un anticorps qui est utilisé durant tout ce manuscrit pour marquer les protéines lamines A/C.
- TRITC (longueur d'onde rouge), un anticorps qui marque la lamine B.

La figure 1.3 montre différentes acquisitions avec le microscope à fluorescence en utilisant les marqueurs précédemment cités. Le microscope à fluorescence et les marqueurs sont présentés en détail dans l'annexe C.

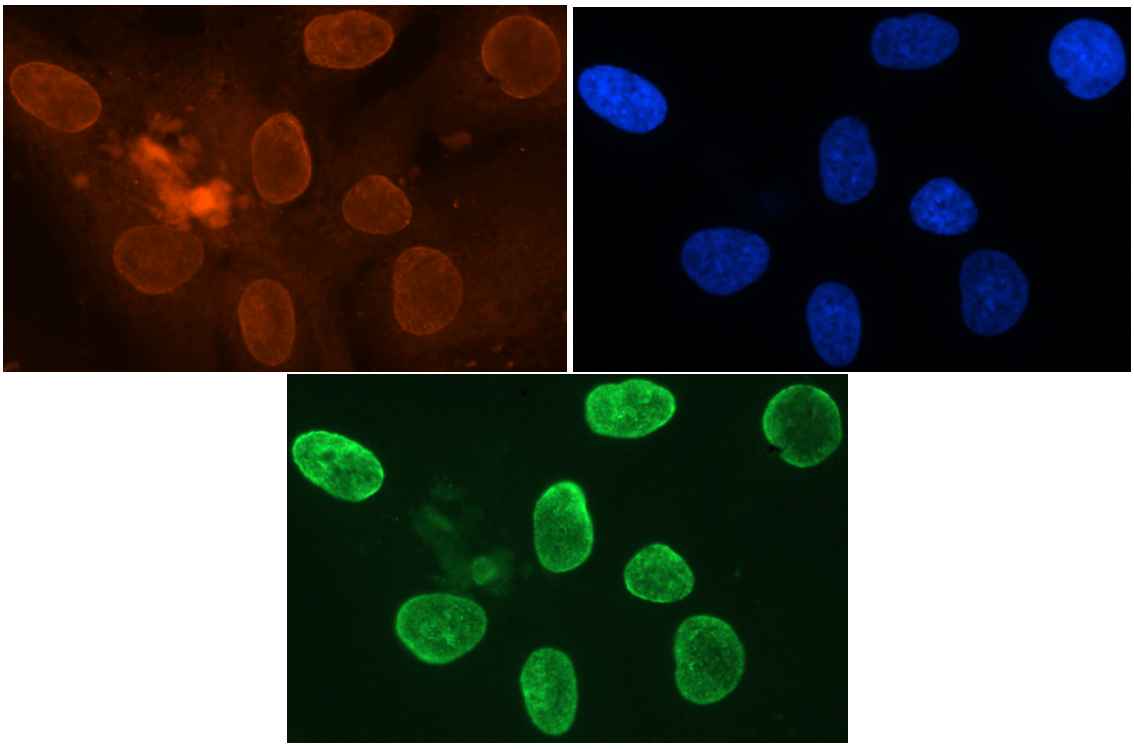


Figure 1.3. Exemples de résultats des différents marqueurs (FITC, TRITC, DAPI) avec un microscope à fluorescence classique.

Dans la section 1, il a été expliqué que la cause de la Progéria provenait d'un défaut de lamine A. Pour cela, le diagnostic s'effectue en observant les acquisitions avec le FITC (marqueur vert). Dans certains cas très particuliers et plutôt rares, les autres marqueurs peuvent apporter des informations complémentaires afin d'affiner le diagnostic préalablement établi. Mais nous ne disposons pas de diagnostic des noyaux à partir des marqueurs DAPI et TRITC (cf. section 1.5), qui ne sont donc pas utilisés dans ce manuscrit.

1.4 Acquisition 2.5D à l'aide du microscope confocal

Les experts utilisent parfois le microscope confocal pour acquérir des informations visuelles sur les noyaux. Ce microscope fournit des images de coupe des noyaux de cellules. Lorsque l'on "empile" ces coupes, on obtient une représentation du volume cellulaire. C'est ce que l'on nomme des informations 2.5D : des informations 2D qui renseignent sur la 3D. Mais ce microscope comporte plusieurs inconvénients :

- Il existe une grande différence de résolution entre les dimensions : la précision de l'acquisition en largeur et hauteur (axes X et Y) est beaucoup plus importante qu'en profondeur (axe Z). Cette différence engendre un volume anisotrope.
- L'observation de coupes pénalise la qualité de l'image : la précision du contour est très faible et la qualité de la texture incertaine.
- Le temps d'acquisition des données (plusieurs passages du faisceau laser par coupe) altère la durée de vie et la qualité des marqueurs (donc la qualité de l'image), ainsi que le nombre de saisies possibles (très peu d'échantillons). Cette altération provoque un écart d'intensité entre la première et la dernière coupe.
- Le temps d'acquisition d'un seul noyau de cellule est supérieur à celui nécessaire pour acquérir un échantillon complet à l'aide du microscope à fluorescence classique.

Outre les avantages et inconvénients techniques de ce microscope, un problème fondamental est issu de la nature même des cellules : la forme des cellules testées n'est pas "naturelle". En effet, ces cellules sont cultivées *in vitro* et par conséquent soumises à deux forces : la gravité et la viscosité. Ces forces empêchent le développement naturel de la cellule et altèrent sa forme générale. Lorsque la cellule se développe *in vivo*[9] sa forme est proche d'un ellipsoïde, mais dans notre cas (in vitro) elle est très aplatie. De plus, le haut et le bas des noyaux sont "coupés" lors de l'acquisition car ils n'offrent pas d'intérêt pour l'analyse d'un point de vue biologique. Toutes ces contraintes conduisent à l'obtention de noyaux de cellules qui ont une forme similaire à celle de la figure 1.4.

De plus les différences de résolution sur les trois axes et en particulier la faible résolution en Z, font perdre de l'information sur le contenu du noyau entre deux coupes successives. Donc bien que le résultat apparaisse en trois dimensions, il est la conséquence de l'empilement des différentes coupes d'acquisition 2D.
Cette acquisition 2.5D à l'aide du microscope confocal était originale et a motivé le début de notre travail. Nous souhaitions caractériser ces volumes afin d'apporter un nouveau type de caractérisation spécifique à la 3D. Mais les différents problèmes et inconvénients qui viennent d'être présentés rendent extrêmement difficile la caractérisation 3D dans ce type d'application.

[9] *in vivo* signifie que cela se passe dans l'organisme vivant, ici dans le patient. C'est le contraire de *in vitro*.

Figure 1.4. Exemple de résultat de l'acquisition au microscope confocal après segmentation.

1.5 Diagnostic des noyaux

Lorsque l'on diagnostique visuellement l'état d'un noyau de cellules à partir d'une acquisition au microscope à fluorescence, on souhaite le classer dans une des deux catégories : noyau sain ou noyau pathologique. Pour cela, les experts marquent le noyau à l'aide du FITC et utilisent les deux éléments de diagnostic fondamentaux qui sont les suivants :

1. La forme.
2. La texture.

Le rôle de ces deux éléments est expliqué de manière détaillée dans les sous-sections suivantes.

1.5.1 Classement en fonction de la forme

Le critère de forme est le principal élément de diagnostic des noyaux. Comme on a pu le voir sur le noyau pathologique de la figure 1.2, les noyaux atteints ont *généralement* une forme fortement non convexe[10]. La convexité est le principal critère lors de l'étude de la forme.

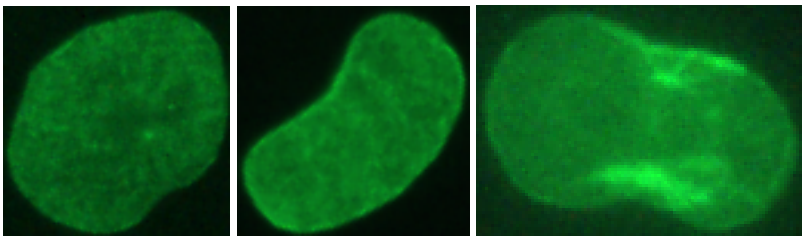


Figure 1.5. Trois exemples de noyaux normaux mais ayant une forme non convexe.

Toutefois, certains noyaux non convexes sont pourtant considérés comme sains (figure 1.5). Bien que la forme de ces trois noyaux soit non convexe, ils sont classés parmi les noyaux normaux.

[10]La définition de la convexité est donnée dans la section A.5 de ce manuscrit.

Donc, tous les noyaux avec une forme non convexe ne peuvent pas être considérés comme pathologiques. Il faut prendre en considération le degré de convexité d'un noyau : si celui-ci est inférieur à un certain seuil, alors il est considéré comme *fortement non convexe* et par conséquent pathologique. Actuellement, ce seuil n'est pas quantifié par les experts ; c'est-à-dire que chaque expert évalue le degré de convexité du noyau à partir de ses connaissances et de son expérience.
Dans la suite du document, les noyaux ayant une forme anormale seront appelés *boursouflés*.

1.5.2 Classement en fonction de la texture

La texture d'un noyau de cellule est la répartition du marqueur à l'intérieur du noyau. Dans notre cas, c'est la répartition des lamines A marquées par le FITC. Dans le cas idéal (un noyau sain pouvant être classé sans aucune ambiguïté), la répartition des lamines A doit être homogène à l'intérieur du noyau et plus dense sur la *périphérie* (en anglais *rim*, le noyau sain dans la figure 1.2). Lorsque les experts analysent la texture, ils sont attentifs à différents critères :

- L'homogénéité.
- Les focis.
- Les trous.
- La périphérie.

1.5.2.1 L'homogénéité

Un noyau sain a une texture homogène (figure 1.2). Une non homogénéité de la texture peut être de plusieurs types (figure 1.6). Mais comme pour l'évaluation du degré de convexité de la forme, aucun critère formalisé, ni aucune technique ne sont utilisés pour estimer l'homogénéité de la texture. Cette évaluation est propre à chaque expert.

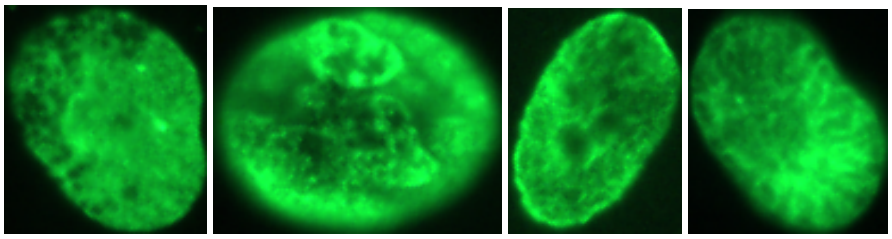


Figure 1.6. Quatre exemples de noyaux avec une texture non homogène.

De plus, il arrive que le marqueur ne se répartisse pas correctement dans l'échantillon, c'est-à-dire que certaines parties d'un noyau ne sont pas mises en contact avec le marqueur ou au contraire que le marqueur se concentre sur certaines zones. Il est également possible que le marqueur "précipite" ; il crée une pigmentation sur la texture. Ces défauts de marquage introduisent alors une erreur dans le diagnostic du noyau à partir de ce critère.

1.5.2.2 Les focis

Les *focis* sont des "tâches" ou des pics d'intensité dans la texture du noyau (figure 1.7). Les experts n'ont pas d'explication précise sur la présence de focis dans les noyaux mais ils savent que leur présence est anormale s'ils dépassent un certain nombre ou une certaine taille. Ils estiment que le noyau doit être considéré comme pathologique si le nombre de focis est supérieur à cinq ou si le noyau contient deux "gros" focis ou plus. Mais l'appréciation de la taille d'un focis est propre à chaque expert.

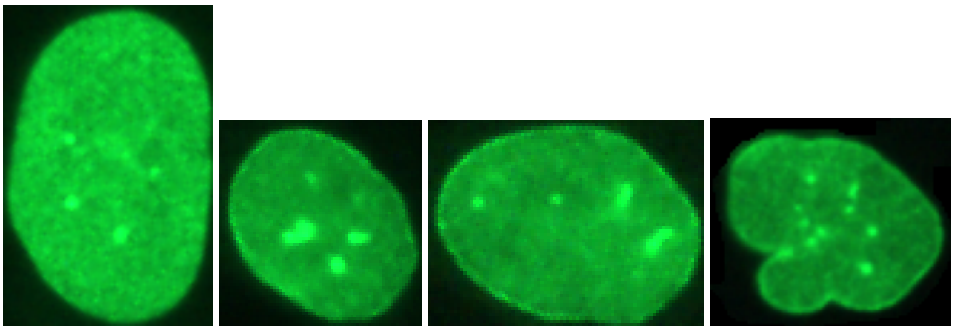


Figure 1.7. Quatre exemples de noyaux dont la texture comporte des focis.

1.5.2.3 Les trous

La répartition de la lamine A dans le noyau doit être homogène. Une absence de lamine A sur une partie de la texture forme un *trou* qui peut être interprété comme un défaut (figure 1.8). Dans le paragraphe précédent il a été dit que le marqueur peut parfois mal se répartir sur le noyau, mais cela ne se produit pas au point de créer des trous dans la texture pour des cellules saines. Le nombre et la taille des trous influent dans le diagnostic, mais comme précédemment, il n'y a pas de critère pour évaluer des seuils.

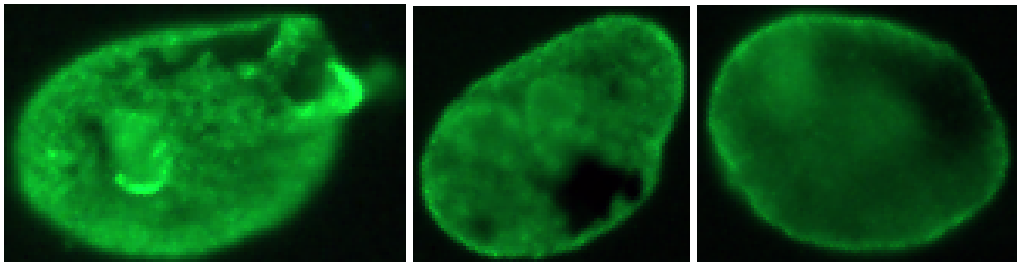


Figure 1.8. Trois exemples de noyaux avec au moins un trou dans la texture.

1.5.2.4 La périphérie

La périphérie d'un noyau n'est pas toujours marquée (figure 1.9).

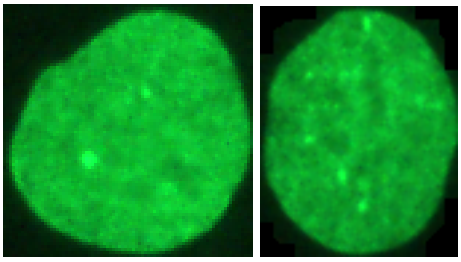


Figure 1.9. Deux exemples de noyaux sains avec une périphérie non marquée.

Si la périphérie d'un noyau n'est pas marquée, cela ne signifie pas obligatoirement que le noyau a une texture anormale et qu'il est pathologique. En revanche, si un noyau a une périphérie bien marquée, mais que sur une zone de la frontière il y a un défaut de marquage, alors la périphérie est anormale et le noyau pathologique (figure 1.10).

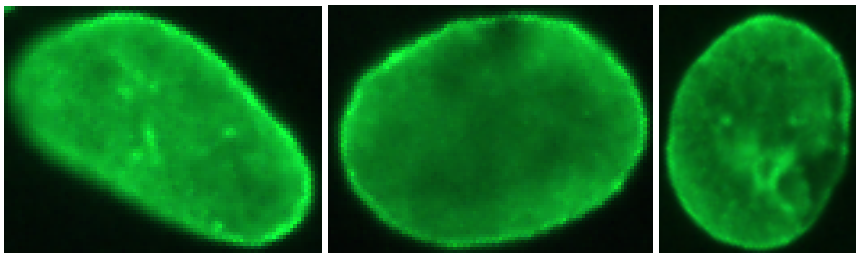


Figure 1.10. Trois exemples de noyaux pathologiques avec une périphérie marquée, mais pas sur la totalité.

1.6 Discussion et objectifs de notre contribution

Les différents outils d'acquisition et de marquage utilisés pour l'analyse des noyaux de cellules viennent d'être présentés, ainsi que leurs avantages et inconvénients. Il en résulte que pratiquer une expertise à l'aide de noyaux acquis au microscope confocal apporte de nombreux avantages mais pose également certains problèmes (en particulier le temps d'acquisition élevé qui engendre un nombre insuffisant de noyaux disponibles) pour lesquels un système de classement automatique ne peut être d'une grande utilité. Par conséquent les acquisitions utilisées dans le reste de ce document sont effectuées avec un microscope à fluorescence et le marqueur FITC.
Le protocole d'évaluation d'un traitement qui est entièrement manuel et répétitif, engendre une perte de temps considérable et pourrait être automatisé. C'est sur ce point qu'intervient notre contribution. Nous souhaitons construire une méthodologie qui utilise les éléments de diagnostic afin de donner une probabilité de classement pour chaque noyau. Cette probabilité doit permettre de classer correctement les noyaux dans les deux classes : noyaux pathologiques ou noyaux sains.
Mais lors de l'explication des différents éléments de diagnostic utilisés par les experts, il revient régulièrement que des critères de seuil et d'estimation peuvent varier en fonction de l'expert. Pour cette raison, nous souhaitons tout au long de ce document apporter des réponses fiables à ces absences de seuil et de critère en exploitant au mieux les différents éléments de diagnostic.

CHAPITRE 2

CLASSEMENT ET CLASSIFICATION

2.1 Introduction

Effectuer un diagnostic automatique des noyaux de cellules consiste à construire un algorithme qui définisse automatiquement si un noyau est sain ou pathologique. C'est ce que l'on appelle un classement, c'est-à-dire attribuer une classe d'appartenance aux individus étudiés (ici les noyaux). Tout au long de ce manuscrit, des modèles de classement et de classification sont construits afin de répondre aux différents problèmes rencontrés.
Ce chapitre présente, définit et différencie les notions de classement et de classification. Pour chacune de ces notions, les méthodes utilisées dans notre travail sont présentées en détaillant leurs utilisations, avantages et inconvénients.

2.2 Le classement

2.2.1 Définitions

Dans le problème présent, l'objectif du classement est de déterminer la classe d'appartenance des noyaux de cellules : *noyau sain* ou *noyau pathologique*.

Définition 2.2.1 (Classement) – *Le classement ("classification" en anglais) est l'opération qui vise à répartir des objets (individus ou variables) dans des classes prédéfinies par l'expert en fonction des caractéristiques de l'objet.*

Cette définition fait intervenir la notion de description d'un individu à l'aide de caractéristiques. Pour mettre en œuvre un classement, il faut au préalable construire un *vecteur caractéristique* de l'individu étudié. Ce vecteur à N composantes (attributs, caractéristiques, en anglais *features*) décrit l'individu. Par exemple, il peut contenir sa surface, son périmètre, etc. Donc un individu est perçu comme un point dans l'espace des caractéristiques (espace à N dimensions) et une classe est alors une partition ou une région de l'espace.
Toute la difficulté dans la construction d'un vecteur caractéristique est de le construire avec des descripteurs suffisamment discriminants afin que les individus d'une même classe soient proches dans l'espace des caractéristiques (cf. définition 2.4.3). De même, les individus appartenant à des classes différentes doivent occuper des régions disjointes et les plus éloignées possibles dans l'espace (cf.

définition 2.4.4)[1]. L'efficacité de discrimination des caractéristiques est déterminée par la capacité et la facilité de la méthode de classement à séparer des formes provenant de différentes classes.

Les méthodes de classement sont ce que l'on appelle des méthodes par apprentissages supervisés : elles font intervenir une expertise préalable des individus. Leur objectif est de construire un modèle de classement en fonction des individus à classer. A partir d'un ensemble d'individus provenant de chaque classe, elles établissent les frontières de décision (*decision boundary*) dans l'espace des caractéristiques permettant de séparer les individus qui appartiennent à des classes différentes. Dans la théorie de la décision, les frontières sont déterminées par les distributions des probabilités de chaque classe [Jain et al. 2000].
Dans notre travail, nous bénéficions de l'expertise de spécialistes qui ont déterminé les classes (*Sain* et *Pathologique*) et les sous-classes (*forme normale* et *forme boursouflée*, *texture homogène* et *texture non homogène*, etc.), ce qui permet d'utiliser des méthodes de classement.

Définition 2.2.2 (Classifieur) – *Le classifieur est l'algorithme mettant en œuvre la méthode de classement.*

Définition 2.2.3 (Capacité de prédiction) – *La capacité de prédiction d'une méthode est le pourcentage d'individus bien classés lors de la prédiction.*

NOTE - *Dans la littérature, les expressions "taux de classement", "capacité de validation" ou "pouvoir de prédiction" sont employées pour parler de la capacité de prédiction d'un classifieur.*

Les méthodes de classement (les classifieurs) utilisent le principe d'apprentissage sur les données. Mais tout comme un élève qui apprendrait sa leçon, l'apprentissage par cœur n'apporte que très peu d'intérêt. Ce que l'enseignant souhaite c'est que son élève puisse généraliser ce qu'il connaît, c'est-à-dire qu'il puisse appliquer ses connaissances sur un exercice sur lequel il ne s'est jamais entraîné. Il en est exactement de même des attentes des utilisateurs vis-à-vis des méthodes de classement ; il est possible d'avoir une procédure qui classe bien tous les individus lors de l'apprentissage mais qui ait un mauvais pouvoir de prédiction. Donc l'objectif d'un système d'apprentissage est de construire une procédure de classement qui doit non seulement classer correctement les individus lors de l'apprentissage mais qui ait en plus un bon pouvoir de prédiction. Ainsi, bien que s'appliquant à un problème spécifique, on souhaite que le classifieur puisse généraliser au sens des individus. On dit alors que le classifieur possède un *pouvoir prédictif*.

[1] Le choix de ses caractéristiques est l'enjeu majeur pour résoudre notre problème de classement et il occupe les parties I et II de ce manuscrit.

Les méthodes de classement utilisées

Les différentes définitions et autres notions liées au classement viennent d'être présentées. Les sous-sections suivantes décrivent les méthodes de classement utilisées tout au long de ce manuscrit. Afin d'obtenir les meilleurs modèles de classement, nous testons systématiquement quatre méthodes :

1. Les *k*-plus proches voisins.
2. La régression logistique.
3. Les forêts aléatoires.
4. Les réseaux de neurones.

Pour calculer les modèles, nous utilisons une implémentation logicielle de ces méthodes qui est réalisée dans la bibliothèque de fouille de données (*data mining*) *weka*[2].

2.2.2 Les *k* plus proches voisins

La méthode des *k plus proches voisins* (*k nearest neighbor*) est une des plus anciennes, plus simples et plus intuitives méthodes de classement [Fix and Hodges 1951; Fix and Hodges 1952; Shakhnarovich et al. 2006]. Elle peut se résumer par ce simple principe : *Dis moi qui sont tes voisins, je te dirai qui tu es !* Elle est motivée par le fait que des entrées semblables (proches) doivent avoir la même classe d'appartenance.
La première étape consiste à bien définir ce que l'on appelle *semblables* ou *proches*. Les individus sont représentés par leur vecteur caractéristique, donc ces notions se traduisent par la distance entre les individus. Pour cela on peut utiliser les distances existantes, mais chacune apporte un résultat différent. Il convient cependant de centrer/réduire les différents attributs (ou de les standardiser [Milligan and Cooper 1988]) pour que leur contribution à la distance soit comparable.
Pour classer un individu, la méthode consiste à rechercher les *k* plus proches voisins en terme de distance et de prédire le classement de l'individu à partir du classement des voisins. Chaque voisin apporte un vote utilisé dans le classement. En général le résultat du vote se fait à la majorité, mais il est possible d'affiner le résultat pour tenir compte du déséquilibre entre les classes [Zhang and Mani 2003].
Dans cette méthode, il est nécessaire de fixer le nombre de voisins *k* que l'on doit utiliser pour évaluer la classe de l'individu étudié. Ce paramètre influe grandement sur la qualité du classement. Si l'on considère un nombre réduit de voisins en choisissant une valeur de *k* trop petite, la frontière de décision[3] devient plus irrégulière et on risque d'être confronté à un sur-apprentissage. En revanche si l'on prend *k* trop grand, la frontière de décision est trop simplifiée. De même, si on choisit *k* égal au nombre d'individus dont on connaît le classement, on prédit juste la classe majoritaire (la plus fréquente) dans les individus. Il est donc nécessaire de tester différentes valeurs de *k* et de retenir la meilleure (au sens du taux de prédiction). Dans la littérature [Duda et al. 2000], on trouve deux heuristiques pour le choix de *k* :

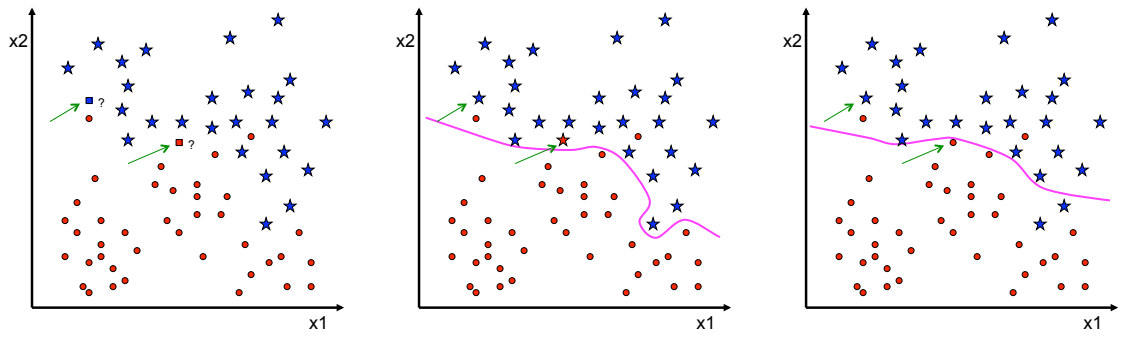
- la racine carrée du nombre d'individus.
- le nombre d'attributs additionné de 1.

[2] http://www.cs.waikato.ac.nz/~ml/weka/
[3] C'est la frontière séparant les classes d'individus dans l'espace des caractéristiques.

Données : x l'individu à classer, l'ensemble des individus E constituant l'échantillon d'apprentissage, le nombre de voisins k, une distance d.
Résultat : La classe d'appartenance supposée de x.
Début
  $kNN \leftarrow$ Null ;
  Initialiser tous les éléments du tableau $Vote$ à 0 ;
  Pour $i \leftarrow 1$ **à** k **Faire**
    Ajouter E_i dans kNN ;
  Pour $i \leftarrow k+1$ **à** $|E|$ **Faire**
    à
  Si $d(x,E_i) < max(d(x,kNN))$ **Alors**
    Supprimer l'individu le plus éloigné de x dans kNN ;
    Ajouter E_i dans kNN ;
  Pour $i \leftarrow 1$ **à** k **Faire**
    Incrémenter Vote[Classe(E_i)] ;
  Retourner l'indice de la valeur maximale contenue dans $Vote$;
Fin

Algorithme 1 : k plus proches voisins.

Si l'échantillon d'apprentissage est constitué de N individus décrits par un vecteur caractéristique de longueur d, la complexité de cet algorithme est en $O(kdN)$. La complexité et la recherche du paramètre k sont les principaux inconvénients de cette méthode car les temps de calcul peuvent rapidement devenir prohibitifs. De plus, elle ne fournit pas de modèle après la phase d'apprentissage et doit toujours utiliser un échantillon constituant la base d'apprentissage. Ce problème peut s'avérer gênant voire critique dans le cas de très grandes bases de données. En revanche, cette approche permet de modéliser des phénomènes non linéaires en se basant sur des informations locales. Plus les individus appartenant à la même classe sont proches (même s'ils sont séparés en plusieurs groupes dans l'espace des caractéristiques) plus la méthode modélise efficacement le problème. De plus cette méthode permet de travailler avec des populations contenant des classes déséquilibrées sans avoir à corriger le déséquilibre.



(a) (b) (c)

Figure 2.1. Illustrations 2D des résultats de classement par k-plus proches voisins. (a) échantillon d'apprentissage et deux individus à classer (la couleur montre la classe à trouver), (b) résultat du classement par 3-plus proches voisins et (c) résultat du classement par 20-plus proches voisins.

2.2.3 La régression logistique

La principale méthode de classement utilisée tout au long de notre travail est la régression logistique [Berkson 1944; McFadden 1973; Hosmer and Lemeshow 1989]. C'est un modèle linéaire multi-variables qui permet d'exprimer sous forme de probabilité la relation entre une variable Y catégorielle à deux modalités qui représente la cible et une ou plusieurs variables explicatives X_i. Dans notre problème, la variable cible est le diagnostic du noyau (sain ou pathologique) et les variables explicatives sont les classes d'altération. La régression logistique réalise une analyse statistique des individus de l'ensemble d'apprentissage et utilise une fonction de distribution logistique pour discriminer les individus :

$$P = P(Y/x) = \frac{e^{f(x)}}{1+e^{f(x)}} \quad \text{avec} \quad f(x) = \sum_i \alpha_i x_i$$

avec $x = (x_1, \ldots x_n)$ le vecteur caractéristique de l'individu en entrée et $P(Y/x)$ la probabilité conditionnelle P de la variable x d'appartenir à la classe Y. Cette méthode effectue une régression sur les individus, afin de séparer l'espace des caractéristiques à l'aide d'un plan d'équation $\alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n = 0$ (cf. figure 2.2*b*). Pour estimer les coefficients α_i du modèle, on utilise le plus souvent la méthode du maximum de vraisemblance qui maximise la probabilité d'obtenir les valeurs observées sur les échantillons de l'ensemble d'apprentissage. Elle consiste à rechercher les paramètres qui optimisent la fonction de vraisemblance :

$$\pounds(\alpha, Y) = P^Y [1-P]^{1-Y}$$

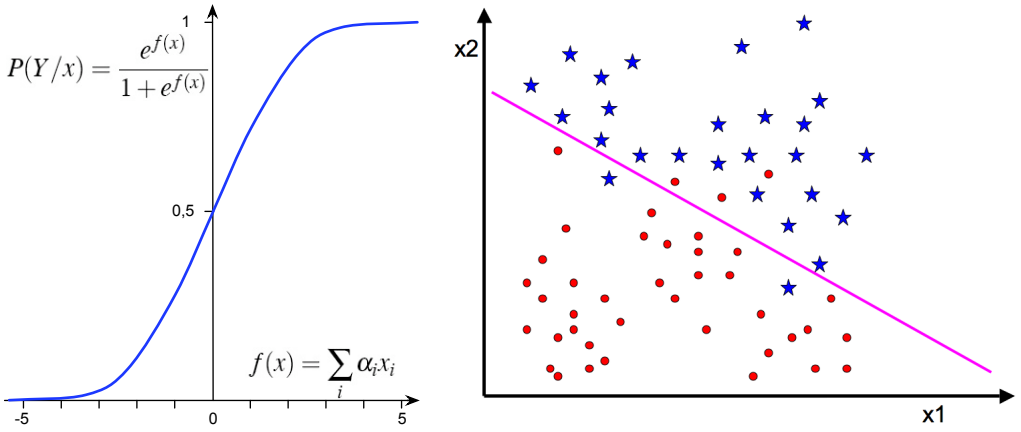


Figure 2.2. (a) Tracé de la fonction logistique. (b) Illustration 2D du résultat de la régression logistique qui construit une séparation linéaire (en violet) dans l'espace des caractéristiques.

Cette méthode s'adapte bien à notre problème à deux classes et elle permet en plus d'utiliser des variables explicatives (les caractéristiques) continues et discrètes (cf. chapitre 5). De plus, son résultat sous forme de probabilité est directement interprétable (cf. figure 2.2*a*) et elle permet même de modéliser certains problèmes non linéaires simples. Pour conclure, c'est l'une des méthodes de modélisation les plus fiables et de nombreux indicateurs faciles à mettre en place permettent de vérifier cette fiabilité.
En revanche, il est impératif que les variables explicatives ne soient pas linéairement liées, car deux variables colinéaires génèrent des problèmes numériques. Ces problèmes surviennent également si

les données sont linéairement séparables. Cette méthode ne permet pas de gérer les valeurs manquantes. En pratique dans de nombreuses applications, lors de l'acquisition il est possible que des mesures soient manquantes. Ceci empêcherait la méthode de classer l'individu, mais ce n'est pas le cas avec nos données. De plus, la régression logistique est sensible aux individus hors normes.

Tout au long de ce manuscrit, nous classons les attributs utilisés dans la régression logistique par ordre décroissant d'importance. Ce classement est effectué à l'aide du test du χ^2 (*khi-2*, *chi-2* ou *chi-carré*) qui vérifie qu'une variable suit une loi probabiliste donnée, en comparant les effectifs observés et théoriques. Ce test mesure la distance entre l'effectif *réellement observé* et l'effectif hypothétique *théorique* (*espéré*), celui qui serait observé si l'hypothèse nulle (H_0, cf. section 2.3.2) était vraie. Il y existe trois manière de calculer le χ^2 à partir de la régression logistique [Saporta 2006; Tufféry 2007; Wonnacott and Wonnacott 1998] :

- Le test de Wald [Cox and Hinkley 1978; Fears et al. 1996].
- Le taux de vraisemblance.
- Le test du score.

Dans ce manuscrit, le χ^2 est calculé à l'aide du taux de vraisemblance.

La généralité, la robustesse et l'interprétabilité sont les trois atouts majeurs de cette technique fondamentale qui en font une des méthodes linéaires les plus utilisées en classement.

2.2.4 Les forêts aléatoires

Les forêts aléatoires [Breiman 2001] (*random forests*) sont une méthode de classement non linéaire basée sur l'utilisation d'arbres de décision [Morgan and Sonquist 1963] de type *Classification And Regression Trees* (CART) [Breiman et al. 1984]. C'est l'un des derniers aboutissements dans la recherche d'agrégation d'arbres de décision *randomisés*. Elle synthétise les approches développées dans [Breiman 1996] et [Amit and Geman 1997].

Un arbre de décision réalise une décomposition du problème de classement en une suite de tests (des questions sur les caractéristiques) correspondant à une partition hiérarchique descendante de l'espace des caractéristiques en sous-régions rectangulaires homogènes en terme de classe. Il est constitué de la manière suivante :

- On choisit la caractéristique qui par ses modalités sépare au mieux les individus de chaque classe, puis on réitère sur chacune des sous-régions jusqu'à ce que la séparation ne soit plus possible ou souhaitable.
- Chaque question sur une caractéristique est représentée par un nœud et dépend de la réponse précédente.
- Le résultat de chaque question doit conduire à la construction de sous-régions de plus grande pureté (au sens des classes).
- Une feuille est un nœud terminal majoritairement constitué d'une seule classe. La proportion de la classe majoritaire de la feuille désigne la probabilité de l'individu à classer d'appartenir à la classe majoritaire.
- L'ensemble des règles de toutes les feuilles constitue le modèle de classement.

Un arbre développé jusqu'à ce que l'on ne puisse plus améliorer les régions sur la base des variables disponibles est dit *complet*, mais il apprend par cœur sur les individus. Pour généraliser, une opération d'élagage est effectuée sur les branches de l'arbre. Cette opération diminue les capacités d'apprentissage, mais accroît le pouvoir de prédiction.
Dans la littérature, les types d'arbres se différencient par les techniques permettant le choix de la meilleure variable séparatrice (indice de Gini [Gini 1921; Tufféry 2007], entropie de Shannon[Pun 1980; Pun 1981], test du χ^2, etc.), les critères d'arrêts (qualité de l'arbre, profondeur de l'arbre, nombre de feuilles, effectif des nœuds, etc.) et les techniques d'élagage.

Une forêt aléatoire est un classifieur constitué d'une collection de prédicteurs structurés en arbres dans lequel chaque arbre fournit un vote unitaire et la classe la plus populaire[4] est le résultat de la prédiction. La particularité des forêts aléatoires est la double "randomisation" utilisée lors de la construction :

- La première consiste à sélectionner aléatoirement avec remise un sous-échantillon de l'échantillon d'apprentissage pour construire chaque arbre.
- La deuxième est l'utilisation aléatoire d'un sous-ensemble de variables explicatives pour construire chaque nœud de chaque arbre. En général, la taille de ce sous-ensemble est égal à la racine carrée du nombre de variables prédictives.

Tirer aléatoirement des sous-échantillons d'apprentissage et les variables explicatives permet d'obtenir des arbres peu corrélés afin d'accroître la puissance de prédiction. Chacun des arbres construits est moins performant qu'un arbre de décision (car il dispose de moins de variable pour son élaboration), mais *l'union fait la force*.

Données : L'ensemble des individus I décrits par D descripteurs (variables explicatives), le nombre d'arbres à construire A, le nombre de descripteurs à utiliser d.
Résultat : A arbres que l'on fait voter pour le classement.
Début
  Pour tous les *arbres* $a \leftarrow 1 \dots A$ **Faire**
    Tirer aléatoirement avec remise un échantillon I_a parmi I ;
    Estimer un arbre complet sans élagage sur I_a avec randomisation des variables :
    – **Pour tous les** *nœuds* $n \in a$ **Faire**
      Tirer aléatoirement et uniformément d variables parmi D pour construire la décision associée à n ;
Fin

Algorithme 2 : Forêts aléatoires.

En utilisant les ensembles d'arbres on obtient une amélioration significative de la prédiction par rapport aux techniques classiques basées sur les arbres de décision. En effet, cette structure permet d'éviter le sur-apprentissage. Dans [Breiman 2001] l'auteur démontre empiriquement que lorsque le nombre d'arbres de la forêt augmente, le taux d'erreur converge vers une limite. Une borne supérieure de cette limite peut être estimée à partir des caractéristiques intrinsèques de la forêt. De plus, le nombre de variables explicatives fixé avant la construction de la forêt (d dans l'algorithme) n'a que peu d'influence sur le taux d'erreur. Lorsque ce nombre croît, le taux d'erreur atteint un minimum avant d'augmenter progressivement en restant dans des valeurs proches. De plus l'auteur démontre que si on choisit $d = 1$, donc si on construit des arbres dans lesquels chaque nœud est

[4] La classe ayant reçu le plus grand nombre de votes.

construit à partir d'une variable explicative aléatoire, la prédiction de la forêt ainsi construite est inférieure de moins de 1% par rapport aux forêts construites avec des valeurs de d plus grandes. La très faible sensibilité de cette méthode quant aux choix des paramètres fixés, lui confère un grande stabilité qui est son principal avantage en plus de sa puissance de prédiction.

2.2.5 Les réseaux de neurones

Il est aujourd'hui impossible de parler de classement ou de classification sans parler de réseaux de neurones. Ils sont largement répandus grâce à leur puissance de modélisation (ils peuvent approcher n'importe quelle fonction suffisamment régulière), qui fait merveille pour résoudre une grande variété de problèmes, face à des phénomènes complexes, des données difficiles à appréhender et ne suivant pas de lois probabilistes particulières (figure 2.3).

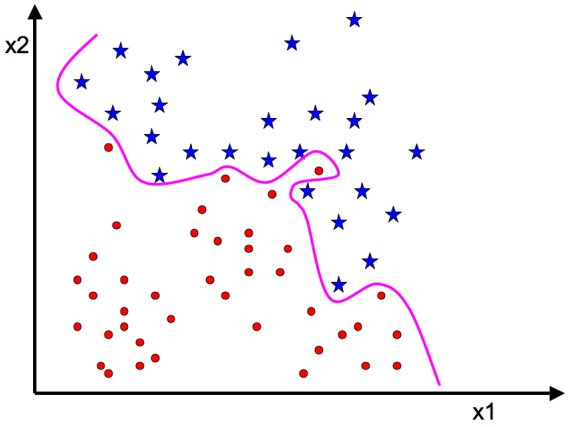


Figure 2.3. Illustration 2D du résultat d'un réseau de neurones ; une séparation non linéaire complexe (en violet) de l'espace des caractéristiques.

Les réseaux de neurones (appelés parfois *réseaux de neurones artificiels*, en anglais *artificial neural network*) sont nés en 1943 [McCulloch and Pitts 1943] dans une tentative de modélisation mathématique du cerveau humain. Ils ont une architecture calquée sur celle d'un cerveau, organisée en neurones et en synapses. Ils se présentent sous forme de nœuds (les neurones, figure 2.4) connectés entre eux par des fonctions de transfert (les synapses). Le plus souvent ces fonctions sont des sigmoïdes comme le montre la figure 2.4. Dans la forme la plus courante de réseaux de neurones, le *perceptron* [Rosenblatt 1958], les nœuds sont regroupés en couches. Un réseau est en général composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente. Les réseaux comportent donc au minimum deux couches :

- Une couche d'entrée qui reçoit les informations (les caractéristiques) de l'individu à classer. Chaque information est transmise à un seul nœud et le nombre de nœuds est par conséquent égal à la dimension du vecteur caractéristique (cf. section 2.2.1). Les nœuds de la couche d'entrée sont triviaux dans la mesure où ils ne combinent rien et ne font que transmettre la valeur de la variable qui leur correspond.
- Une couche de sortie qui contient autant de nœuds que de classes.

Entre la couche d'entrée et la couche de sortie des nœuds sont souvent connectés. Ils appartiennent à un niveau intermédiaire : la couche cachée. Il peut parfois exister plusieurs couches cachées.

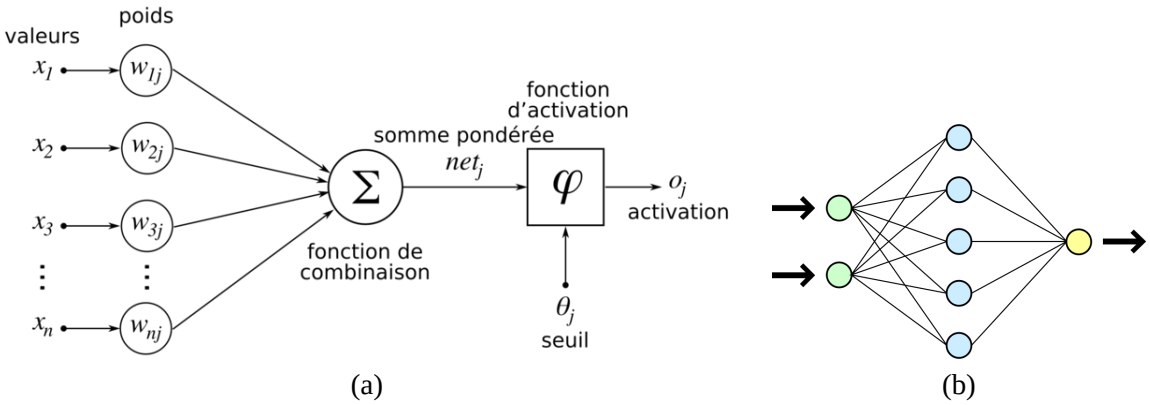


Figure 2.4. (a) Représentation d'un neurone ; les valeurs transmises par les synapses sont traitées par la fonction de combinaison (pour le perceptron, il s'agit d'une somme pondérée), puis une fonction d'activation détermine la sortie en fonction du seuil. (b) Schéma d'un réseau de neurones contenant deux caractéristiques en entrée, puis une couche cachée.

L'apprentissage du réseau s'effectue en ajustant le poids w_i des connexions entre les nœuds. Au cours de l'apprentissage, la valeur renvoyée par le nœud en sortie est comparée à la valeur réelle souhaitée et les poids w_i de tous les nœuds sont ajustés de façon à améliorer la prédiction par un mécanisme dépendant du type de réseau de neurones. Un mécanisme encore couramment utilisé est la rétro propagation du gradient, mais il en existe d'autres plus récents et plus performants : les algorithmes de Levenberg-Marquardt, quasi Newton, du gradient conjugué, de propagation rapide ou encore les algorithmes génétiques. L'échantillon d'apprentissage est parcouru de nombreuses fois, parfois plusieurs milliers de fois. L'apprentissage s'arrête lorsqu'une solution optimale a été trouvée et que les poids w_i ne sont plus modifiés significativement ou lorsque qu'un nombre d'itérations préfixé a été atteint.

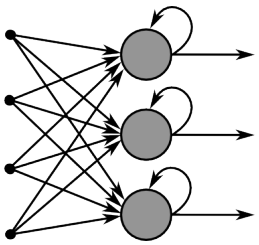


Figure 2.5. Illustration du principe de la rétro-propagation.

La structure d'un réseau de neurones (encore appelée topologie ou architecture) est constitué du nombre de couches et de nœuds, la façon dont sont interconnectés les différents nœuds (choix des fonctions de combinaison, de transfert et d'activation) et le mécanisme d'ajustement des poids. Le choix de cette structure détermine grandement les résultats qui seront obtenus et constitue le point délicat dans la mise en œuvre d'un réseau de neurones.
On augmente le pouvoir d'apprentissage en ajoutant une ou plusieurs couches cachées entre les couches d'entrée et de sortie. Bien que le pouvoir d'apprentissage augmente avec le nombre de couches cachées et de nœuds dans ces couches, ce nombre doit néanmoins être limité pour que le réseau de neurones ne se contente pas de mémoriser l'ensemble d'apprentissage mais puisse

le généraliser, en évitant ce que l'on appelle le sur-apprentissage (cf. section 2.3) qui survient lorsque les poids ne font qu'apprendre les particularités de l'ensemble d'apprentissage au lieu d'en découvrir les structures générales.
Dans notre travail, nous utilisons la version standard des réseaux de neurones implémentés dans la bibliothèque *weka*. Ces derniers sont des perceptrons multicouches dont le nombre de neurones de la couche cachée est déterminé pour chaque modèle par le calcul suivant :

$$\frac{N_{Input} + N_{Output}}{\nu}$$

avec $\nu \in \mathbb{N}^*$ le paramètre à déterminer pour chaque modèle et N_{Input} (resp. N_{Output}) le nombre de neurones de la couche d'entrée (resp. de sortie).
Malgré la puissance de modélisation des réseaux de neurones, leur utilisation est parfois freinée par les difficultés qu'elle présente : le coté boite noire des réseaux, la délicatesse des réglages à effectuer dans le choix de l'architecture (nombre de couches cachées, nombre de neurones par couches, méthode d'optimisation, fonction de transfert et condition d'arrêt), la puissance informatique requise, le risque de convergence vers une solution globalement non optimale. Un risque de sur-apprentissage survient lorsque la taille de l'ensemble d'apprentissage est trop faible par rapport à la complexité/topologie du modèle. C'est pourquoi il est préférable de commencer par tester des modèles de classement moins complexes (linéaires), puis si la modélisation (le résultat) n'est pas satisfaisante, on peut alors essayer d'utiliser un réseau de neurones.

2.3 La validation

Pour vérifier les capacités d'un classifieur, les individus sont séparés en deux groupes : un échantillon d'apprentissage et un échantillon de validation, c'est ce que l'on nomme la *validation croisée*. Le classifieur effectue son apprentissage sur un certain nombre d'individus (l'échantillon d'apprentissage) puis son efficacité est validée à l'aide d'individus sur lesquels le classifieur n'a pas appris (l'échantillon de validation). Il doit garder des performances proches lors de l'apprentissage et de la validation.
Ces performances sont en partie liées à la taille du vecteur caractéristique qui joue un rôle important dans un modèle de classement. Le risque majeur lorsque l'on fournit trop de caractéristiques au classifieur est le *sur-apprentissage* ou *apprentissage par cœur*. Plus la dimension du vecteur est grande, plus le modèle est adaptable et donc plus le classement est bon, mais plus la validation du modèle à l'aide d'individus non utilisés dans la phase d'apprentissage est mauvaise. Il faut alors systématiquement valider chaque modèle construit et obtenir l'écart le plus faible entre bon classement et bonne prédiction.

Il existe plusieurs protocoles de validation parmi lesquels ceux utilisés dans ce manuscrit sont présentés dans les sous-sections suivantes.

2.3.1 Les protocoles de validation

2.3.1.1 K-fold validation croisée

Le *K-Fold* [Stone 1974; Kohavi 1995; Dietterich 1998] ou *cross validation* est un protocole dans lequel les individus sont séparées en K groupes de tailles identiques. L'apprentissage s'effectue sur $K-1$ groupes et la validation sur le groupe écarté lors de l'apprentissage. Cette opération est réitérée pour tous les groupes, puis on effectue la moyenne des taux de bon classement et de bonne prédiction. Ces deux moyennes forment le taux d'apprentissage et de validation du modèle. Dans ce protocole, chaque individu sert $K-1$ fois à l'apprentissage et une fois à la validation. On s'affranchit ainsi des contraintes liées à la sélection aléatoire des individus, comme c'est le cas dans la validation croisée.

2.3.1.2 Leave-one-out

Le *Leave-One-Out* [Dietterich 1998; Martens and Dardenne 1998] est une protocole de validation de type *k-fold*, avec k égal au nombre d'individus : à chaque itération, un individu est retiré, puis l'apprentissage est effectué sur les individus restants et la validation sur l'individu préalablement retiré. Ensuite, on réitère sur tous les individus et on calcule le pourcentage d'individus bien classés. Ce protocole est particulièrement conseillé lorsque le nombre d'individus est faible (inférieur à 400). L'inconvénient est que l'on obtient autant de modèles que d'individus. Pour démontrer sa performance, il convient de construire un dernier modèle sur tous les individus puis de valider sur ces mêmes individus. Si l'écart de performance entre ce dernier modèle et le pourcentage obtenu par *leave-one-out* est important, alors cela implique que l'on est en situation de sur-apprentissage car retirer un seul individu influence grandement les résultats. En revanche, des performances équivalentes démontrent un bon pouvoir de généralisation.

2.3.1.3 Holdout

Lors de la mise en place d'un protocole de *holdout* [Kohavi 1995], on réitère plusieurs fois un protocole *2-fold*. Afin d'assurer une bonne variabilité des individus dans les échantillons, l'opération est réitérée un grand nombre de fois[5] et la moyenne des pourcentages de prédiction est effectuée. Les individus étant choisis aléatoirement, il est possible que certains soient systématiquement dans le même échantillon. Effectuer plusieurs itérations permet de réduire la probabilité qu'un tel cas survienne, mais celle-ci n'est jamais nulle. Pour s'assurer de la robustesse de ce protocole, il est possible de mesurer les écarts de prédiction entres les différentes itérations. Un écart important signifierait une défaillance du protocole.

2.3.1.4 Bootstrap

Le protocole *bootstrap* [Efron 1979; Dietterich 1998] consiste à construire un échantillon d'apprentissage de la taille de l'échantillon initial par tirage aléatoire avec remise des individus. Les individus n'ayant pas été tirés constituent l'échantillon de validation. Si l'échantillon initial est de taille N, la probabilité d'un individu donné d'être tiré est $1-(1-\frac{1}{N})^N$ et tend vers $0,632$ lorsque N tend vers l'infini. Le principal inconvénient de cette méthode est sa lenteur d'exécution car il faut réitérer l'opération un grand nombre de fois (au minimum de l'ordre de 200 fois). Tout comme le

[5] Cela rend le protocole lent, ce qui constitue son principal inconvénient.

leave-one-out, il est possible d'évaluer la robustesse du protocole en analysant les variations des performances sur l'ensemble des itérations.

2.3.2 Hypothèse nulle

L'étape de validation permet de déterminer les performances d'un modèle. Mais on peut se demander si cette valeur est pertinente. Autrement dit, quelle serait la probabilité d'obtenir les mêmes performances en choisissant aléatoirement les résultats de classement ? L'hypothèse nulle (notée H_0) [Wonnacott and Wonnacott 1998; Saporta 2006] permet d'estimer cette probabilité.
Pour obtenir une estimation, on choisit un protocole de validation, durant lequel les étiquettes des classes d'appartenance des individus de l'échantillon d'apprentissage sont aléatoirement permutées. Donc le classifieur apprend sur des données erronées. Puis on valide sur l'échantillon de validation dont les étiquettes n'ont pas été permutées et le pourcentage de prédiction obtenu est sauvegardé. Cette opération est réitérée N fois. Ensuite, le pourcentage de prédiction du modèle est calculé (de manière classique, sans permutation). Si le pourcentage de prédiction est supérieur au meilleur des pourcentages obtenus lors des N itérations avec permutations, alors on peut estimer que la probabilité d'obtenir cette performance grâce au hasard est inférieure à $\frac{1}{N}$. En revanche, si le pourcentage de prédiction du modèle est le n-ième plus grand, alors la probabilité est de l'ordre $1 - \frac{n}{N}$.

Données : L'ensemble des individus I, un protocole de validation PV, le nombre d'itération N.
Résultat : P la probabilité du modèle.
Début
  Appliquer PV sur I et construire le modèle de classement ;
  $p \leftarrow$ Pourcentage de prédiction du modèle ;
  Pour $i \leftarrow 1$ **à** N **Faire**
    Appliquer PV sur I en permutant aléatoirement les étiquettes de l'échantillon d'apprentissage et construire le modèle de classement ;
    $Prédictions[i] \leftarrow$ Pourcentage de prédiction du modèle qui vient d'être calculé ;
  Trier($Prédictions$) ;
  Si $Prédictions[N] < p$ **Alors**
    P $\leftarrow \frac{1}{N}$;
  Sinon
    $n \leftarrow$ Position(p dans $Prédictions$) ;
    $P \leftarrow 1 - \frac{n}{N}$;
Fin

Algorithme 3 : Calcul de la probabilité d'un modèle par l'hypothèse nulle.

Dans la suite de ce manuscrit, la probabilité de chaque modèle est systématiquement calculée à l'aide de cet algorithme pour 10000 itérations.

2.3.3 Intervalle de confiance

Lorsque l'on construit les échantillons d'apprentissage et de validation, on effectue un tirage aléatoire de deux sous-ensembles sur une population. Le pourcentage de prédiction calculé à partir de ces deux échantillons est soumis en partie au hasard. Ainsi il peut exister un écart de performance

pour deux modèles de classement d'une population utilisant la même technique de classement, mais construit sur deux échantillons d'apprentissage différents. Donc on obtient une estimation du pourcentage de prédiction du modèle, mais pas sa valeur exacte. Il est souhaitable de pouvoir dire à partir de l'estimation que le pourcentage de prédiction exact est dans l'intervalle $[a,b]$ avec un risque d'erreur inférieur ou égal à ε%. On dit alors que $[a,b]$ est un *intervalle de confiance* avec un risque d'erreur de ε% ou un degré de confiance de $(100-\varepsilon)$%.

D'une manière générale en statistique et en particulier dans la théorie des sondages, quand on cherche à estimer la valeur d'un paramètre, on parle d'intervalle de confiance lorsque l'on donne un intervalle qui contient avec un certain degré de confiance (exprimé sous la forme d'une probabilité) la valeur à estimer. Par exemple, un intervalle de confiance à 95% (ou au risque d'erreur de $\varepsilon = 5\%$) a une probabilité égale à $0,95$ de contenir la valeur du paramètre que l'on cherche à estimer. Une estimation par intervalle de confiance est d'autant meilleure que l'intervalle est petit pour un degré de confiance grand. La longueur de l'intervalle de confiance est donc une mesure de l'incertitude sur la position de la valeur exacte du paramètre estimé.

Afin de calculer l'intervalle de confiance d'un paramètre X, on effectue N estimations (au minimum 200, par *bootstrap*, cf. algorithme 4) de ce dernier. A partir de la distribution des estimations, il existe différents algorithmes pour calculer l'intervalle de confiance [Gosh 1979; Newcombe 1998]. Une solution simple est de supprimer ε% des valeurs extrêmes ($\frac{\varepsilon}{2}$% des valeurs à chaque extrémité de la distribution) parmi les N estimations afin d'obtenir un intervalle de confiance à $(100-\varepsilon)$%. Parmi les valeurs restantes, les extrêmes forment l'intervalle de confiance du paramètre.
Si la distribution des estimations est normale, il y a une relation entre le pourcentage d'erreur et l'écart type. Un calcul de l'intervalle de confiance est alors :

$$\left[\overline{X} - \alpha\sigma\left(\overline{X}\right), \overline{X} + \alpha\sigma\left(\overline{X}\right)\right]$$

avec $\alpha > 0$ le coefficient à choisir en fonction du pourcentage d'erreur souhaité (ε), $\overline{X}$ et $\sigma(\overline{X})$ sont respectivement la moyenne et l'écart type empirique des estimations de X. Le tableau 2.1 donne quelques couples de valeurs de relation entre le coefficient α et le pourcentage d'erreur ε.

ε (%)	0,3	1	2	5	10	32
α	3	2,5758	2,3263	1,96	1,645	1

Table 2.1. Relations entre pourcentage d'erreur ε et coefficient α de l'écart type empirique pour une distribution normale des estimations.

Dans notre travail, nous utilisons l'algorithme 4 afin de calculer l'intervalle de confiance avec un pourcentage d'erreur $\varepsilon = 5\%$.

Données : Un échantillon d'individus I, une technique de classement C, un protocole de validation V, le nombre N d'échantillons *BootStrap* à calculer, la valeur du coefficient α ou le pourcentage d'erreur ε.
Résultat : Les bornes Inf et Sup de l'intervalle de confiance.
Début
  Créer aléatoirement E échantillons en fonction de V ;
  Pour $i \leftarrow 1$ **à** E **Faire**
    $ValidationSet \leftarrow E_i$;
    Pour $j \leftarrow 1$ **à** N **Faire**
      $LearningSet \leftarrow$ Echantillon d'apprentissage construit par *BootStrap* sur $I \setminus \{E_i\}$;
      $Classifier \leftarrow$ Modèle de classement construit sur $LearningSet$ en utilisant C ;
      $Predictions[i][j] \leftarrow$ Pourcentage de prédiction de $Classifier$ sur $ValidationSet$;
  Pour $j \leftarrow 1$ **à** N **Faire** `/* Moyenne des` $j^{ième}$ `prédictions */`
    $Distributions[j] \leftarrow$ Moyenne ($Predictions[*][j]$) ;
  Si *Distribution est une distribution normale* **Alors**
    $\overline{X} \leftarrow$ Moyenne($Distributions[*]$) ;
    $S(\overline{X}) \leftarrow$ Ecart type($Distributions[*]$) ;
    $Inf \leftarrow \overline{X} - \alpha S(\overline{X})$;
    $Sup \leftarrow \overline{X} + \alpha S(\overline{X})$;
  Sinon
    Trier($Distributions$) ;
    $Inf \leftarrow Distributions[\frac{\varepsilon}{2}N]$;
    $Sup \leftarrow Distributions[(1-\frac{\varepsilon}{2})N]$;
Fin

Algorithme 4 : Calcul de l'intervalle de confiance.

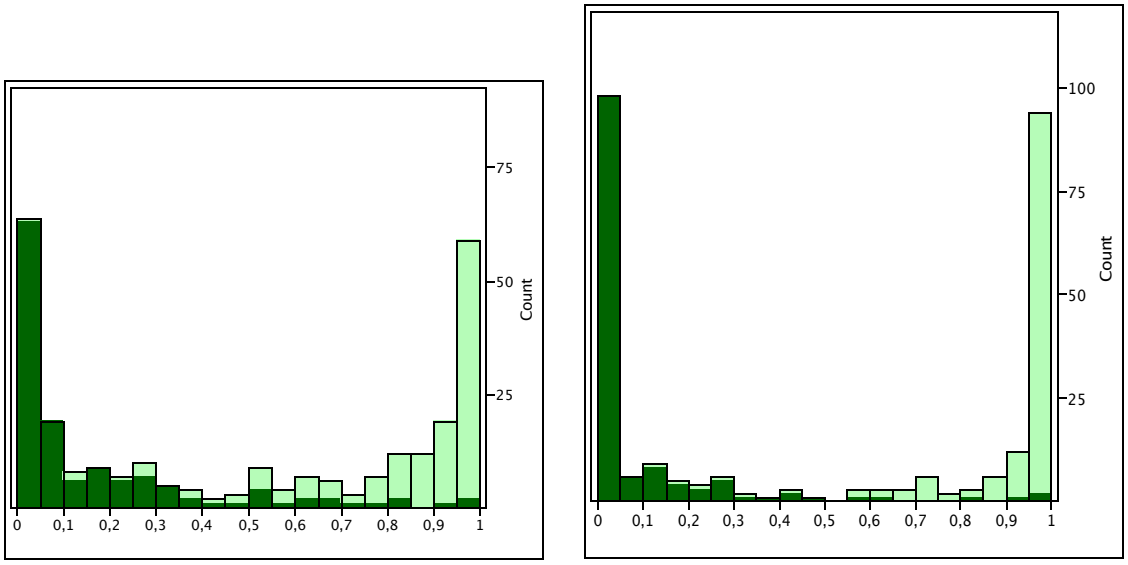
2.3.4 Performance optimale d'un modèle

Il reste une question essentielle à laquelle il faut répondre : *à partir de quel taux de bon classement peut-on considérer que le modèle est satisfaisant ?* Autrement dit, quel pourcentage de prédiction doit-on dépasser pour être au moins satisfait par le modèle ?
Une réponse est apportée par les experts eux-mêmes, par ce que l'on nomme *le taux de répétabilité*. Le taux de répétabilité est le taux de concordance obtenu par les experts lors de deux classements successifs des individus. Si par exemple un expert obtient un taux de répétabilité de 85%, cela implique qu'il n'est d'accord avec lui même que sur 85% des individus. Donc un modèle est vraiment satisfaisant lorsque celui-ci obtient un taux de classement supérieur ou égal au taux de répétabilité des experts.

2.3.5 Histogramme des probabilités

Un histogramme est un outil visuel et qualitatif qui permet l'étude de la dispersion (répartition statistique) d'une variable.
Un classifieur attribue une probabilité de classement à chaque individu. Représenter ces probabilités par un histogramme apporte des informations sur l'efficacité du classifieur :

1. Au plus les probabilités sont distribuées sur les extrémités, au mieux le classifieur parvient à séparer les individus dans les classes d'appartenance.
2. Le classifieur doit générer un minimum de *cas ambigus*. Dans le cas d'un problème de classement à deux classes, le seuil de décision est $0,5$. Donc plus la probabilité attribuée par le classifieur est proche de $0,5$, plus l'incertitude de classement est élevée. Pour notre travail, nous définissons ainsi un intervalle d'incertitude $[0,3 \ldots 0,7]$: tout individu ayant une probabilité de classement comprise dans cet intervalle est considéré comme un cas ambigu.
3. A l'inverse plus la probabilité est proche des extrémités (0 et 1), plus l'incertitude est faible. Malgré la probabilité forte d'appartenance, il peut survenir que la classe d'attribution soit erronée. C'est ce que l'on nomme des *erreurs graves*. Ces erreurs graves portent sur des individus auxquels le classifieur attribue la mauvaise classe d'appartenance et une probabilité forte (dans notre travail ce sont des probabilités inférieures à $0,2$ ou supérieures à $0,8$). Donc un classifieur doit générer un minimum d'erreurs graves.



(a) (b)

Figure 2.6. Exemples d'histogrammes de probabilités attribuées par deux modèles de classement sur une même population. L'histogramme (a) comporte un nombre de cas ambigus plus important et l'histogramme (b) montre une meilleure distribution sur les extrémités ainsi qu'un nombre d'erreurs graves plus faible. Ces différences permettent de conclure que le classifieur qui a engendré les probabilités de l'histogramme (b) est plus efficace.

2.3.6 Echantillon de données déséquilibrées

Dans la pratique[6], il est fréquent d'avoir des échantillons de données dans lesquels une ou plusieurs classes d'appartenance sont sous-représentées en terme d'individus. C'est ce que l'on nomme des données déséquilibrées. Dans un problème à deux classes, la classe la plus (resp. moins) fréquente est dite majoritaire (resp. minoritaire). Si l'on ne tient pas compte de la probabilité a priori des classes, ce genre de situation biaise les performances des classifieurs : si on construit un modèle

[6]Nous rencontrons ce problème dans la partie II de ce manuscrit.

de classement sur un échantillon dont la classe majoritaire contient 95% des individus, alors le classifieur *pourrait se contenter* de classer tous les individus dans la classe majoritaire et ainsi obtenir 5% d'erreur de classement. Le classifieur n'est alors pas capable de généraliser. En particulier son taux d'erreur de classement sur la classe minoritaire est important. Or en général, la classe minoritaire est la classe d'intérêt, d'où l'importance de calculer le taux d'erreur sur les deux classes. Les problèmes engendrés sont étudiés dans [Japkowicz 2000a; Japkowicz and Stephen 2002]. Ce problème est considéré comme l'un des dix problèmes les plus importants en fouille de données [Yang and Wu 2006]. Les solutions [Japkowicz 2000b; Weiss and Provost 2003; Visa and Ralescu 2005] peuvent intervenir à deux niveaux : soit sur les algorithmes, soit sur les individus.

Pour les algorithmes, une première famille de solutions rééquilibre le taux d'erreur par une pondération de chaque type d'erreur [Domingo 1999]. Une comparaison de ces solutions est étudiée dans [Liu and Zhou 2006]. D'autres méthodes basées sur les arbres de décisions agissent sur l'ajustement des estimations de probabilité dans les feuilles et les seuils de décision ou encore sur le choix de la mesure de chaque nœud par décentrage de l'entropie [Lallich et al. 2007; Ritschard et al. 2007; Zighed et al. 2007; Do et al. 2008; Lenca et al. 2008; Marcellin et al. 2008a; Marcellin et al. 2008b].

Il existe trois solutions pour intervenir sur les individus :

1. Le sur-échantillonnage (*over-sampling*) qui consiste à dupliquer de façon aléatoire ou dirigée les individus de la classe minoritaire jusqu'à équilibre des effectifs [Liu et al. 2007]. Certaines méthodes sont basées sur la génération d'individus en effectuant un tirage aléatoire dans la distribution (supposée normale) de chacune des variables.
2. Le sous-échantillonnage (*down-sizing* ou *under-sampling*) qui réduit de manière aléatoire ou dirigée la taille de la classe majoritaire [Kubat and Matwin 1997; Liu et al. 2006b; Liu et al. 2009]. Il fait perdre de l'information sur la classe majoritaire.
3. L'apprentissage sur les individus par l'utilisation d'un réseau de neurones auto-associateur [Japkowicz 2000b].

Les difficultés d'apprentissage portent sur des individus proches de la frontière de décision. Dans le cas du sur-échantillonnage il est donc préférable de dupliquer majoritairement ces individus. De même, lors du sous-échantillonnage garder les individus proches de la frontière améliore l'apprentissage. Malheureusement, la frontière n'est connue qu'après apprentissage.

Sélectionner aléatoirement les individus à supprimer/dupliquer n'est pas pertinent et peut générer des écarts de performance importants lors de plusieurs tests successifs. De même, l'utilisation de méthodes complexes dans le choix des individus à supprimer/dupliquer n'améliore pas de manière significative les résultats [Japkowicz 2000b]. En général, le sous-échantillonnage apporte de meilleurs résultats que le sur-échantillonnage [Japkowicz 2000b; Drummond and Holte 2003; Liu et al. 2006b], mais cela dépend essentiellement des données.

Dans notre travail (partie II), nous utilisons le sous-échantillonnage. Nous souhaitons supprimer les individus les moins représentatifs de la classe majoritaire. Pour cela, nous utilisons une méthode de classification (cf. section 2.4) afin de sélectionner un nombre d'individus représentatifs (les *parangons*, cf. définition 2.4.5) égal à la taille de la classe minoritaire. La méthode de classification choisie est les *k*-moyennes (*k-means*, cf. section 2.4.2) pour son efficacité. Elle est initialisée par formes fortes pour accroître sa stabilité (réduire son aspect stochastique) [Diday 1972].

L'algorithme de construction de l'échantillon de travail dans le cas de données avec classes déséquilibrées est le suivant :

Données : L'ensemble des individus I
Résultat : Un échantillon de travail équilibré E
Début
  Ajouter les individus de la classe minoritaire dans E ;
  $N = card(E)$;
  Supprimer les individus de la classe minoritaire de I ;
  Calculer N formes fortes par k-moyennes sur I ;
  Effectuer les K-moyennes à N classes initialisées avec les formes fortes ;
  Extraire les N parangons et les ajouter à E ;
Fin

Algorithme 5 : Construction de l'échantillon de travail dans le cas de données déséquilibrées.

2.4 La classification

2.4.1 Définitions

Définition 2.4.1 (Classification) – *La classification [Jain et al. 1999] ("clustering") est l'opération statistique qui consiste à regrouper des objets (individus ou variables) en un nombre limité de groupes (les classes) qui ont deux propriétés :*

1. *Elles ne sont pas prédéfinies par l'expert, mais découvertes au cours de l'opération, contrairement aux classes du classement.*
2. *Elles regroupent les objets ayant des caractéristiques similaires et séparent les objets ayant des caractéristiques différentes (homogénéité interne et hétérogénéité externe), ce qui peut être mesuré par des critères tels l'inertie inter-classe (définition 2.4.4) et/ou l'inertie intra-classe (définition 2.4.3).*

Comme le classement, la classification consiste à répartir des objets en groupes (figure 2.8). Toutefois, cette répartition n'est pas effectuée en fonction d'un critère prédéfini et ne vise pas à rassembler les objets possédant la même valeur pour ce critère. Donc on ne sait pas à l'avance à quelle classe appartient chaque objet. Ainsi, la classification est descriptive et non pas prédictive.
Le nombre de partitions possibles pour une classification est donné par le nombre de Bell :

$$B_n = \frac{1}{e}\sum_{k=1}^{\infty}\frac{k^n}{k!}$$

Par exemple, pour $n = 4$ individus, on a $B_4 = 15$.
- 1 partition à 1 classe $(abcd)$.
- 7 partitions à 2 classes (ab,cd), (ac,bd), (ad,bc), (a,bcd), (b,acd), (c,abd), (d,abc).
- 6 partitions à 3 classes (a,b,cd), (a,c,bd), (a,d,bc), (b,c,ad), (b,d,ac), (c,d,ab).
- 1 partition à 4 classes (a,b,c,d).

Pour $n = 30$, on obtient $B_{30} = 8.47\ 10^{23}$ et en règle générale $B_n > e^n$. Il est par conséquent nécessaire de définir des critères de bonne classification car on ne peut tester toutes les combinaisons possibles.

Définition 2.4.2 (Inertie totale) – *L'inertie totale I_T d'une population est la moyenne des carrés des distances des individus au centre (barycentre) de la population. Soit P une population constituée de N individus $P = \{I_1, I_2, ...I_N\}$ et de centre B_P (le barycentre non pondéré), alors :*

$$I_T(P) = \frac{1}{N} \sum_{k=1}^{N} d(B_P, I_k)^2$$

L'inertie totale d'une classe permet d'évaluer l'hétérogénéité de la classe. Une classe est d'autant plus homogène (homogénéité interne, cf. définition 2.4.1) et cohérente que son inertie est faible.

Définition 2.4.3 (Inertie intra-classe) – *L'inertie intra-classe I_A (appelée aussi erreur intra-classe) est la somme des inerties totales de chaque classe. Soit P une partition constituée de γ classes $P = \{C_1, C_2, ...C_\gamma\}$, alors :*

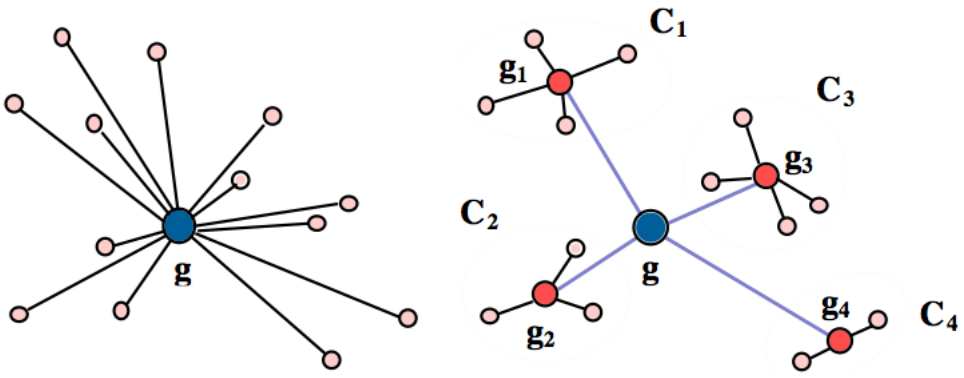
$$I_A = \sum_{k=1}^{\gamma} I_T(C_k)$$

L'inertie intra-classe permet d'évaluer l'hétérogénéité à l'intérieur des classes. La classification de la population est d'autant meilleure que I_A est faible (cf. figure 2.8*a*).

Définition 2.4.4 (Inertie inter-classe) – *L'inertie inter-classe I_E est la moyenne (pondérée par l'effectif de chaque classe) des carrés des distances des centres de chaque classe au barycentre global. Soit P une partition constituée de γ classes $P = \{C_1, C_2, ...C_\gamma\}$, B_{C_i} le centre de la classe C_i (le barycentre non pondéré) et B le barycentre global, alors :*

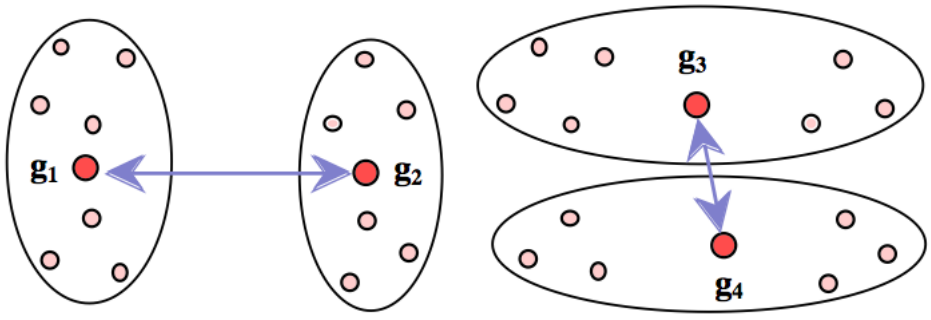
$$I_E = \sum_{k=1}^{\gamma} d(B, B_{C_k})^2 \text{ , avec } B = \frac{1}{\sum_{k=1}^{\gamma} |C_k|} \sum_{k=1}^{\gamma} |C_k| \cdot B_{C_k}$$

Plus l'inertie inter-classe est grande, plus les classes sont séparées, donc hétérogènes (hétérogénéité externe, cf. définition 2.4.1) et par conséquent meilleure est la classification (cf. figure 2.8*a*).



(a) (b)

Figure 2.7. Illustrations des inerties : (a) l'inertie totale d'une population, (b) l'inertie inter-classe (bleu) et intra-classe. Avec g le barycentre de la population et g_i le barycentre de la classe c_i.



(a) (b)

Figure 2.8. Deux exemples de classification d'une population : (a) une classification avec une inertie intra-classe faible et une inertie inter-classe élevée, (b) le contraire.

2.4.2 Les K-moyennes

La méthode des *k-moyennes* (*K-means*) [MacQueen 1967; Hartigan and Wong 1979; Kanungo et al. 2002] est une extension des *centres mobiles* [Forgy 1965]. C'est une méthode de classification automatique (*clustering*) qui regroupe les individus dans un nombre de classes prédéfini, par rapport à leurs distances avec les barycentres des classes.

L'algorithme se déroule en plusieurs étapes (cf. algorithme 6) :

1. On tire au hasard le barycentre de chacune des N classes (clusters).
2. On affecte chaque individu à la classe dont le barycentre est le plus proche en terme de distance
3. Après affectation de l'individu on recalcule le barycentre de la classe à laquelle on a affecté l'individu.
4. On recommence à l'étape 2 jusqu'à l'obtention d'une solution stable (tant que l'on modifie la classe d'appartenance des individus).

REMARQUES - *La construction même de l'algorithme permet de déduire immédiatement plusieurs propriétés :*

1. *Le partitionnement de l'espace est d'autant plus facile que sa dimension est élevée.*
2. *L'algorithme des k-moyennes fait intervenir une notion de distance. Or comme cela est expliqué dans l'annexe A.3, il existe différentes distances qui apportent différentes valeurs et qui engendrent par conséquent différents résultats. Dans notre travail, nous utilisons toujours la distance euclidienne qui est la distance la plus couramment utilisée.*
3. *Tirer aléatoirement les positions des barycentres lors de l'initialisation engendre des résultats différents (cf. figure 2.9).*

Cette dernière remarque peut poser parfois des problèmes, comme nous allons le voir.

Données : L'ensemble des individus I, le nombre de classes N, une distance d
Résultat : La classe d'appartenance de chaque individu et l'ensemble des classes C
Début
  Pour tous les $c \in C$ **Faire**
    Affecter aléatoirement le barycentre b_c ;
  Répéter
    Pour tous les $x \in I$ **Faire**
      Affecter x à la classe c tel que $\min_{c \in C} d(b_c, x)$;
      Recalculer b_c ;
  Jusqu'à *ce que l'on ne change plus d'individus de classe* ;
Fin

Algorithme 6 : K-moyennes (*K-means*).

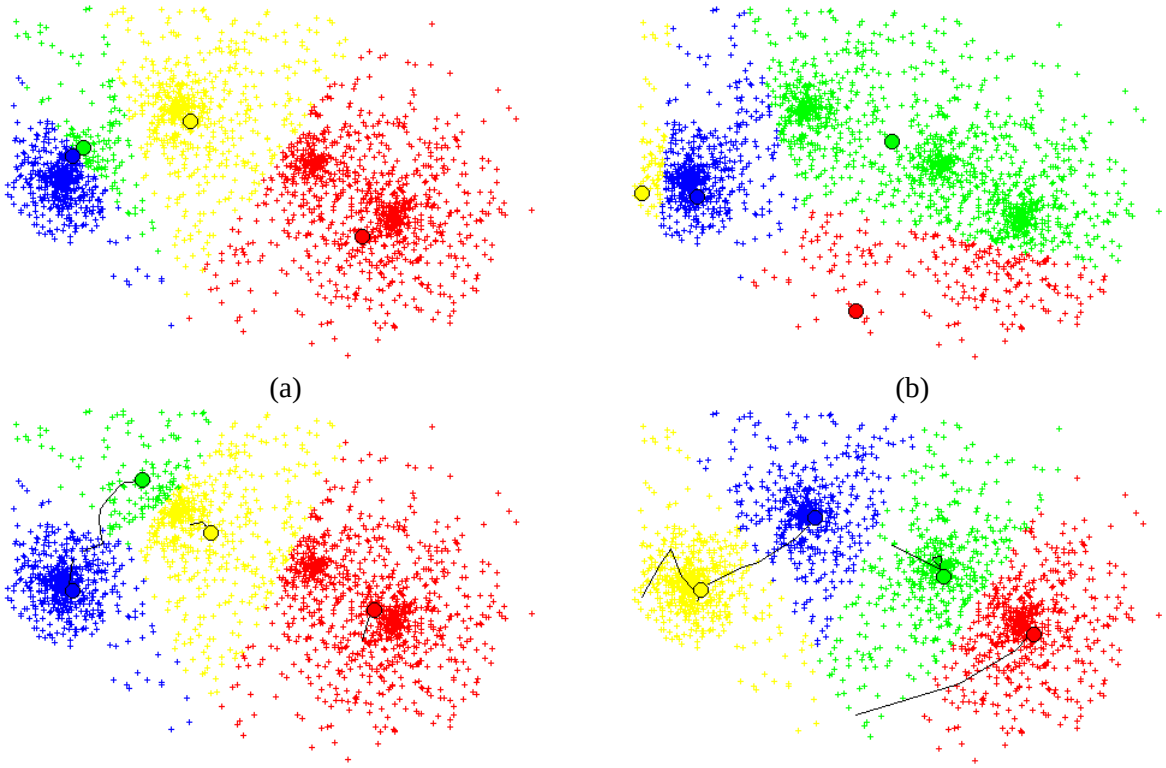


Figure 2.9. Illustration des différents résultats que produit la méthode des k-moyennes sur une même population en fonction de l'initialisation. (a) et (b) deux initialisations différentes, (c) et (d) les résultats différents engendrés. On peut remarquer que les barycentres et les classes sont totalement différents.

Les méthodes de classification sont parfois utilisées afin de trouver les individus les plus représentatifs d'une population, ce qui est le cas dans la partie II de ce travail relative à la caractérisation de la texture. En effet, on souhaite trouver un échantillon de N individus qui seraient les plus représentatifs d'une population. Pour cela, on utilise la méthode des k-moyennes afin de diviser la population en N classes. Pour chaque classe l'individu le plus proche du barycentre est l'individu le plus représentatif de la classe.

Définition 2.4.5 (Parangon) – *Le parangon est l'individu le plus représentatif d'une classe. Par définition, c'est celui qui est le plus proche du barycentre de la classe.*

Mais la remarque 3 implique que deux classifications consécutives engendrent deux résultats différents et donc deux échantillons de parangons différents. Une solution à ce problème est d'effectuer le calcul des *formes fortes* de la population.
Les formes fortes [Diday 1972; Diday et al. 1982] sont les groupes d'individus qui sont le plus souvent réunis dans la même classe lors de plusieurs classifications successives. Si l'on souhaite une classification à N classes, on commence par effectuer plusieurs classifications, puis on recherche les groupes d'individus les plus souvent ensemble comme le montre la figure 2.10. Cette opération nécessite d'énumérer toutes les combinaisons de cluster possible. On retient alors les N groupes d'individus les plus fréquents et on en calcule les barycentres. Ces N barycentres servent à initialiser une dernière classification dont on tire le résultat souhaité.

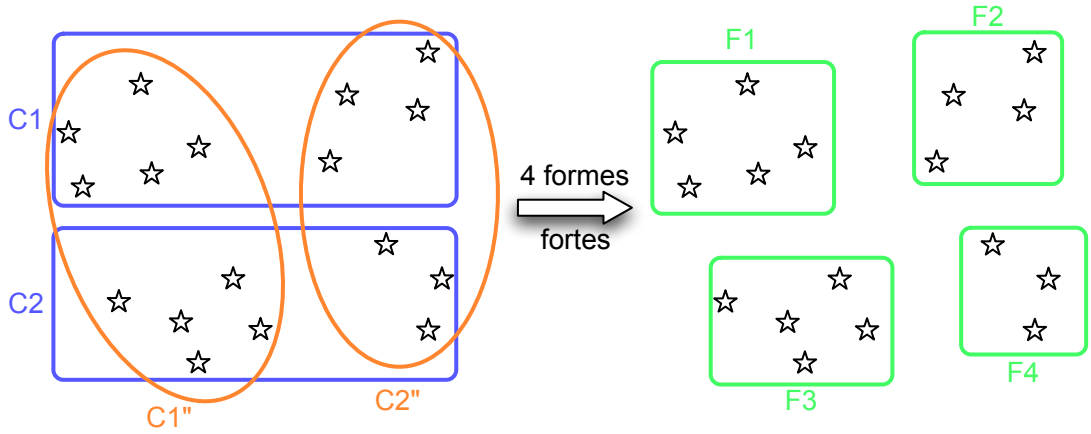


Figure 2.10. Illustration du calcul des formes fortes : deux classifications sont effectuées puis on recherche les individus les plus souvent classés ensemble.

Sa simplicité fait la popularité de cette approche. Autre avantage notable, il n'est pas nécessaire de calculer au préalable les distances deux à deux entre tous les individus, opération très gourmande en temps et en espace mémoire dans certains algorithmes de classification.
Mais la méthode des K-moyennes présente néanmoins des inconvénients. On cite notamment l'obligation de fixer au préalable le nombre de groupes, sans qu'aucune indication ne soit fournie. Toutefois certaines méthodes permettent de déterminer le nombre de cluster optimal [Li et al. 1999; Halkidi and Vazirgiannis 2001; KIM et al. 2001; Kima et al. 2004; Ammor et al. 2006; Ammor et al. 2008]. On peut aussi citer la forte dépendance de la méthode au choix des barycentres initiaux et la méthode peut converger vers un optimum local, mais ceci est résolu par le calcul des formes fortes.

CHAPITRE 3

LES DONNÉES

Si un des critères précédemment décrits dans la section 1.5 a une valeur anormale, alors le noyau est considéré comme pathologique. Donc chaque noyau pathologique appartient obligatoirement à au moins une de ces classes d'altération : forme boursouflée, texture non homogène, contient un nombre de focis anormal, contient un nombre de trous anormal ou possède un défaut de périphérie. Mais ces classes n'ont pas la même importance dans le diagnostic des noyaux. Afin de pouvoir les trier par ordre d'importance (qui est également l'ordre de priorité de l'étude de chaque classe), il est nécessaire d'analyser la répartition des noyaux pathologiques dans ces classes d'altération. Avant de réaliser l'expertise, il est nécessaire de segmenter les images afin d'extraire les noyaux.

3.1 Segmentation des noyaux

Nous disposons d'un ensemble de 270 images contenant des fibroblats (cf. section 1.2) provenant de sept patients différents : six d'entre eux sont atteints par la Progéria ; le septième est sain. Pour segmenter les images et extraire les noyaux, nous utilisons le procédé suivant :

1. Conversion de l'image en niveau de gris. Dans sa forme originale, l'image est un dégradé de vert.
2. Filtrage de l'image par transformée de Fourier rapide[1] [Chinga et al. 2007] (cf. figure 3.1*b*).
3. Seuillage par maximisation d'entropie [Pun 1980; Pun 1981; Kapur et al. 1985; Sahoo et al. 1988; Chang et al. 2006] (cf. figure 3.1*c*).
4. Etiquetage et calcul de la taille de toutes les composantes connexes présentes.
5. Suppression des composantes dont la taille est de l'ordre de quelques pixels.
6. Bouchage des trous présents dans les composantes. A cette étape on obtient le masque de segmentation.
7. Segmentation de l'image d'origine à l'aide du masque de segmentation (cf. figure 3.1*d*).
8. Séparation des composantes connexes résultantes qui sont les noyaux.

De ces images, nous avons segmenté et extrait environ 4000 noyaux de cellule, parmi lesquels 3300 sont exploitables. Les autres noyaux ont subi différentes altérations qui les rendent inutilisables pour les raisons suivantes :

– Coupés lors du cadrage de la photo (noyaux sur les bords de l'image).

[1] Nous utilisons le plugin disponible sur la plate-forme ImageJ (http ://rsb.info.nih.gov/ij/index.html) développée par le *National Institut of Health* : http ://rsb.info.nih.gov/ij/plugins/fft-filter.html.

– Noyaux déchirés par une mauvaise manipulation de la lame du microscope.
– Noyaux repliés sur eux-mêmes. Leurs formes et leurs textures sont déformées et partiellement cachées.

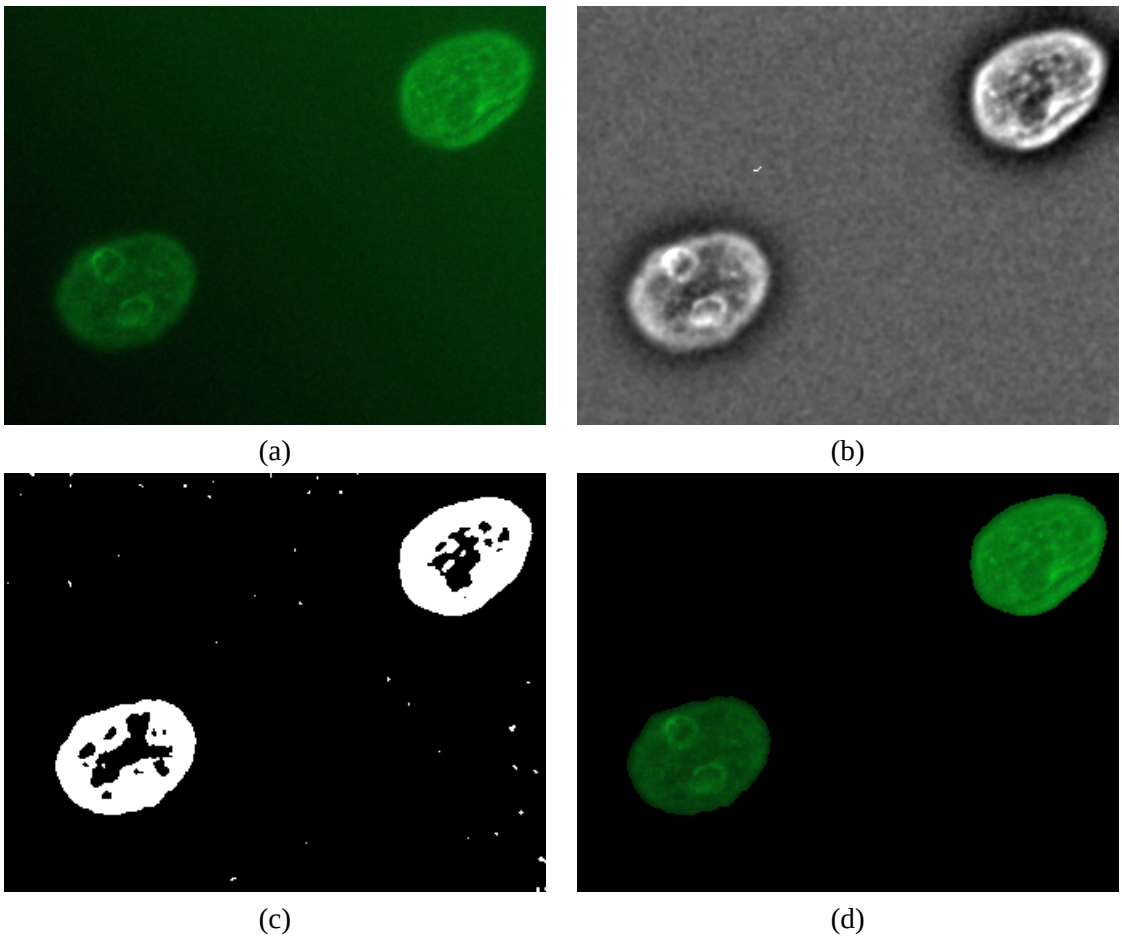


Figure 3.1. Illustration des différentes étapes de la segmentation des noyaux. (a) l'image originale, (b) l'image après filtrage par FFT, (c) l'image filtrée après seuillage par maximisation de l'entropie et (d) résultat de la segmentation.

REMARQUE - *Dans les chapitres 4 et 5 relatifs à la caractérisation de la forme des noyaux, les différentes méthodes caractérisent des formes binaires. Elles travaillent sur le masque de segmentation des noyaux.*
Dans le chapitre 9, nous souhaitons détecter les trous présents dans la texture des noyaux. Cependant lors de la sixième étape de segmentation, les trous de chaque composante sont bouchés : ces trous sont produits par un seuil trop élevé de la méthode de seuillage. En effet, on peut observer sur la figure 3.1a que les noyaux ne possèdent aucun trou.

NOTE - *Toutes les images ont été acquises avec la même échelle, donc un pixel représente toujours la même surface. Cette égalité permet d'utiliser le pixel comme unité de mesure tout au long de ce manuscrit.*

3.2 Distribution des noyaux dans les différentes classes

Ces 3300 noyaux ont été expertisés par les biologistes et généticiens de l'équipe de la Timone (cf. section 1.1) pour les différentes caractéristiques (*sain, pathologique, boursouflée, etc.*) qui viennent d'être citées (cf. section 1.5) et constituent l'échantillon de référence pour l'ensemble des travaux qui sont présentés dans ce manuscrit.
Dans cet échantillon, on peut observer la répartition suivante des noyaux :

1. 1024 noyaux sont considérés comme pathologiques.
2. 811 noyaux ont une forme boursouflée.
3. 135 noyaux ont une texture non homogène.
4. 340 noyaux possèdent des focis, parmi lesquels 123 peuvent être classés comme pathologiques en raison des focis qu'ils contiennent.
5. 181 noyaux contiennent des trous, dont 111 sont classés comme pathologiques en raison des trous qu'ils contiennent.
6. 86 noyaux ont une périphérie marquée mais non régulière.

Cette répartition permet d'observer les déséquilibres entre chaque classe. De tels déséquilibres sont problématiques pour la construction des modèles d'apprentissage : il est nécessaire d'en tenir compte dans ce travail.

3.3 Recouvrement entre les classes

Certains noyaux peuvent présenter plusieurs types d'altérations et il est intéressant d'étudier les recouvrements entre les classes (cf. table 3.1).

Classes	Forme	Texture	Focis	Trous	Périphérie	Pathologiques
Forme	-	10,35	4,68	7,02	5,05	100
Texture	62,22	-	11,85	14,81	5,18	100
Focis	46,34	16,26	-	8,13	6,5	100
Trous	34,23	14,41	0	-	6,5	100
Périphérie	47,67	8,13	9,3	9,3	-	100
Pathologiques	79,19	13,18	10,83	12,01	8,39	-

Table 3.1. Tableau du recouvrement entre les classes d'appartenance des noyaux pathologiques. Les valeurs données correspondent au pourcentage de noyaux de la classe de la ligne y appartenant aussi à la classe de la colonne x.

Le tableau 3.1 montre le pourcentage de recouvrement entre les différentes classes. La case (x,y) contient le pourcentage de noyaux de la classe de la ligne y qui appartiennent également à la classe de la colonne x.
La dernière ligne contient les recouvrements de chaque classe avec la classe "pathologique". La répartition des noyaux pathologiques dans les différentes classes indique l'ordre d'importance des

classes pour le diagnostic des noyaux.
On peut en déduire l'ordre de priorité suivant :

1. La forme.
2. L'homogénéité de la texture.
3. La présence de trous.
4. La présence de focis.
5. La régularité de la périphérie.

La forme est de loin l'élément de diagnostic le plus important puis vient l'homogénéité de la texture. Mais l'écart entre l'homogénéité de la texture et la présence de trous et de focis est très faible. Or lorsque l'on observe la ligne "texture", on peut remarquer que cette classe est intersectée de manière majoritaire avec la classe forme. Ce qui implique que la majorité des noyaux à texture non homogène est déjà classée comme étant pathologique dès l'analyse de leur forme. De manière plus précise, le classement des noyaux par leur texture apporte au mieux une amélioration du classement pour 31 noyaux (ce qui est de l'ordre du pourcent) si le classifieur de forme est efficace.
En revanche, le recouvrement entre les classes "forme pathologique" et "présence anormale de trous" ou "présence anormale de focis" est nettement inférieur. Donc même si la présence anormale de focis et de trous semble moins importante dans le diagnostic des noyaux, elle permet une amélioration bien supérieure à ce que peut apporter l'homogénéité de la texture.
Donc l'étude de la texture commence par une analyse globale (son homogénéité) avant de s'intéresser à des éléments plus spécifiques de son contenu (les trous et les focis).

3.4 Taux de répétabilité

Dans la section 2.3.4, la notion de *le taux de répétabilité* a été définie (concordance obtenue par les experts lors de deux classements successifs des individus). Ainsi pour un élément de diagnostic donné, le classifieur construit pour le modéliser doit obtenir un taux de prédiction au moins égal au taux de répétabilité des experts.
Pour chaque élément de diagnostic utilisé par les experts, le taux est le suivant :

- La forme : 92%.
- L'homogénéité de la texture : 85%.
- Les focis : 91%.
- Les trous : 90%.

PREMIÈRE PARTIE

CARACTÉRISATION ET CLASSEMENT DE LA FORME DES NOYAUX

CHAPITRE 4

MÉTHODES DE CARACTÉRISATIONS DE FORME

4.1 Introduction

L'analyse de l'expertise des noyaux dont nous disposons a mis en exergue l'importance de la *forme* dans le diagnostic (cf. section 3.2). A lui seul le critère de forme permet de diagnostiquer l'état de 87% des noyaux. La caractérisation et le classement des noyaux par leurs formes est donc l'étape la plus importante (premier sous-problème) dans le problème que nous souhaitons résoudre.

Dans le sous-problème de classement qui nous préoccupe, les méthodes que nous devons mettre en place doivent permettre de caractériser la forme afin de discriminer les formes différentes (inertie inter-classes élevée) et repérer les formes proches (inertie intra-classe faible). De plus, on a vu dans le chapitre 1 que des problèmes peuvent survenir lors de l'acquisition des noyaux et surtout lors de leur marquage par fluorescence. Du bruit peut être présent sur le contour ou sur la texture du noyau. Les méthodes que nous utilisons, doivent être peu sensibles aux déformations et aux bruits qui sont susceptibles de détériorer l'objet. De plus, les images d'acquisition contiennent parfois plusieurs dizaines de noyaux à classer. Bien que n'ayant pas de contrainte particulière concernant la complexité (coût de calcul) des algorithmes à employer, il est préférable d'utiliser des méthodes qui fournissent un résultat dans un temps raisonnable.

Des méthodes de reconnaissances des formes ont été développées pour répondre à un problème souvent spécifique et avec des contraintes tout aussi particulières. La littérature scientifique contient un grand nombre d'articles présentant ces méthodes qui peuvent être classées en deux grandes familles [Zhang and Lu 2004] :

- les méthodes basées sur l'étude du contour.
- les méthodes basées sur la caractérisation de la forme globale.

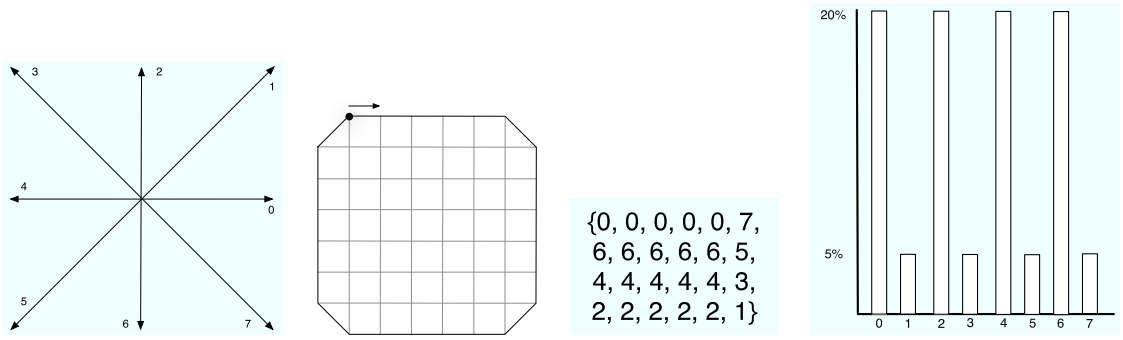
Différentes approches ayant été proposées dans la littérature, ce chapitre en décrit quelques-unes pour chacune des familles qui viennent d'être citées, tout en présentant les algorithmes correspondants. Ces méthodes sont testées en les appliquant au sous-problème auquel nous souhaitons répondre. Leurs avantages et inconvénients sont systématiquement présentés afin de valider ou d'invalider l'utilité de ces méthodes pour notre problème.

4.2 Caractérisation du contour

La première famille des méthodes de caractérisation de forme, concerne celles qui utilisent une description du contour de la forme. Cette section présente deux méthodes représentatives de cette famille.

4.2.1 La chaîne de Freeman

La chaîne de Freeman [Freeman 1961; Freeman 1974; Andriamampianinao et al. 1994; Bribiesca and Guzman 1980] est une des plus anciennes méthodes de description du contour. Si elle est moins employée en reconnaissance de forme qu'auparavant, elle reste encore utilisée dans des méthodes de transmission d'images comme le MPEG.
C'est une technique de représentation du contour basée sur les positions successives des pixels formant le contour ordonné de la forme : on fixe un point de départ, puis chaque point est exprimé par rapport au point précédent en 4- ou 8-connexités. Le résultat est une séquence ordonnée de n liens $\{c_i : i = 1,2\ldots n\}$, où c_i est un numéro de vecteur connectant deux pixels voisins du contour (figure 4.1-a). La direction de c_i est codée avec un entier $k \in [0, K-1]$ dans le sens des aiguilles d'une montre, avec $K = 2^{m+1}$ et $m = 1$ (4-connexité) ou $m = 2$ (8-connexité). La figure 4.1 illustre le calcul d'une chaîne de Freeman.
Dans [Iivarinen et al. 1997; Iivarinen and Visa 1996], les auteurs utilisent cette représentation pour calculer l'histogramme des valeurs c_i (*Chain Code Histogram*, *CCH*), qui est une fonction discrète obtenue comme ceci : $p(k) = \frac{n_k}{n}$, où n_k est le nombre de maillons de valeur k de la chaîne et n la taille de la chaîne totale (figure 4.1*d*).



(a) (b) (c) (d)

Figure 4.1. Exemple pour $K = 8$ (a), d'une forme (b), de sa chaîne de Freeman associée (c) et le CCH (d) engendré.

Le CCH apporte une représentation statistique et permet d'obtenir un codage invariant par rotation et homothétie (l'invariance par translation était déjà acquise). Il est par ailleurs possible de calculer une version normalisée de la CCH (*NCCH*, pour *Normalized CCH*), comme cela est utilisé dans [Iivarinen and Visa 1996]. Cette méthode est facile à mettre en œuvre et extrêmement rapide (temps réel), mais elle n'apporte pas une solution robuste en reconnaissance de forme. Cela provient essentiellement de sa grande sensibilité au bruit (particulièrement dans sa forme originale) et surtout parce que deux formes différentes peuvent avoir le même CCH.

REMARQUE - *La sensibilité au bruit est le principal inconvénient des méthodes de caractérisation du contour. Ce défaut se ressent particulièrement sur cette méthode. On peut aussi constater ce même défaut sur les profils de formes [Trier et al. 1996; Heutte et al. 1998; Tao et al. 2001; Soltanzadeh and Rahmati 2004] ou la représentation polaire du contour. Cette sensibilité provient de la description exacte du contour de la forme. Pour résoudre ce défaut, certaines méthodes comme les descripteurs de Fourier [Zahn and Roskies 1971; Zhang and Lu 2001; Zhang and Lu 2002; Chen and Kegl 2009; Aragon et al. 2007] "lissent" la frontière : cela permet d'éliminer les défauts de contour et offre une certaine souplesse dans le classement.*

4.2.2 Multiscale curve Smoothing for Generalised Pattern Recognition

La méthode *Multiscale curve Smoothing for Generalised Pattern Recognition* (MSGPR) [Kpalma and Ronsin 2003; Kpalma and Ronsin 2006] est une méthode de caractérisation du contour qui effectue l'intersection entre le contour de la forme et ce même contour ayant subi différents niveaux de lissage par filtrage Gaussien.
Elle nécessite préalablement un pré-traitement composé de deux étapes (figure 4.2), afin que le calcul de la courbe paramétrique qui est nécessaire plus tard soit invariant par rotation :

1. calcul du barycentre et de l'axe principal de la forme (défini dans la section A.6).
2. une rotation qui a pour centre le barycentre de la forme, dans le sens des aiguilles d'une montre, afin que l'axe principal soit confondu avec l'axe des X (figure 4.2).

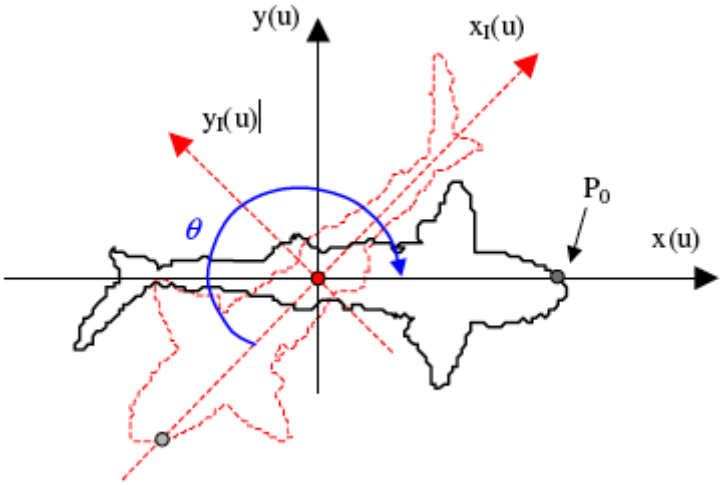


Figure 4.2. Traitement préliminaire sur la forme afin de rendre la méthode invariante par rotation (illustration issue de [Kpalma and Ronsin 2006]).

Une fois le pré-traitement effectué, cette méthode se décompose en quatre étapes qui sont illustrées par la figure 4.3.

1. Le contour de la forme est transcrit en fonction paramétrique $Contour(u) = (x(u), y(u)), u \in [0, 2\pi]$ (figure 4.4).
2. Application d'un filtre passe-bas de type Gaussien sur chacune des composantes de la fonction : $g(\sigma, u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}}$, avec σ la déviation standard du noyau Gaussien.

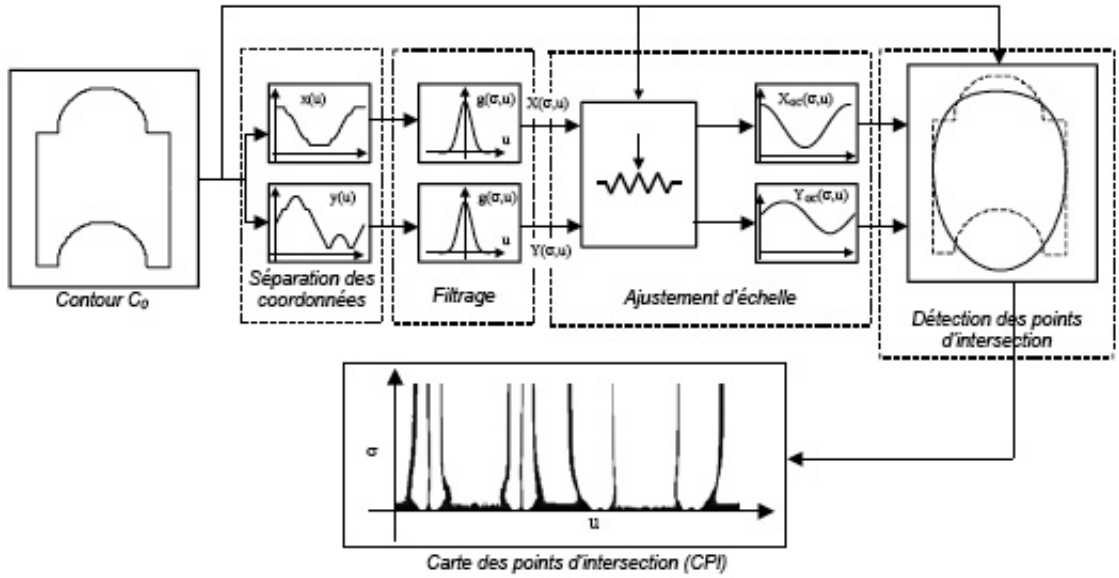


Figure 4.3. Les différentes étapes de la méthode MSGPR (illustration issue de [Kpalma and Ronsin 2006]).

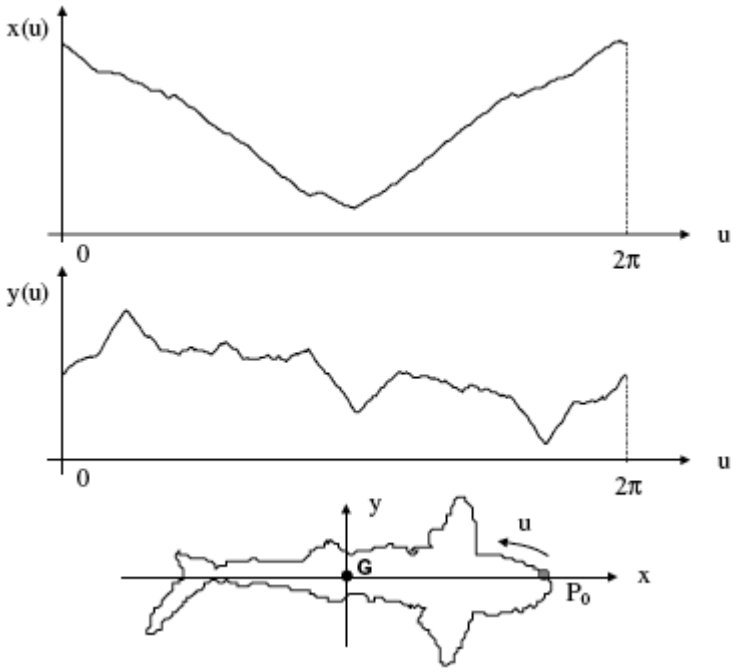


Figure 4.4. Transcription du contour en fonction paramétrique (illustration issue de [Kpalma and Ronsin 2006]).

3. Les deux fonctions lissées sont recomposées pour toutes les valeurs de $\sigma \in [0, 180]$.
4. Calcul de la fonction *Intersection Points Map* (*IPM*) entre la courbe finale et le contour initial

de la manière suivante (résultats figure 4.5) :

$$IPM(u,\sigma) = \begin{cases} 1\ (noir)\ si\ (x(u),y(u))\ est\ un\ point\ d'intersection. \\ 0\ (blanc)\ sinon. \end{cases}$$

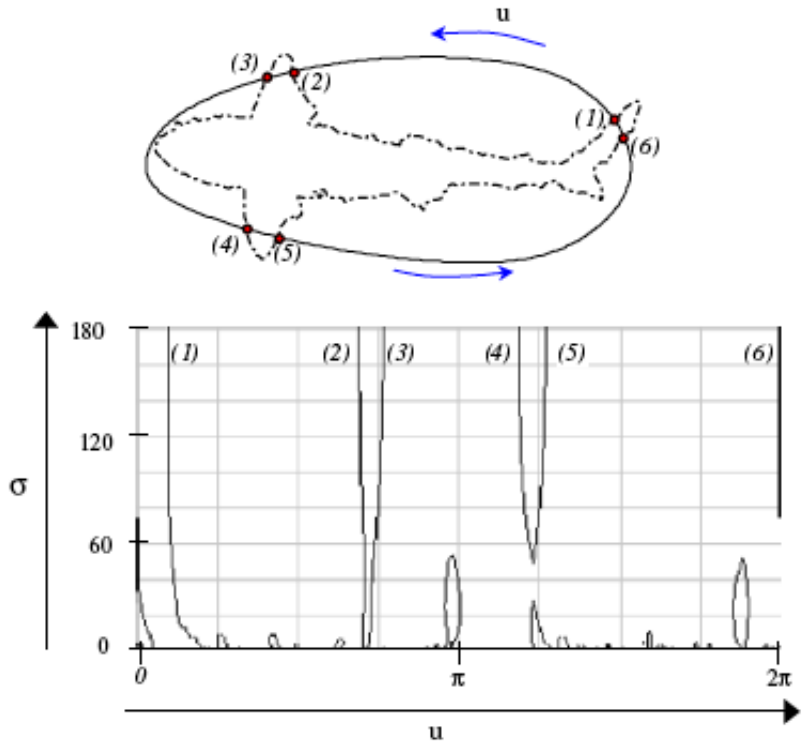


Figure 4.5. Exemple et résultat du calcul de la fonction *IPM* (illustration issue de [Kpalma and Ronsin 2006]).

Les écarts entre les différentes lignes verticales représentent les caractéristiques de l'objet analysé (figure 4.5). Ils sont ainsi comparés à la base de données des caractéristiques des objets référencés afin de classer la forme.
Les avantages majeurs de cette méthode de caractérisation sont : sa robustesse pour n'importe quel type de forme et son insensibilité aux variations de types translations, rotations, homothéties et au bruit. En revanche, sa mise en œuvre est lourde et elle ne peut pas s'appliquer en temps réel. Sa structure ne lui permet pas de caractériser puis d'identifier des formes contenant des parties occultées car il est nécessaire de paramétrer la totalité du contour. Mais lorsque le contour est visible dans sa totalité, c'est une méthode robuste de *shape matching*. De plus, la grande dimension du vecteur résultat (graphique) pose des problèmes spécifiques.

4.3 Caractérisation globale de la forme

Deux méthodes de caractérisation de forme par description et analyse du contour viennent d'être présentées. L'utilisation de ces méthodes pose des problèmes spécifiques : sensibilité au bruit et résultat identique pour des formes différentes ou dimension importante du vecteur caractéristique. Nous présentons donc deux méthodes de caractérisation par analyse de la forme dans sa globalité. Les méthodes de caractérisation globales de type squelette [Blum 1964; Heutte et al. 1998; Remy 2001; Mari 2002; Kégl and Krzyzak 2002; Remy and Thiel 2002; Lorigo and Govindaraju 2006] ne sont pas appropriées à notre problème car la forme géométriquement simple des noyaux fait que les déformations à caractériser sont faibles. Elles ne sont ni présentées ni étudiées dans ce travail.

4.3.1 Signature polaire

La signature polaire est une méthode de caractérisation basée sur l'intersection de la forme avec une série de cercles de rayons différents et centrés sur le barycentre de la forme. Elle se décompose en trois étapes :

1. Choix judicieux du nombre de cercles et de leurs rayons respectifs (généralement une partition du rayon maximum).
2. Calcul du barycentre de la forme.
3. Remplissage de la fonction résultat (figure 4.6) de la manière suivante :

$$S(R,\theta) = \begin{cases} 1 \text{ si } I(R,\theta) \in F \\ 0 \text{ sinon} \end{cases}, \ \theta \in [0, 2\pi[$$

Ce qui donne l'algorithme suivant :

Données : Image I contenant la forme F, le nombre de cercles N et leurs rayons R_c
Résultat : Un tableau S contenant les résultats binaires des intersections
Début
  Calcul du barycentre de la forme ;
  Pour $c \leftarrow 1$ **à** N **Faire**
    Pour $\theta \leftarrow 0$ **à** 359 **Faire**
      Si $I(R_c, \theta) \in F$ **Alors**
        $S[c, \theta] \leftarrow 1$;
      Sinon
        $S[c, \theta] \leftarrow 0$;
Fin

Algorithme 7 : Signature polaire.

Cette technique apporte une représentation caractéristique (une signature) pour tous types de formes ou de volumes (possibilité de l'étendre en 3D). Elle est robuste et invariante pour toutes les transformations :

- Translation, car elle est centrée sur le barycentre.
- Homothétie, si le nombre de cercles est préalablement fixé et si les rayons dépendent de la taille de la forme.
- Rotation, car les fonctions signatures (résultats) sont 2π-périodiques (cf. figure 4.7).

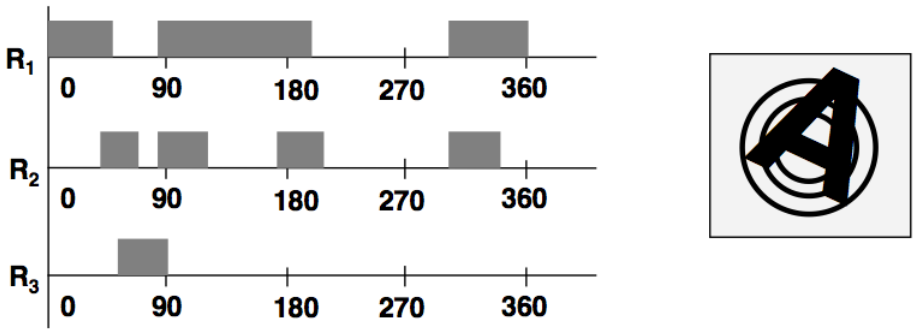


Figure 4.6. Exemple de signature polaire pour le caractère A, avec trois cercles non uniformément répartis.

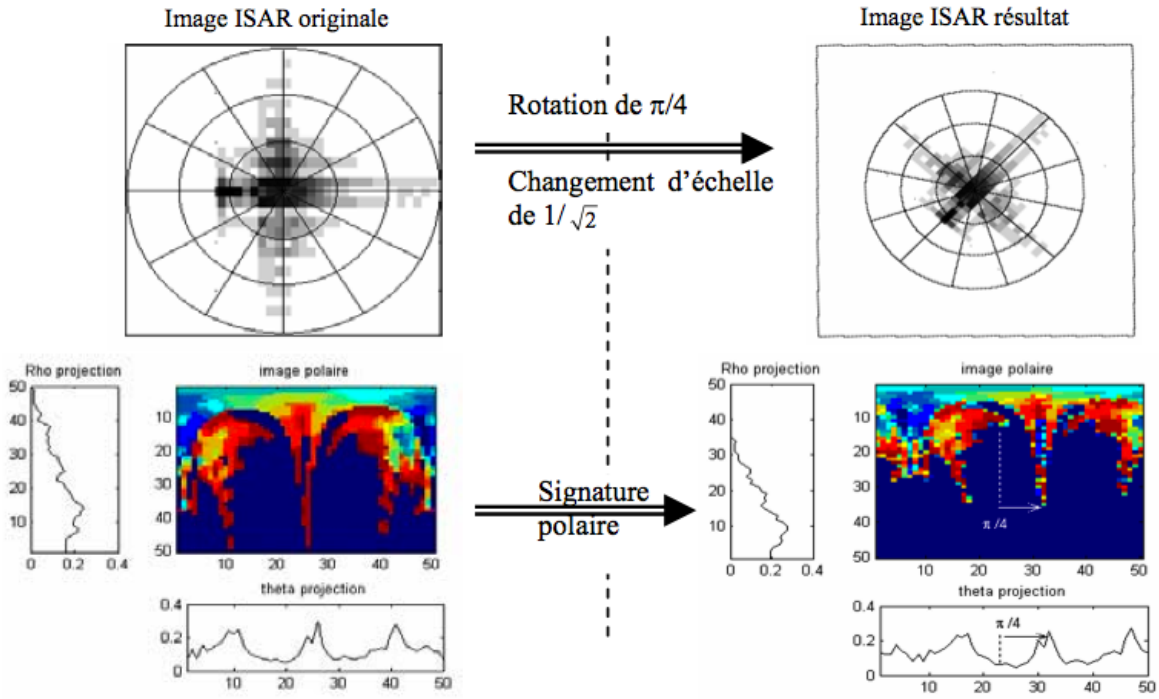


Figure 4.7. Un exemple d'invariance par rotation et homothétie de la signature polaire.

Ses propriétés et sa robustesse permettent de l'employer dans de nombreuses applications en reconnaissances de forme [Collewet 1999], imagerie médicale [Benjelloun et al. 2006], télédétection [Toumi et al. 2006], reconnaissance de caractères, etc. Cette méthode est donc testée pour le classement de la forme des noyaux.
Pour N cercles, la fonction résultat est N ensembles ordonnés de 360 valeurs (une par degré) binaires (0 si on est en dehors de la forme, 1 sinon). Le vecteur caractéristique comporte alors $360 \times N$ variables, ce qui est un nombre excessivement grand qui risque d'engendrer un problème d'apprentissage par cœur. Il est possible de contourner ce problème :

- en diminuant la taille de la fonction solution ; on ne considère que M valeurs régulièrement réparties sur le cercle (par pas de $360/M$) parmi les 360 possibles.
- pour chaque cercle, additionner les 360 valeurs ; c'est-à-dire que l'on compte le nombre de

points de chaque cercle qui appartient à la forme.

Nous appliquons ces deux méthodes afin de les tester et de garder celle qui classe le mieux la forme. Pour cela, nous utilisons l'algorithme suivant :

Données : Image I contenant la forme F, le nombre de cercles N et leurs rayons R_c
Résultat : Un vecteur caractéristique V contenant les résultats binaires de chaque cercle ou leurs sommes
Début
  Calcul du barycentre de la forme ;
  Calcul du l'axe principal et de l'angle θ qu'il fait avec l'axe des X ;
  Rotation de I dans le sens des aiguilles d'angle $-\theta$ centré sur le barycentre ;
  Pour $c \leftarrow 1$ **à** N **Faire**
    Pour $\theta \leftarrow 0$ **à** 359 **Faire**
      Si $I(R_c, \theta) \in F$ **Alors**
        $T[c, \theta] \leftarrow 1$;
      Sinon
        $T[c, \theta] \leftarrow 0$;
  Si *utilisation directe des résultats binaires de* T **Alors**
    Copie de T dans V ;
  Sinon
    Somme de chaque ligne de T ;
    Copie de la somme des V ;
Fin

Algorithme 8 : Utilisation de la signature polaire en classification.

Le meilleur résultat est obtenu en sommant les fonctions résultats, avec 16 cercles non uniformément répartis. En effet, la forme des noyaux étant *pleine*, les cercles ayant un petit rayon apportent une solution toujours égale à 1. Les cercles sont concentrés vers les bords de la forme en prenant comme longueur du premier rayon la longueur du plus petit rayon de la forme (cf. définition B.1.4 en annexe). Malheureusement ce procédé parvient à classer correctement au mieux 75% des noyaux par régression logistique et validation croisée. Il ne permet pas d'apporter de solution satisfaisante au problème de classement de la forme. La principale raison de cet échec est la difficulté à analyser le contour à l'aide de cette méthode.

4.3.2 Histogrammes de projections

La technique des histogrammes de projections [Cakmakov et al. 2002; Lorigo and Govindaraju 2006; Soltanzadeh and Rahmati 2004; Tao et al. 2001] est une méthode de caractérisation globale d'une forme qui renseigne sur son épaisseur dans plusieurs directions. Chaque histogramme est calculé en comptant le nombre de pixels de la forme dans une direction δ : $HP(\delta) = \sum_F I_\delta(x,y)$. Cela revient à *projeter* les pixels de la forme dans une direction et à regarder les variations de la distribution marginale. Pour des objets en deux dimensions, on peut choisir quatre directions de projections : horizontale, verticale et deux diagonales. L'algorithme 9 décrit les étapes de la construction des histogrammes et la figure 4.8 montre un exemple de résultats d'histogrammes des quatre projections.

Données : Image binaire I contenant la forme F
Résultat : Deux tableaux contenant les histogrammes de projections horizontaux et verticaux
Début
  Initialiser les tableaux $Horizontal$ et $Vertical$ à 0 ;
  Pour tous les *points* $p(x,y) \in F$ **Faire**
    Incrémenter($Horizontal[y]$) ;
    Incrémenter($Vertical[x]$) ;
Fin

Algorithme 9 : Calcul des histogrammes de projections dans les directions horizontales et verticales.

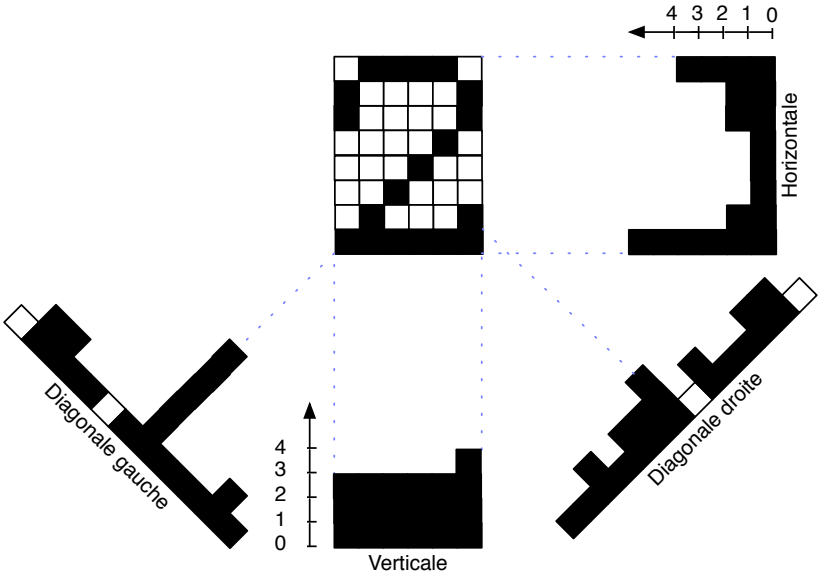


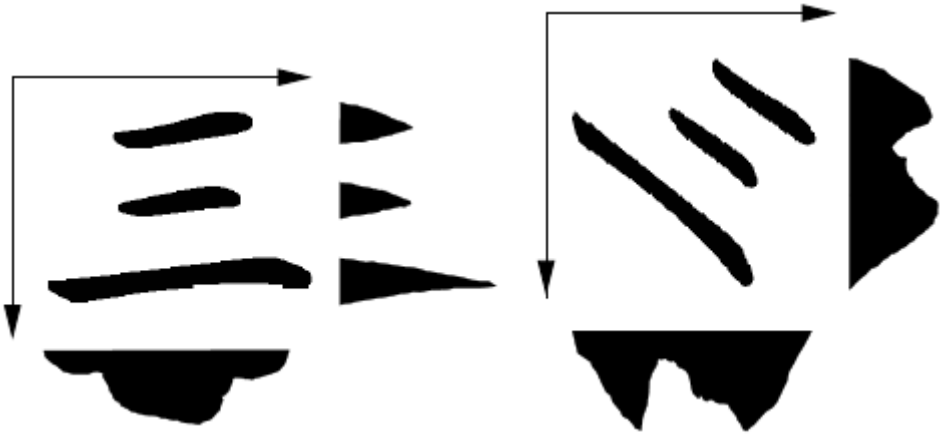
Figure 4.8. Exemples d'histogrammes de projections horizontaux, verticaux et diagonaux pour le chiffre 2.

Cette technique est très répandue en reconnaissance des caractères, en particulier pour les caractères d'imprimerie. Pour certaines lettres de l'alphabet d'imprimerie (par exemple *I,L,T*), les histogrammes horizontaux et/ou verticaux sont identiques ou symétriques à la forme originale.
Elle est insensible aux variations de types :

- translation, la projection est translatée mais les valeurs inchangées.
- homothétie, il y a un rapport constant entre les projections et il est possible de normer les résultats ce qui résout le problème.

En revanche, elle n'est pas invariante par rotation (figure 4.9) et ce cas se produit régulièrement lors de la reconnaissance de caractères manuscrits. Toutefois, ce problème est résolu en effectuant une rotation préliminaire suivant l'axe principal (cf. figure 4.2).

Un autre inconvénient de cette méthode est son absence d'indication sur les éventuels trous se trouvant dans la forme. En effet, un trou se traduit par un minimum sur un ou plusieurs histogrammes, mais ce minimum pourrait tout aussi bien correspondre à une concavité sur le bord de la forme. Ce défaut peut être contourné en combinant les histogrammes de projections avec la méthode du *nombre de segments* (*CCV* pour *Crossing Counts Vector* [Heutte et al. 1998; Soltanza-



(a) (b)

Figure 4.9. Exemple de non invariance par rotation des histogrammes de projections (illustration issue de [Tao et al. 2001]). (a) L'image originale avec ses deux histogrammes "horizontal" et "vertical", (b) l'image ayant subi une rotation d'angle $-\pi/4$ avec les deux histogrammes associés.

deh and Rahmati 2004]). Cette méthode compte le nombre de segments de pixels appartenant à la forme qui sont rencontrés dans une direction donnée. Donc si le nombre de segments dans toutes les directions est égal à 0 ou 1, cela implique que la forme étudiée est convexe et pleine, sinon elle comporte des points de concavité ou des trous (figure 4.10).

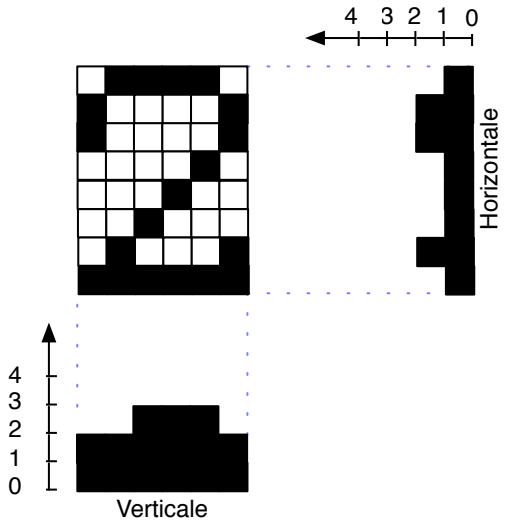


Figure 4.10. Exemples de résultats de la méthode CCV appliquée aux chiffres 2 de la figure 4.8.

La CCV est invariante par translation, rotation et homothétie, mais elle nécessite un débruitage de l'image, sinon tout pixel parasite rencontré est considéré comme un vecteur supplémentaire et altère le résultat.

Il existe un dernier type de projection insensible aux problèmes de transformations, qui projette non pas dans une direction mais vers un point (le barycentre) : la *projection centrale* [Tao et al.

2001] (*CPT* pour *Central Projection Transformation*). Tous les pixels de la forme sont projetés vers le barycentre qui se comporte comme un centre d'attraction en attirant tous les pixels (figure 4.11*b*).

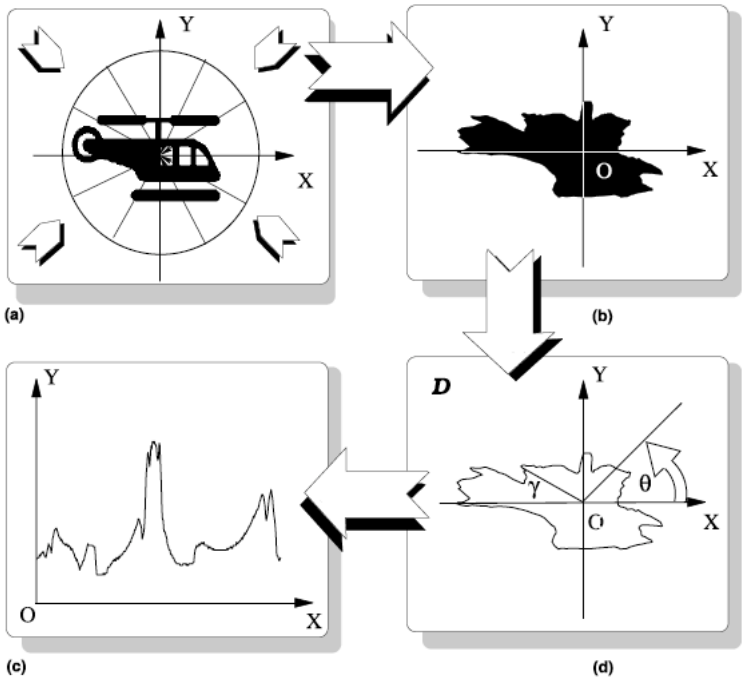


Figure 4.11. Illustration de la projection centrale (issue de [Tao et al. 2001]) : (a) l'image d'origine, (b) la projection centrale, (d) extraction du contour et (c) paramétrisation du contour.

Une paramétrisation du contour est ajoutée après le calcul de la CPT (figure 4.11 c et d). La conjugaison de ces deux transformations permet d'obtenir une courbe paramétrique 2π-périodique qui apporte une invariance par rotation robuste (cf. figure 4.12).
Par la suite, les auteurs décomposent la courbe paramétrique en ondelettes afin d'utiliser la décomposition comme caractéristique dans un classificateur de type "classifieur par distance euclidienne pondérée" (*The weighted euclidean distance classifier*).

Dans notre problème, les noyaux sont pleins (cf. section 3.1), il est par conséquent inutile d'utiliser la CCV. Nous employons les histogrammes de projections dans leurs formes classiques en utilisant les projections horizontales et verticales. Toutefois, il est nécessaire de *préparer* les noyaux afin d'obtenir les invariances par translation, rotation et homothétie. Pour cela, chaque noyau subit tout d'abord une rotation afin de confondre son axe principal avec l'axe des X (cf. figure 4.2), puis une transformation d'échelle afin que la boite englobante soit de dimensions $N \times N$. Ainsi les histogrammes horizontaux et verticaux sont de taille N et apportent $2N$ caractéristiques.

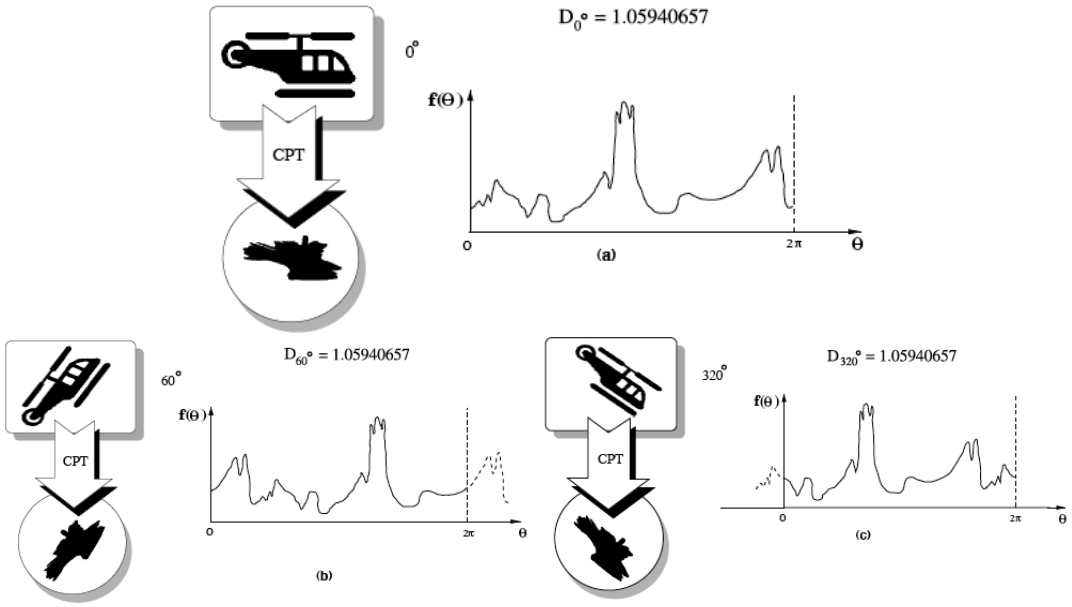


Figure 4.12. Exemple d'invariance par rotation pour la CPT (illustration issue de [Tao et al. 2001]). (a) l'image d'origine, la CPT associée ainsi que la courbe paramétrique extraite. (b) (resp. (c)) l'image ayant subi une rotation de 60˚ (resp. 320˚) avec la nouvelle CPT associée et la nouvelle courbe paramétrique extraite.

Données : Image binaire I contenant la forme F du noyau, le nombre N de caractéristiques
Résultat : Un vecteur caractéristique contenant les informations des histogrammes de projections horizontaux et verticaux

Début
- Initialiser les tableaux $Horizontal$ et $Vertical$ à 0 ;
- Trouver l'axe principal et tourner le noyau ;
- Transformer le noyau pour qu'il soit de dimension $N \times N$;
- **Pour tous les** *points* $p(x,y) \in F$ **Faire**
  - Incrémenter($Horizontal[y]$) ;
  - Incrémenter($Vertical[x]$) ;
- Copier les valeurs de $Horizontal$ et $Vertical$ dans le vecteur caractéristique ;

Fin

Algorithme 10 : Utilisation des histogrammes de projections pour la classification.

Le meilleur résultat est de 83% (par régression logistique et validation croisée) pour un nombre de caractéristiques égal à 32 (noyaux de dimensions 16×16, donc 16 caractéristiques verticales et 16 horizontales). Bien que supérieur au résultat précédemment obtenu, ce pourcentage n'est pas suffisant (car inférieur au taux de répétabiité des experts) et la méthode ne peut donc pas être utilisée pour classer la forme des noyaux.

4.4 Conclusion

Nous venons de présenter un état de l'art général des méthodes de caractérisations de forme. La première famille de méthodes basées sur la description du contour connaît de nombreux défauts qui la rendent difficilement utilisable dans un classifieur. En revanche, la deuxième famille de méthodes comprenant les techniques de caractérisation globale d'une forme est mieux adaptée pour la classification. Nous avons présenté deux de ces méthodes, mais elles ne se sont pas révélées pertinentes ou efficaces pour résoudre notre sous-problème de classement de la forme des noyaux.
Il est donc nécessaire d'utiliser une autre méthode afin de résoudre le sous-problème. Cette méthode doit permettre de classer correctement les noyaux, donc elle doit être facilement utilisable dans un classifieur et permettre une caractérisation robuste de la forme aussi bien dans sa globalité que dans la description du contour.

CHAPITRE 5

MESURES ET INDICES DE FORMES

5.1 Définitions et propriétés

Les indices de forme ont été présentés pour la première fois par Santalo [Santalo 1976] dans un ouvrage relatif aux propriétés mathématiques des formes convexes. On trouve la définition et les propriétés des indices de forme dans [Coster and Chermant 1985; Fillère 1995].

Définition 5.1.1 (Indice de forme) – *On appelle indice de forme tout paramètre, cœfficient ou combinaison de cœfficients permettant de donner des renseignements chiffrés sur la forme.*
De plus, les indices doivent avoir les propriétés suivantes :

1. *Etre sans dimension.*
2. *Etre invariant par homothétie.*
3. *Etre invariant par rotation et translation.*

REMARQUE - *Dans [Coster and Chermant 1985], on trouve également une quatrième propriété : "S'appliquer à des formes connexes simples donc homéomorphes au disque". Mais dans de nombreux articles plus récents [Castanon et al. 2007; Stojmenović et al. 2006; Perner et al. 2002; Liu et al. 2006a; Rosin 2004; Street et al. 1993; Soltanian Zadeh et al. 2004; Thiran and Macq 1996; Zunic and Rosin 2004], les auteurs utilisent les indices de forme pour tous types de formes grâce notamment à leur utilisation avec des méthodes par apprentissage. Ces articles récents tendraient à prouver que cette propriété n'est plus d'actualité.*

Le calcul d'un indice de formes est équivalent au calcul de la valeur d'une fonction à plusieurs variables que l'on nomme *mesures*.

Définition 5.1.2 (Mesure) – *On appelle "mesure" d'une forme toute valeur ou ensemble de valeurs numériques "mesurées" sur la forme.*

REMARQUES - *Les mesures peuvent avoir des dimensions :*
- *trois dimensions (le volume)*
- *deux dimensions (la surface)*
- *une dimension (le périmètre, le diamètre euclidien, la longueur de l'axe principal, la longueur du diamètre géodésique, etc.)*
- *aucune dimension (le nombre de composantes connexes, le nombre de trous, etc.)*

Les mesures sans dimension satisfont la définition 5.1.1 et sont par conséquent des indices de forme.

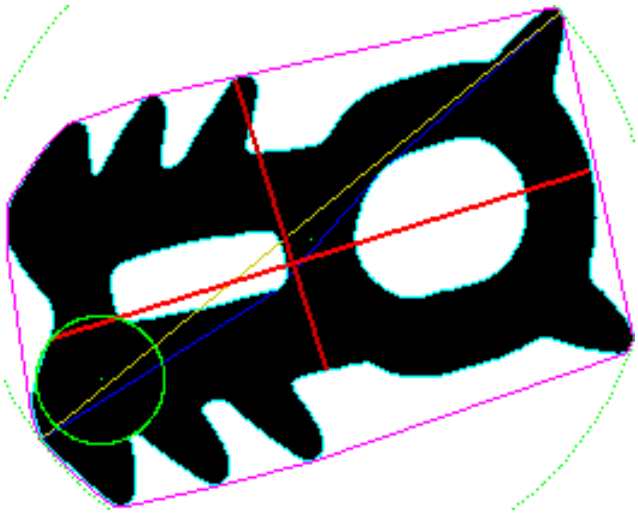


Figure 5.1. Exemples de mesures : surface (noir), périmètre (cyan), axes principaux (rouge), enveloppe convexe (violet), diamètre géodésique (bleu) et diamètre euclidien (jaune), plus petite (resp. grande) boule circonscrite (resp. inscrite) (vert).

L'extraction des mesures représente l'étape la plus importante dans le calcul d'un indice de forme, car de la valeur des mesures dépend la valeur de l'indice. Le temps de calcul et le comportement d'un indice dépendent du temps de calcul et du comportement des mesures qui le composent. Par exemple, le périmètre ne possède pas de bonne propriété de continuité [Coster and Chermant 1985] et le plus petit rayon est sensible au bruit[1], ce qui affecte de la même façon tout indice utilisant ces mesures.
La liste complète des mesures et des indices utilisés dans notre travail est en annexe B.1 et B.2 de ce document. Mais certaines mesures bien connues comme le diamètre de Ferêt [Tuset et al. 2003] ou le rayon de courbure maximum (resp. minimum) du contour ne sont pas utilisées. Ces mesures sont absentes de ce document, soit parce que leur extension en trois dimensions est difficile soit parce que leur temps de calcul est trop important (complexité élevée).

Par construction, les indices de forme se calculent comme une fonction à plusieurs variables. On aimerait donc retrouver certaines propriétés des fonctions et notamment la bijection (section A.4). Mais on constate qu'il n'y a pas de bijection entre l'espace des formes et celui des indices. Les indices de convexité et d'allongement par le diamètre apportent des contre-exemples :

$$\forall F \text{ convexe}, \; Convexité_{Périmétrique}(F) = Convexité_{Surfacique}(F) = 1$$

$$Allongement_{Diamètre}(Carré) = Allongement_{Diamètre}(Disque) = \frac{1}{2}$$

Donc les indices de forme ne sont pas des fonctions bijectives car elles ne sont pas injectives (définition A.4.2). En revanche, un indice de forme permet de différencier deux formes.

Propriété 5.1.3 (Indice de forme) – *Soit F_1 et F_2 deux formes quelconques, s'il existe un indice de forme I tel que $I(F_1) \neq I(F_2)$ alors $F_1 \neq F_2$.*
Ce qui peut s'écrire : $\forall F_1, F_2$ *si* $\exists I / I(F_1) \neq I(F_2) \Rightarrow F_1 \neq F_2$

[1] On parle ici du bruit de type *poivre et sel*.

Le principal avantage des indices de forme est leur grande souplesse. En effet, il est aisé de construire de nouveaux indices en fonction du problème que l'on souhaite traiter. Ces nouveaux indices spécifiques auront ainsi une grande capacité de description et permettront un meilleur classement. De plus, chaque indice apporte une valeur ou un ensemble de valeurs qui sont directement utilisables dans un classifieur.

5.2 Quatre nouveaux indices

La souplesse des indices de forme est un avantage majeur qu'il faut exploiter. Nous utilisons cette propriété afin de construire trois nouveaux indices spécifiques à la caractérisation des noyaux de cellules.

5.2.1 Indices de caractérisation d'une ellipse

Les noyaux sains ont une forme allongée, régulière et quasi elliptique (cf. figure 5.2). C'est la forme résultante de la déformation d'une forme ellipsoïdale (forme naturelle d'un noyau) aplatie (lors de l'observation sur une lame de microscope) et observée en 2D. Nous en avons déduit que construire des indices qui caractérisent des ellipses permettrait d'avoir des indices qui caractérisent les noyaux avec une forme normale.

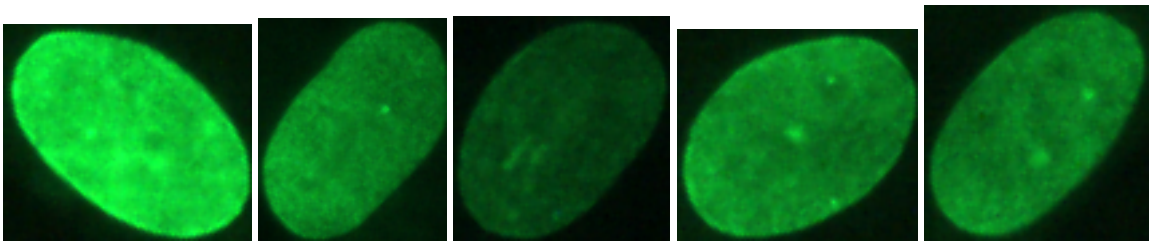
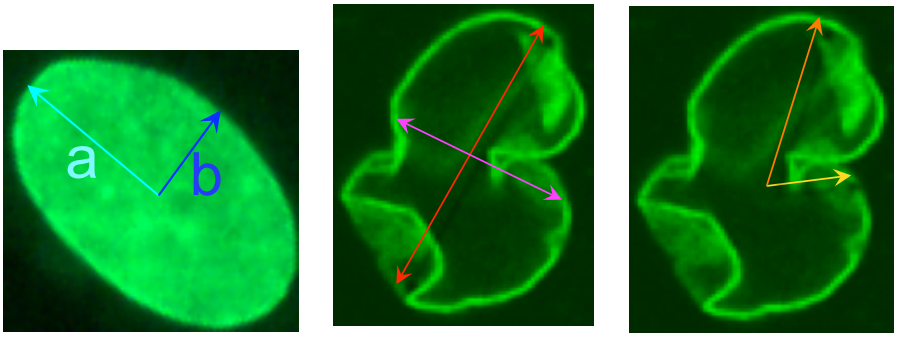


Figure 5.2. Exemples de noyaux possédant une forme normale. On peut aisément constater que la forme s'apparente à celle d'une ellipse.

Les mesures les plus simples à obtenir sur une forme sont le périmètre et l'aire. Nous nous intéressons à l'aire d'une ellipse qui se calcule de la façon suivante : $A = \pi ab$, avec a le demi grand axe et b le demi petit axe (cf. figure 5.3).
Dans une ellipse on constate que :

- Le grand axe est confondu avec l'axe principal et le diamètre.
- Le demi grand axe est égal au plus grand rayon.
- Le petit axe est porté par l'axe secondaire (cf. section A.6).
- Le demi petit axe est égal au plus petit rayon et à l'épaisseur issue du diamètre.

Nous pouvons en déduire les égalités suivantes : $a = \frac{1}{2}L_{AP} = \frac{1}{2}D = R_{max}$ et $b = \frac{1}{2}L_{AP\perp} = E_D = R_{min}$ (cf. figure 5.3).



(a) (b) (c)

Figure 5.3. (a) Illustration des demi axes sur un noyau avec une forme normale. (b) Axe principal (en rouge) et axe secondaire (en violet). (c) Plus grand rayon (orange) et plus petit rayon (jaune).

Ces inégalités permettent de construire trois nouveaux indices permettant de caractériser des ellipses, par :

Les rayons

$$\Psi_{EllipseR} = \pi \frac{R_{min} R_{max}}{A} \in [0,1]$$

L'axe principal

$$\Psi_{EllipseAP} = \frac{\pi}{4} \frac{L_{AP} L_{AP\perp}}{A} \in [0,1]$$

Le diamètre

$$\Psi_{EllipseD} = \frac{\pi}{2} \frac{E_D D}{A} \in [0,1]$$

L'indice d'ellipse par les rayons caractérise une ellipse en utilisant des mesures dépendantes du contour. En revanche, l'indice d'ellipse par l'axe principal utilise des mesures qui prennent en compte la totalité des points du noyau. Par construction, ces indices valent 1 pour des ellipses. Les intervalles d'appartenance des indices sont calculés pour des formes convexes variant du segment au disque [Coster and Chermant 1985].

5.2.2 Indice de caractérisation de la convexité

Nous venons de construire trois indices qui caractérisent des noyaux possédant une forme normale. Il serait maintenant intéressant d'élaborer un indice pour caractériser les noyaux ayant une forme anormale (*boursouflée*).
Le critère de décision principal dans le diagnostic de la forme des noyaux est la convexité. En effet, les noyaux boursouflés possèdent des zones de concavité en taille et en nombre différents. Pour compter ces zones de concavité, il est possible de calculer le *nombre de composantes connexes d'écart* N_{Cce} issues de la soustraction de la forme à son enveloppe convexe $N_{Cce} = card(C_H(F) \backslash F)$.

Pour construire cet indice, nous utilisons la forme normée de la mesure N_{Cce} :

$$\psi_{N_{Cce}} = \frac{1}{1+N_{Cce}} \in]0,1]$$

Figure 5.4. Illustration du calcul de la mesure N_{Cce}. On compte le nombre de composantes connexes d'écarts (le nombre de composantes en violet).

Cet indice vaut 1 si la forme est convexe car aucune composante d'écart n'est trouvée et plus la forme possède des composantes d'écart, plus l'indice tend vers 0.

Mais dans la pratique, on ne peut considérer comme composante connexe d'écart, des composantes dont la taille est de l'ordre du pixel et qui sont dues à des imprécisions de discrétisation. De plus, dans les éléments de diagnostic des noyaux, la taille et le nombre des composantes connexes doivent être pris en compte. Donc la mesure N_{Cce} nécessite un calibrage. A partir de l'expertise de la forme des noyaux, il faut trouver un ou plusieurs seuils concernant la taille et le nombre des composantes qui permettent de décider si une composante connexe d'écart doit être comptabilisée. Pour cela, nous avons réalisé une étude systématique du pourcentage de bon classement (obtenu en utilisant uniquement l'indice $\psi_{N_{Cce}}$) en fonction de la taille et du nombre de composantes connexes d'écarts (figure 5.5). Cette étude consiste à calculer systématiquement le pourcentage de bon classement en faisant varier les seuils de taille et de nombre des composantes. C'est-à-dire que pour une taille t et un nombre n, on ne comptabilise que les composantes ayant une surface plus grande que t, puis un noyau est considéré comme boursouflé si le nombre de composantes connexes d'écart est supérieur ou égal à n. Il s'agit donc d'étudier une fonction discrète à deux variables qui produit une surface représentant le taux de classement. La figure 5.5 montre le résultat de cette étude.

Cette analyse met en exergue l'utilité de l'indice $\psi_{N_{Cce}}$ dans le cas de noyaux non convexes ayant au minimum soit une zone de concavité d'au moins trente-deux pixels, soit deux zones de concavité d'au moins douze pixels (figure 5.6).

On peut également remarquer sur la surface résultat, la partie totalement plane. Elle correspond en fait à des seuils en taille ou/et en nombre trop haut qui ont engendré le classement de tous les noyaux dans la classe "forme normale", ce qui vérifie que nous possédons près de 70% de noyaux avec une forme normale. A l'inverse, des seuils trop bas ont conduit à un classement de tous les noyaux dans la classe "noyaux boursouflés".

Les deux seuils extraits de la surface résultat sont utilisés de manière équivalente dans $\psi_{N_{Cce}}$ (sans pondération) et leur combinaison permet ainsi d'obtenir un taux de bon classement de plus de 90% pour le sous-problème de forme avec ce seul indice.

Cet indice est utilisable dans tous les problèmes de caractérisation de la convexité. Il peut être employé directement sans effectuer l'étude relative à la taille des composantes connexes. Mais ce calibrage permet de l'adapter de manière spécifique et ainsi de répondre au mieux au problème

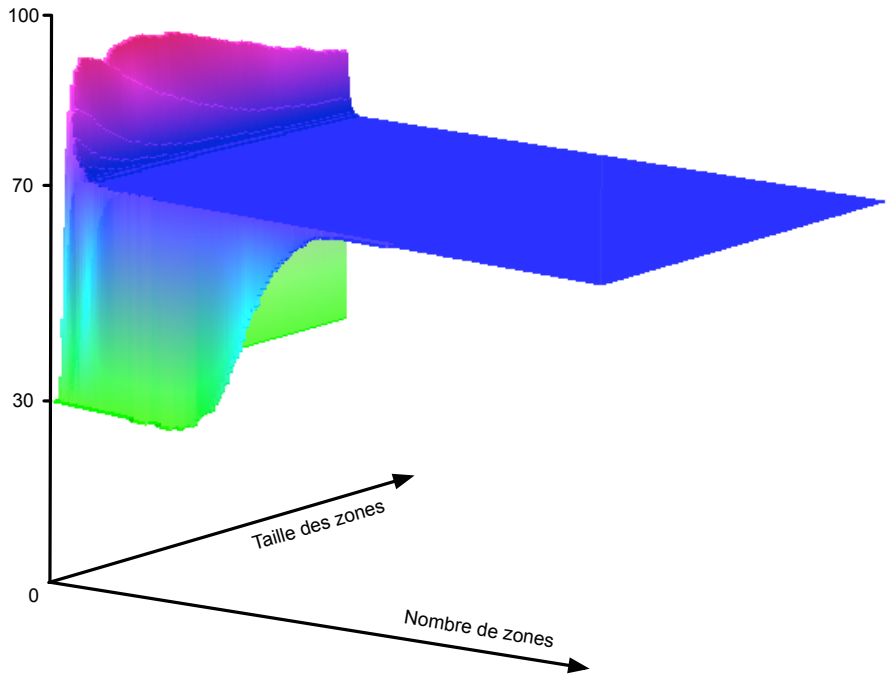
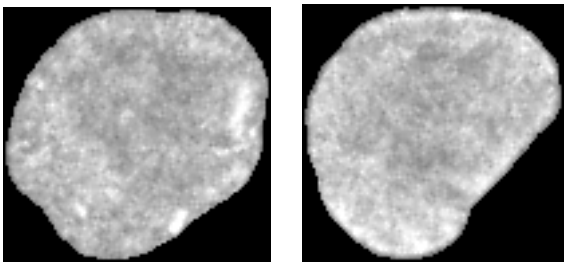


Figure 5.5. Surface représentant le pourcentage de bon classement des noyaux en fonction du nombre et de la taille des composantes connexes d'écart.



(a) (b)

Figure 5.6. Noyaux possédant des points de concavité. (a) un noyau avec deux points de concavité d'au moins 12 pixels, (b) un noyau avec un seul point de concavité d'au moins 32 pixels.

dans lequel il est utilisé. Le calibrage représente donc l'inconvénient de cet indice en terme de complexité et d'étude, mais il constitue également son principal avantage car il permet d'améliorer son efficacité.

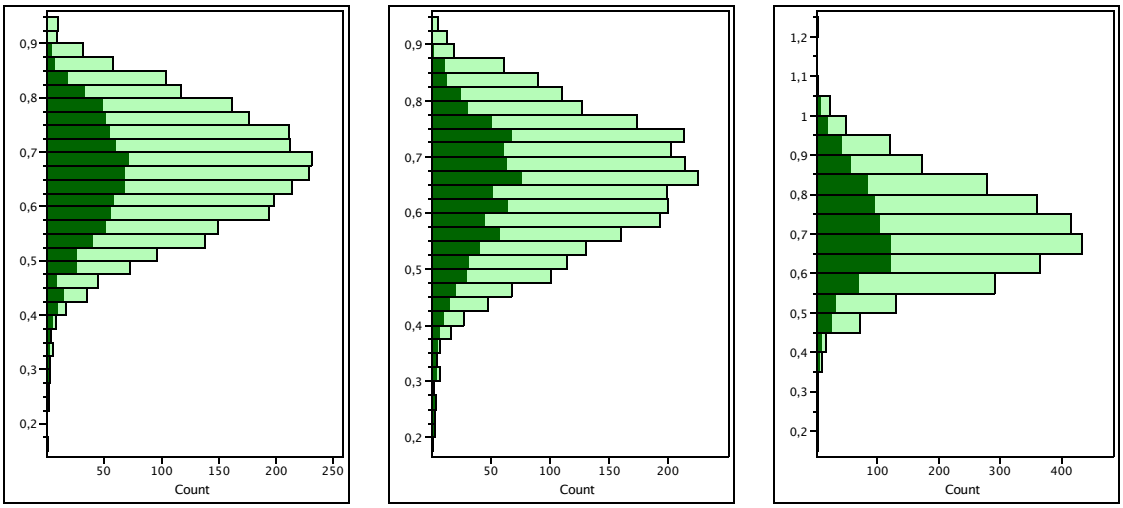
5.3 Analyse mono-variable

La littérature fournit quelques indices de forme parmi lesquels nous en avons retenu douze (cf. annexe B.2). Leurs constitutions (les mesures qui les composent) sont différentes ou combinées différemment. De plus, ils sont facilement calculables en 2D et par la suite en 3D (cf. section 9.3.2).

A ces indices s'ajoutent les quatre qui viennent d'être élaborés ainsi que l'indice de caractérisation de la courbure (cf. annexe B.3.1). Donc nous disposons d'un ensemble de 17 indices pour caractériser la forme des noyaux. Mais avant de construire le sous-modèle, il est nécessaire d'étudier ces indices.
Cette section réalise une étude préliminaire de l'intérêt des différents indices dans le problème de caractérisation de la forme. Ce travail porte sur l'analyse de la distribution des valeurs des indices qui sont engendrées par les noyaux, sur l'étude des *outliers*[2][Moore and McCabe 1998], sur leur potentiel de classement et les différentes corrélations entre les indices.

5.3.1 Les distributions

La première étape d'une analyse mono-variable est l'étude de la répartition des valeurs (la distribution) de chaque attribut. L'observation de l'histogramme d'une variable permet d'observer la répartition des individus de chaque classe. Plus la variable est pertinente, plus son histogramme révèle une distribution bien séparée des individus en fonction de leur classe.
On peut remarquer sur la figure 5.7 les distributions de différentes variables qui décrivent mal le problème.



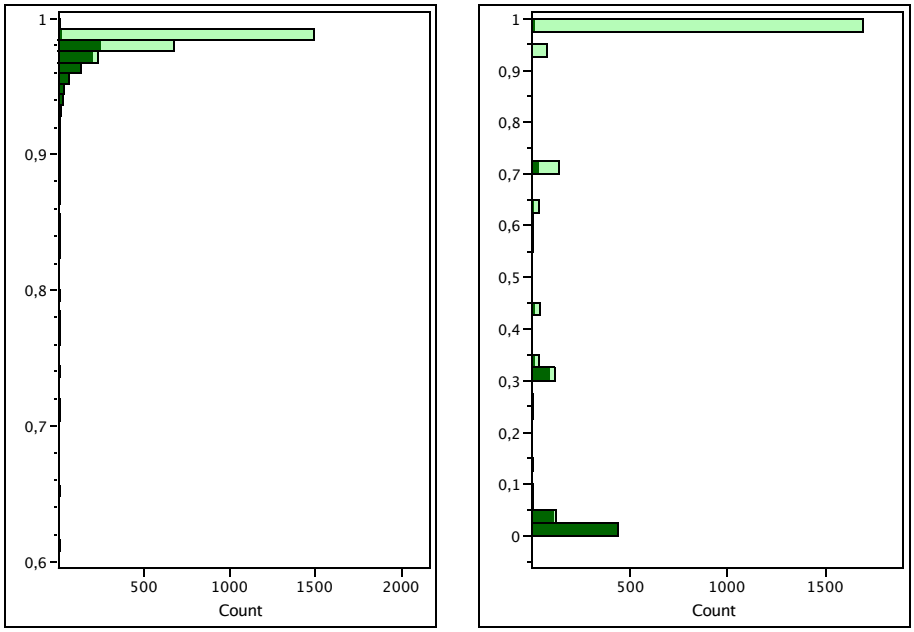
(a) (b) (c)

Figure 5.7. Histogrammes montrant la distribution des attributs : allongement par les rayons (a), écart au disque inscrit (b) et étalement de Morton (c). En vert foncé (resp. clair) les individus à forme anormale (resp. normale). On peut observer que ces trois variables ne permettent pas de séparer les classes d'individus.

Les distributions révèlent que seulement deux attributs sont intéressants pour décrire la forme des noyaux (cf. figure 5.8) : l'indice de convexité surfacique et ψ_{Ncce}. Sur leur histogramme on peut observer une séparation des individus appartenant à des classes différentes : plus la valeur d'un des indices est faible, moins on trouve d'individus ayant une forme normale et inversement. On peut également observer que ces indices discriminent fortement les individus : peu d'individus appartiennent à des classes différentes et ont des valeurs proches. Mais la présence d'erreurs est quasiment inévitable en analyse des données. S'il existait systématiquement une variable des-

[2]Ce sont les individus ayant des valeurs extrêmes. La définition exacte est donnée dans la section 5.3.2.

criptive pour laquelle il n'y aurait aucune erreur, alors la variable seule permettrait de résoudre le problème. Il ne serait alors plus utile de construire un modèle de classement par apprentissage, mais il suffirait de prendre une valeur seuil de la variable qui servirait de décision.



(a) (b)

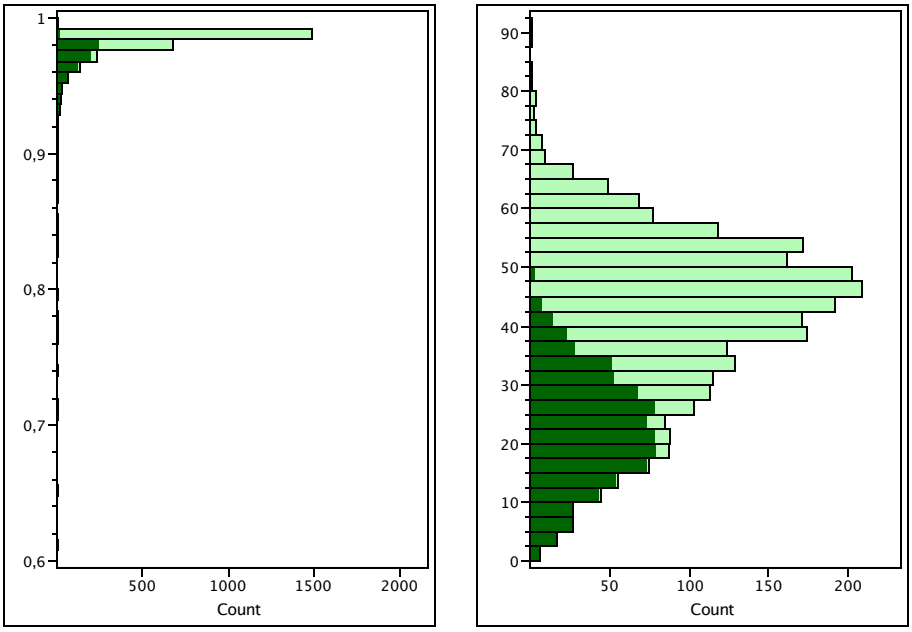
Figure 5.8. Histogrammes des valeurs de l'indice de convexité surfacique (a) et de l'indice ψ_{Ncce} (b). En vert foncé (resp. clair) les individus à forme anormale (resp. normale). On remarque la séparation des individus appartenant à des classes différentes : plus la valeur d'un indice est faible (resp. élevée), moins on trouve d'individus à forme normale (resp. anormale).

Sur la figure 5.8*a*, on peut observer une concentration extrême vers 1 de la distribution des valeurs de l'indice de convexité surfacique. Cette densité élevée sur un petit intervalle peut pénaliser une variable et rendre son utilisation plus difficile dans un classifieur. En effet le classifieur devrait déterminer une valeur de seuil ayant une précision très élevée. Pour corriger la mauvaise répartition de cette variable, nous utilisons une fonction d'étalement à base de tangente : $f(x) = tan(\frac{\pi}{2}x)$. La figure 5.9*b* montre la répartition résultat après utilisation de cette fonction d'étalement.

5.3.2 Les outliers

Les *outliers*[3] sont des individus ayant des valeurs extrêmes, distantes des autres. Pour définir le seuil de distance qui fait qu'un individu est un *outlier*, on utilise couramment la notion de *quartile*. Les quartiles sont trois seuils (haut, milieu et bas, notés Q_H, Q_M et Q_B) qui permettent de séparer une population en quatre groupes de taille égale. Les "boites" (*outliers boxes*) contenues dans la figure 5.10 représentent l'intervalle entre les seuils haut et bas des quartiles. Les *outliers* sont les individus situés en dehors de l'intervalle suivant : $[Q_B - k(Q_H - Q_B), Q_H + k(Q_H - Q_B)]$ avec k une constante. En statistique il est fréquent de choisir $k = 1,5$, comme c'est le cas dans le logiciel *SAS/JMP* que nous utilisons dans cette étude. Sur la figure 5.10 la droite grise représente cet intervalle et les individus affichés sont les *outliers*.

[3]Mot qui pourrait se traduire par *valeurs aberrantes*.



(a) (b)

Figure 5.9. Histogrammes des valeurs de l'indice de convexité surfacique lors du calcul de l'indice (a) et le résultat de l'étalement des valeurs (b). En vert foncé (resp. clair) les individus à forme anormale (resp. normale).

La présence d'*outliers* dans un échantillon d'apprentissage pose des problèmes aux classifieurs. En effet un classifieur doit tenir compte de ces individus dans la construction du modèle. Or prendre en compte une valeur extrême engendre une perte de précision pour le classement des individus ayant des valeurs "normales". Il est par conséquent nécessaire d'étudier puis, si nécessaire de traiter, voire supprimer les *outliers* de chaque indice, afin de vérifier qu'ils ne perturbent pas l'apprentissage du modèle et donc le résultat.

Un phénomène intéressant qui se produit sur certaines variables est révélé par les *outliers*, comme cela peut être observé sur la figure 5.10 : au delà d'un certain seuil, les *outliers* appartiennent tous à la même classe.

En remarquant ce phénomène, il est alors possible d'envisager de traiter tous les individus concernés (c'est-à-dire 93 sur l'ensemble des variables, soit 3% des individus). Pour chaque variable où l'on observe le phénomène, on trouve une valeur de seuil qui permet de classer les individus. Donc au delà de ce seuil, on est certain que les individus appartiennent à la même classe. Ce pré-traitement permettrait sans doute un meilleur apprentissage pour le modèle et notamment lors de l'emploi de méthodes sensibles aux outliers. Mais en pratique ce traitement n'est nécessaire car le sous-modèle de classement de la forme des noyaux qui est présenté (cf. section 5.4) obtient de très bons résultats et surtout classe correctement tous les *outliers* concernés par ce phénomène.

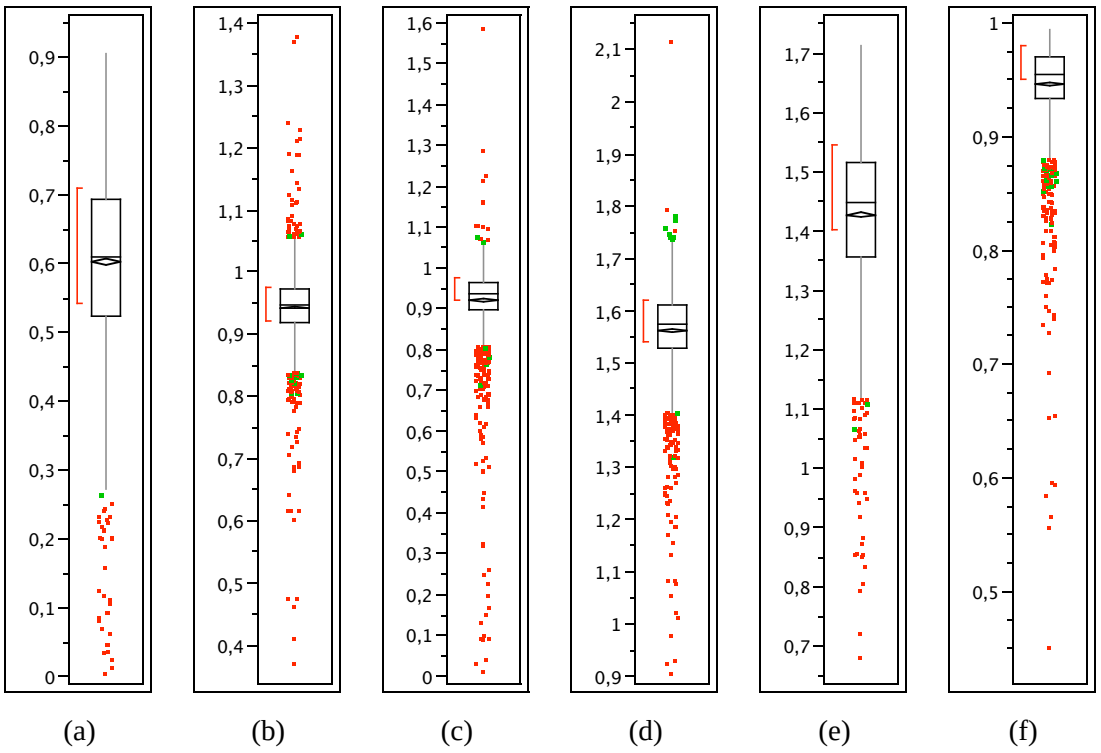


Figure 5.10. Etudes des *outliers* pour six attributs : la circularité (a), les indices d'ellipse par l'axe principal (b) et les rayons (c), les indices de parallélogramme par l'axe principal (d) et les rayons (e), l'indice de symétrie (f). On peut observer les individus à forme normale (vert) et à forme boursouflée (rouge).

5.3.3 Les classements mono-variables

Avant d'utiliser tous les attributs ou une partie des attributs dans le modèle, il peut s'avérer intéressant d'observer l'efficacité de chaque variable. Pour cela un modèle de classement mono-variable est réalisé par régression logistique avec chacune des variables. Les résultats de cette analyse se pressentent après l'observation des distributions. En effet, plus la distribution des valeurs d'une variable montre une bonne séparation des classes, plus la variable est pertinente pour classer les individus.
Pour comparer les performances des variables, on peut utiliser la valeur du test du χ^2 (cf. section 2.2.3). Ce test apporte une indication sur l'efficacité de la variable. Plus la valeur est proche est grande, plus la variable est pertinente.
La figure 5.11 montre la pertinence de classement de plusieurs variables. La courbe (en bleu) est une sigmoïde qui donne la probabilité (en ordonnée) de classement de la forme de noyaux en fonction de la valeur de la variable (en abscisse). Les noyaux ayant une forme normale (resp. boursouflée) sont représentés par des points verts (resp. rouges). Pour un noyau donné, sa position en abscisse dépend de la valeur de l'indice, mais en ordonnée il est placé aléatoirement sous (resp. sur) la courbe si sa forme est normale (resp. boursouflée). Plus les noyaux sont répartis sur les extrémités en fonction de leur classes d'appartenance, plus l'indice est pertinent.

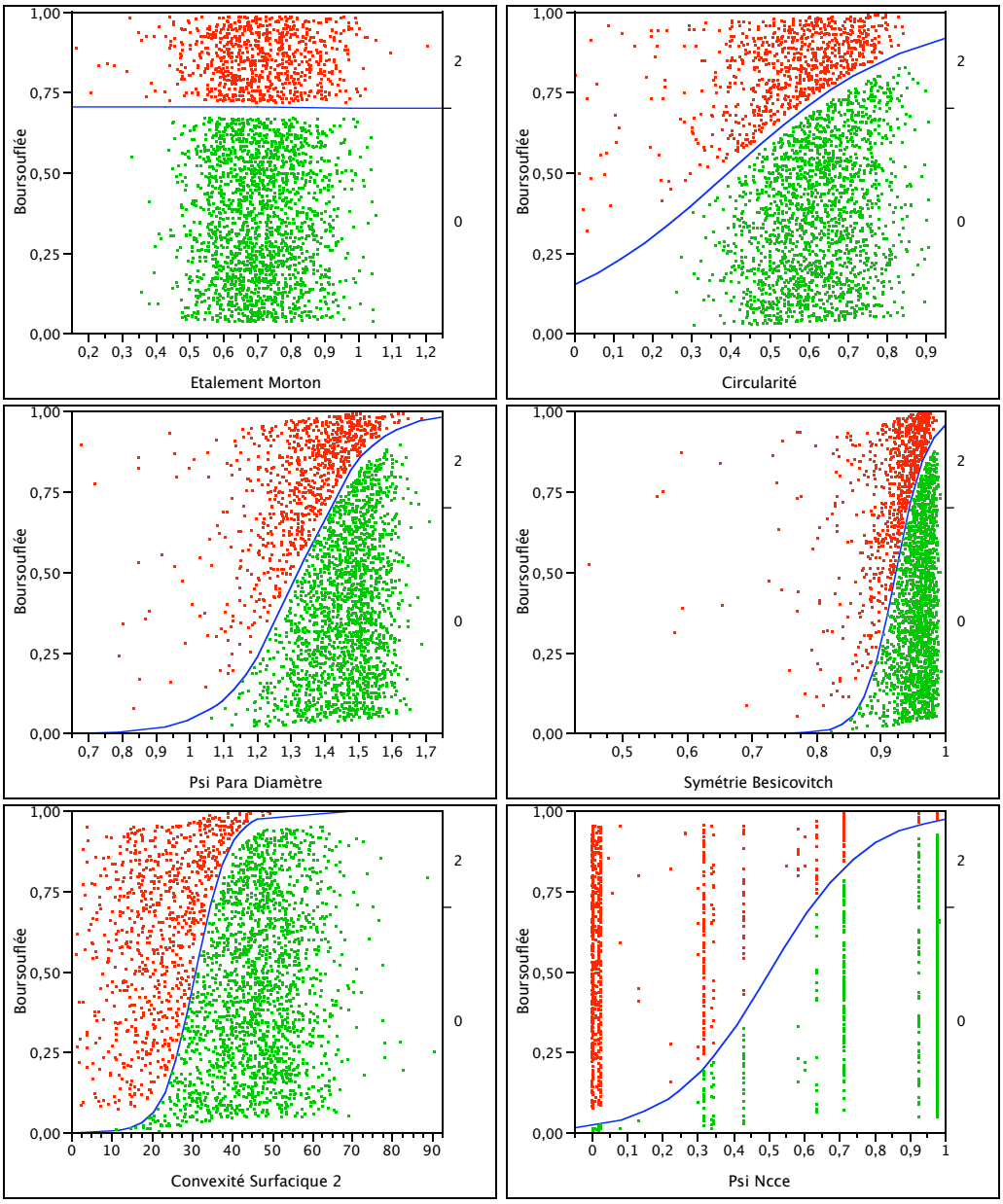


Figure 5.11. Illustrations des performances de différentes variables. En abscisse les valeurs des indices et en ordonnée la probabilité d'appartenance. En rouge (resp. vert), les noyaux ayant une forme boursouflée (resp. normale).

Les résultats des classements mono-variable sont les suivants :
- ψ_{Ncce}, $\chi^2 = 2323$.
- Convexité surfacique, $\chi^2 = 2044$.
- Convexité périmétrique, $\chi^2 = 886$.
- Symétrie de Besicovitch, $\chi^2 = 742$.
- $\psi_{Ellipse}$ par le diamètre, $\chi^2 = 604$.
- $\psi_{Ellipse}$ par l'axe principal, $\chi^2 = 497$.
- Nouveau déficit isopérimétrique, $\chi^2 = 342$.

Cette analyse met en exergue sept indices dont deux sont particulièrement efficaces. Les variables qui n'ont pas été citées ont un χ^2 presque nul, comme l'indice d'étalement de Morton dont on peut voir l'efficacité sur la figure 5.11.

REMARQUE - *L'analyse mono-variable de l'étalement de Morton (cf. figure 5.11) montre une courbe constante. Donc quelle que soit la valeur de l'indice, la probabilité de classement est la même. Cet indice n'apporte aucune information ($\chi^2 = 54.10^{-4}$) pour le classement de la forme des noyaux. La courbe montre la proportion de répartition de la population des noyaux dans les deux classes "forme normale" (*$70,5\%$*) et "forme boursouflée" (*$29,5\%$*).*

5.3.4 Les corrélations

Nous venons de présenter trois études préliminaires mono-variables pour déterminer ou améliorer leur efficacité dans le modèle de classement. Il est aussi intéressant d'étudier les corrélations entre les variables elles-mêmes. L'utilisation de deux variables fortement corrélées dans le modèle de classement n'apporterait que très peu d'informations pour comprendre les résultats obtenus par le modèle de classement. En effet, la valeur de l'une pourrait être déduite de la valeur de l'autre. La figure 5.12 présente les corrélations existantes les plus fortes. Plus le nuage de points formé par les variables a une forme proche d'une courbe que l'on peut calculer analytiquement en minimisant l'erreur, plus les variables sont corrélées. La forme de la courbe détermine le type de corrélation.

La probabilité d'une corrélation est la probabilité que deux variables soient naturellement disposées ainsi sans existence de corrélation entre elles. Dans ce manuscrit, les probabilités non précisée sont inférieures à 10^{-4}. L'analyse de la figure 5.12 met en exergue l'existence de corrélations très fortes (avec des cœfficients de corrélations supérieurs à $0,85$) entre couples ou groupes de variables :

- L'allongement par l'axe principal, l'allongement géodésique, l'allongement par les rayons, la circularité et l'écart au disque inscrit ont un cœfficient de corrélation linéaire supérieur à $0,92$.
- L'étalement de Morton a un coefficient de corrélation de $0,88$ avec le groupe précédent.
- L'allongement par le diamètre est corrélé avec les variables précédentes avec des coefficients se situant entre $0,8$ et $0,9$ à l'exception de la circularité $0,71$.
- On peut également observer une corrélation non linéaire entre les indices de convexité.

Cette dernière étude révèle de nombreuses corrélations qui influent dans la construction du modèle et surtout dans le choix des meilleures variables explicatives. Il est peu probable que deux variables fortement corrélées interviennent dans le modèle.

5.4 Modèle final de caractérisation de la forme

Nous bénéficions de quatre nouveaux indices spécifiquement construits pour caractériser la forme des noyaux. A ces indices s'ajoute également l'indice de caractérisation de la courbure qui est décrit dans l'annexe B.3.1. Avec les douze indices sélectionnés dans la littérature scientifique, nous disposons désormais de dix-sept indices de forme.

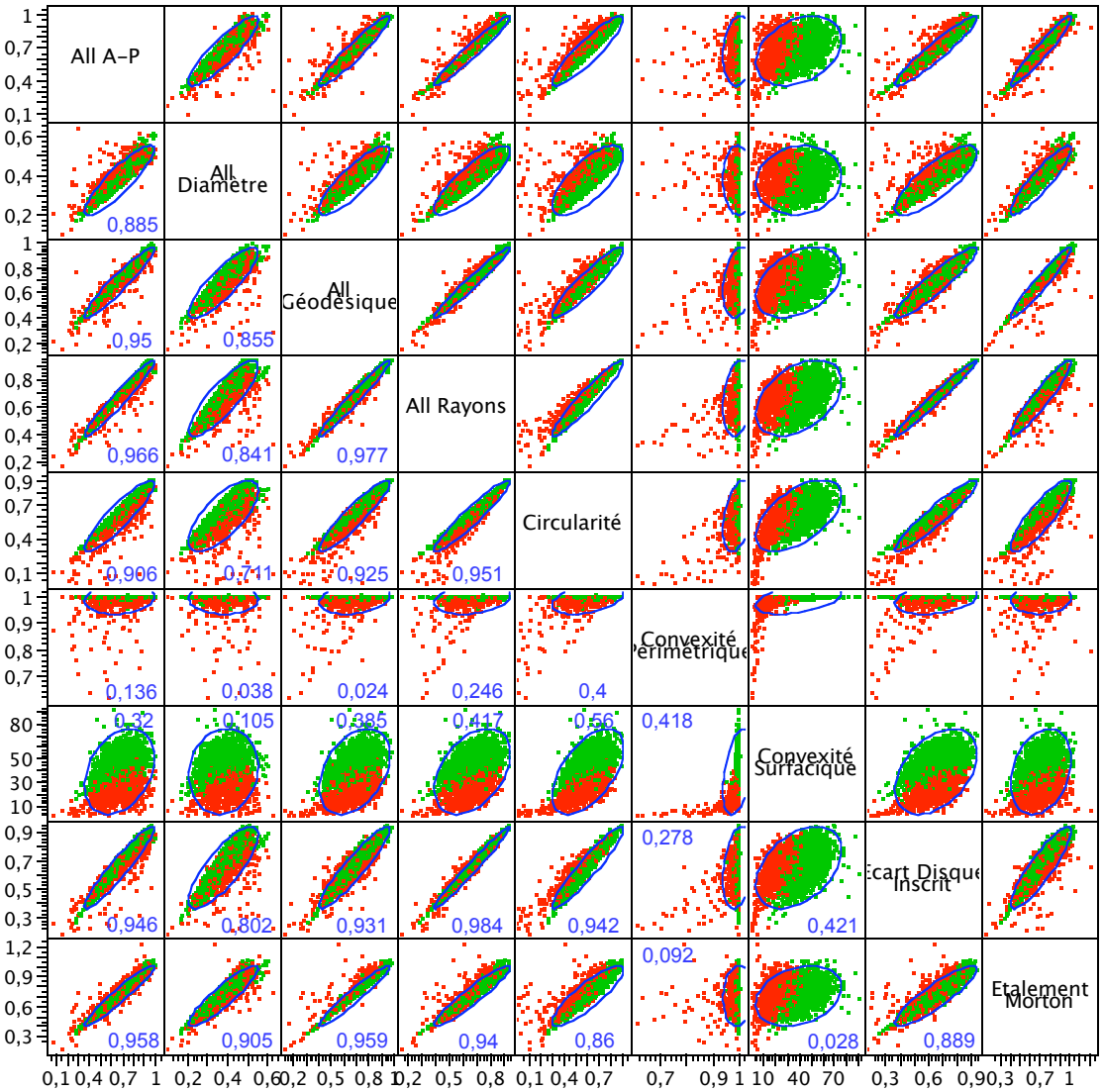


Figure 5.12. Illustrations et coefficients des corrélations les plus fortes entre les variables. Plus le nuage de points a une forme allongée et régulière proche d'une ellipse, plus la corrélation est importante. Les noyaux ayant une forme normale (resp. boursouflée) sont en vert (resp. rouge). L'ellipse (en bleu) contient 95% des noyaux sous l'hypothèse de binormalité.

Pour chaque classifieur, nous avons effectué une recherche (avec une validation croisée) afin de trouver le meilleur sous-ensemble d'indices (cf. tableau 5.1). Pour la régression logistique et les forêts aléatoires, une recherche exhaustive est réalisée. En revanche, il n'est pas possible d'appliquer une telle recherche dans un temps raisonnable pour les *k*-plus proches voisins et les réseaux de neurones. Nous utilisons une méta-heuristique de type *tabou* [Glover 1986; Glover 1990; Michel and Hentenryck 2004].

Le réseau de neurones utilisé est un perceptron multi-couches avec une couche cachée pour lequel nous testons différentes valeurs de ν ($\nu = 2...6$, cf. section 2.2.5).

Pour les k-plus proches voisins, nous testons différentes valeurs du paramètre k :
- k fixe pour des valeurs impaires de 1 à 19, ainsi que $k = 55$ (racine carrée du nombre d'individus).
- k variable, égal à $N_i + 2a + 1$, avec N_i le nombre d'indices utilisés dans le sous-ensemble testé et a variant de 0 à 10.

Nombre d'indices N_i \ Méthodes	$N_i + 5$-PPV	RL	FA	PMC / 4
1	92,24	93,28	92	93,27
2	92,44	93,86	91,97	93,93
3	93,67	94,91	92,73	94,86
4	93,71	95,02	93,6	95,09
5	93.31	95,14	93,93	95,09
6	93,42	95,24	94,11	95,13
7	93,89	95,24	94,36	95,27
8	93,93	95,31	94,33	95,31
9	93,57	95,35	94,33	95,35
10	93,71	95,41	94,4	95,38
11	93,53	95,39	94,29	95,41
12	93,53	95,37	94,22	95,37
13	93,64	95,35	94,18	95,35
14	93,38	95,35	94,33	95,27
15	93,49	95,24	93,71	95,27
16	93,35	95,16	93,89	95,27
17	93,17	95,13	92,8	95,27

Table 5.1. Pourcentage de prédiction obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement de la forme. Deux sous-ensembles de taille n et $n+1$ peuvent n'avoir aucun indice en commun. Les abréviations correspondent aux méthodes suivantes : les k-plus proches voisins ($N_i + 5$-PPV), la régression logistique (RL), les forêts aléatoires (FA) et le perceptron multi-couches (PMC/4).

La régression logistique et le réseau de neurones (avec $\nu = 4$) apportent des résultats comparables. Les meilleurs résultats sont : un sous-ensemble composé de dix indices pour la régression logistique et un sous-ensemble de onze indices pour le réseau de neurones. Ils permettent un pourcentage de prédiction de la forme des noyaux de $95,4\%$: $68,4\%$ de prédiction des noyaux à forme normale (soit 3% d'erreur) et 27% de prédiction des noyaux boursouflés (soit $2,5\%$ d'erreur). La probabilité d'obtenir ce résultat de manière aléatoire est inférieure à 10^{-4} et l'intervalle de confiance à 95% (calculé avec l'algorithme 4) est $[95,2 \ldots 95,6]$. Les performances des deux techniques étant comparables, nous construisons le sous-modèle à l'aide de la régression logistique avec dix indices. Il est préférable d'utiliser un modèle plus simple (moins complexe car linéaire), utilisant moins d'indices (réduction de la taille du vecteur caractéristique) et dont la probabilité associée à un individu est plus rapide à calculer.

NOTE - *Dans la suite de ce manuscrit, nous utiliserons l'abréviation VP pour désigner les vrais positifs et VN les vrais négatifs.*

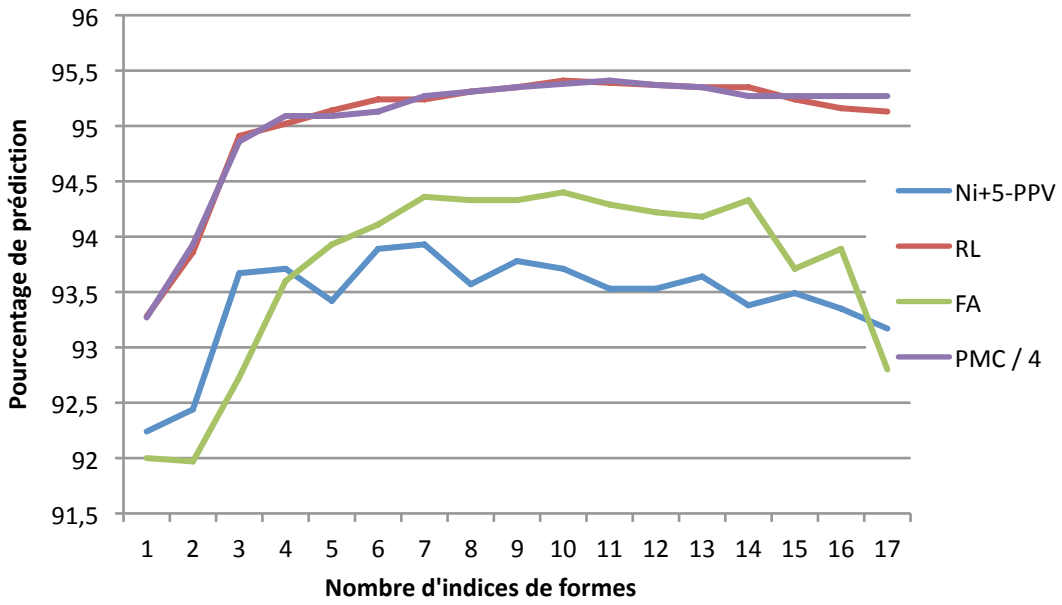


Figure 5.13. Comparaison graphique des performances des différentes méthodes de classement en fonction du nombre d'indices pour le classement de la forme. En abscisse le nombre d'indices de forme utilisés et en ordonnée le pourcentage de prédiction obtenu.

REMARQUE - *Sans l'utilisation des quatre indices que nous avons créés, le meilleur résultat est un sous-ensemble composé de dix indices sur les douze disponibles. Il permet* $93,6\%$ *de prédiction avec un intervalle de confiance de* $[93,3\ldots93,9]$. *Ces résultats sont* 2% *plus faibles avec un intervalle de confiance plus grand et l'intersection des deux intervalles de confiance est nulle. Ceci démontre l'efficacité de nos indices dédiés.*

Le figure 5.14 présente l'histogramme des distributions des probabilités attribuées aux noyaux par le sous-modèle issu de la régression logistique.

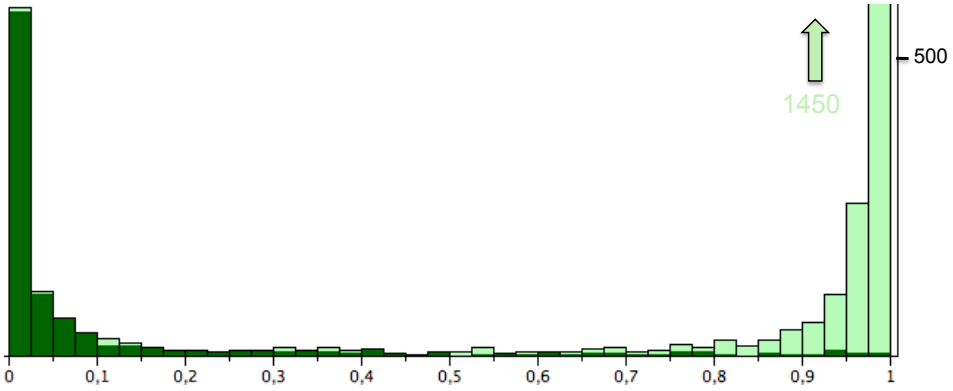


Figure 5.14. Distribution des probabilités attribuées aux noyaux par le sous-modèle de classement de la forme. En vert clair (resp. vert foncé) les individus ayant une forme normale (resp. boursouflée).

Sur l'histogramme (cf. figure 5.14) on remarque :
- La forte répartition des probabilités sur les extrémités de l'histogramme.
- La présence de seulement quelques cas ambigus.

Ces deux informations montrent le peu d'ambiguïté dans le classement et par conséquent confirment le pouvoir de classement du sous-modèle ainsi que son efficacité.

La liste suivante montre le meilleur sous-ensemble d'indices utilisés, classés par ordre décroissant d'importance dans le modèle. Ce classement a été effectué en fonction du test du χ^2 :

1. Ψ_{Ncce}, $\chi^2 = 167$
2. Convexité surfacique, $\chi^2 = 128$
3. $\Psi_{Ellipse}$ par le diamètre, $\chi^2 = 16$
4. Symétrie de Besicovitch, $\chi^2 = 10$
5. Allongement par le diamètre, $\chi^2 = 7$
6. $\Psi_{Ellipse}$ par l'axe principal, $\chi^2 = 5,3$
7. Convexité périmétrique, $\chi^2 = 5$
8. Allongement par les rayons, $\chi^2 = 3$
9. Déficit, $\chi^2 = 2,6$
10. Circularité, $\chi^2 = 1$

REMARQUES - *L'ordre d'importance des indices apparaît comme très logique :*
- *L'indice Ψ_{Ncce}, car il a été conçu spécifiquement pour répondre au problème. L'analyse mono-variable a montré que c'était l'élément le plus pertinent et il est relativement peu corrélé avec les autres indices*[4]*). L'indice de caractérisation de la convexité est l'élément de diagnostic le plus important.*
- *L'indice de convexité surfacique apporte des informations complémentaires non corrélées avec la convexité. De plus, cet indice est fortement lié à son homologue périmétrique. Donc utiliser les deux indices dans le modèle n'apporte que peu d'informations supplémentaires par rapport à l'utilisation d'un seul des deux. Ceci relègue l'indice de convexité périmétrique parmi les indices les moins importants pour ce modèle au bénéfice de l'indice de convexité surfacique.*
- *En troisième position se trouve l'indice de caractérisation des ellipses par le diamètre car il permet de caractériser les noyaux à forme normale.*
- *Les noyaux à forme normale, donc quasi elliptique, possèdent un point de symétrie. Ainsi l'indice de symétrie de Besicovitch (cf. section B.2) permet aussi de discriminer une partie des formes normales. Son analyse en classement mono-variable montrait sa pertinence.*
- *En sixième position, on retrouve un autre indice de caractérisation des ellipses, mais qui utilise des mesures différentes. Il est le seul indice parmi les dix à utiliser l'axe principal.*
- *On remarque également la présence de deux des indices dont on a étudié les fortes corrélations. L'indice de circularité et l'allongement par le diamètre appartiennent au groupe des indices les plus corrélés. Ces indices utilisent en fait des mesures différentes qui apportent des informations complémentaires.*

[4] Coefficient de corrélation égal à $0,799$ avec l'indice de convexité surfacique et $0,497$ avec l'indice de symétrie de Besicovitch.

5.5 Conclusion

Dans ce chapitre nous avons présenté l'utilisation des indices de forme pour la caractérisation et le classement de la forme des noyaux de cellule (cf. figure 5.15).
La souplesse des indices de forme a permis de construire trois nouveaux indices afin de permettre une meilleure discrimination des noyaux. L'utilisation de ces nouveaux indices dédiés dans le sous-modèle permet d'obtenir un pourcentage de prédiction de $95,4\%$ de la forme des noyaux sur l'échantillon de travail.

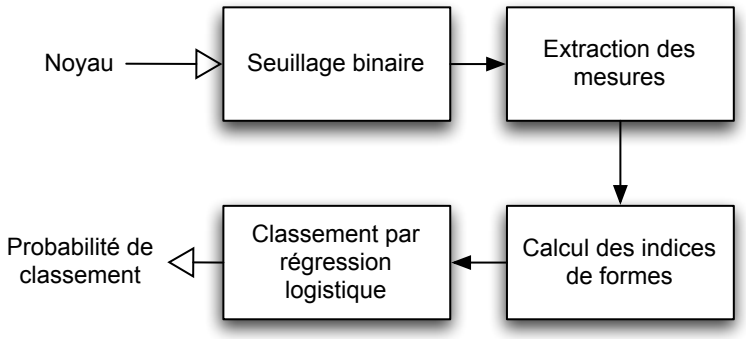


Figure 5.15. Schéma récapitulatif des différentes étapes nécessaires à la construction du sous-modèle de classement de la forme.

L'étude des faux (positifs et négatifs) n'a pas permis de révéler une structure particulière de ces noyaux.
Ce résultat est tout à fait satisfaisant (cf. section 2.2) car il est nettement supérieur au taux de répétabilité des experts qui est de l'ordre de 92% pour la forme. L'utilisation d'une méthode de classement complexe telle que les réseaux de neurones n'a pas permis d'améliorer les résultats, mais a apporté des performances comparables.
Ces résultats ont donné lieu au deux publications suivantes [Thibault et al. 2007; Thibault et al. 2008a].

Malgré ce bon pourcentage de prédiction et le fait que la forme représente l'élément de diagnostic le plus important, ce sous-modèle permet de classer uniquement $86,9\%$ des noyaux pour le problème sain/pathologique. Ce résultat était prévisible à partir de l'analyse des données (cf. chapitre 3). Il est par conséquent nécessaire de construire un sous-modèle de caractérisation de la texture des noyaux afin d'améliorer le classement.

DEUXIÈME PARTIE

CARACTÉRISATION ET CLASSEMENT DE LA TEXTURE DES NOYAUX

CHAPITRE 6

ANALYSE, TRAITEMENT PRÉLIMINAIRE ET ÉCHANTILLON DE TRAVAIL DANS L'ÉTUDE DE LA TEXTURE

6.1 Introduction

L'expertise menée sur les noyaux utilise comme critère secondaire une analyse de la texture et plus particulièrement de son homogénéité. Ce sont deux notions qu'il faut comprendre et définir afin d'appréhender les difficultés du sous-problème de classement de la texture.

La notion de texture est particulièrement délicate à aborder car il n'existe pas de définition générique. De plus, elle est liée au facteur d'échelle : une texture peut ne pas avoir les mêmes propriétés pour des facteurs différents[1]. Le nombre infini de textures et le peu de classes de texture possédant une définition formelle, font que la majorité des définitions et méthodes de caractérisation sont inhérentes au problème que les auteurs souhaitent résoudre. La quantité des méthodes de caractérisation témoigne de l'absence de définition ou de taxinomie voire d'organisation précise [Tuceryan and Jain 1998].
Dans de nombreux travaux, la notion de texture est liée à la notion de perception [Gagalawicz 1983] : *"le concept de texture est intimement lié à l'observateur humain [...] il est évident que l'on ne peut dissocier une texture de la manière dont celle-ci est perçue par le système visuel"*. Dans [Costa 2001], l'auteur décrit les deux phases de la perception par le système visuel :

- la vision pré-attentive, qui dure quelques milli secondes. C'est une *première impression* qui fait appel à la présence de formes élémentaires nommées les *textons* (droites, cercles, etc.).
- la vision prolongée qui nécessite une focalisation sur les détails et les éventuels arrangements.

De nombreux travaux ont tenté de transposer en algorithme le système visuel humain dans le cadre de la vision artificielle. Soit en modélisant la première phase [Hough 1962; Duda and Hart 1972], soit la deuxième [Haralick et al. 1973].

Dans ce chapitre, nous présentons tout d'abord la notion de texture et montrons quelques exemples de textures spécifiques. Ensuite, nous étudions la notion de texture homogène qui est au cœur du problème de classement de cette partie. Pour finir nous détaillons le processus de construc-

[1] Cette propriété a été largement utilisée dans les œuvres de l'artiste *Octavio Campo*. Ses œuvres représentent des scènes différentes en fonction du facteur d'échelle utilisé : http ://www.octavioocampo.com.mx/

tion de l'échantillon de travail, car l'expertise opérée sur les données a conduit à des classes totalement déséquilibrées pour le sous-problème de la texture et il n'est donc pas possible de travailler avec la totalité des noyaux.

6.2 La texture

Il n'existe pas de définition générique d'une texture. Mais on retrouve parfois certaines définitions dont notamment la suivante :

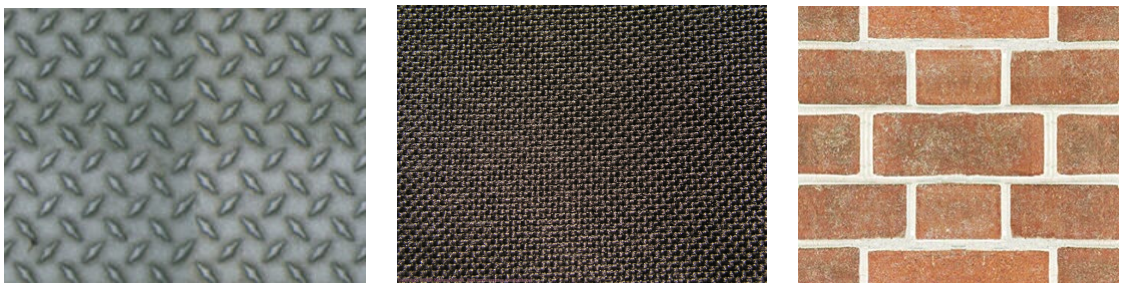
Définition 6.2.1 (Texture) – *Une texture est une image représentant une surface et permettant de simuler l'apparence de celle-ci quand on la plaque sur un objet tridimensionnel.*

Cette définition permet de comprendre que toute image est une texture à partir du moment où elle peut décrire une surface. Donc toute image est une texture, ce qui explique qu'il est quasiment impossible de ranger efficacement toutes les textures à l'aide de classes pertinentes.
Mais il existe tout de même une classification constituée de trois classes et contenant les textures :

1. structurelles, appelées aussi périodiques ou macro-textures ;
2. aléatoires ;
3. directionnelles.

6.2.1 Les textures structurelles

On peut considérer les textures structurelles comme étant la répétition de motifs élémentaires. La répartition spatiale de ces motifs de base suit des règles simples de direction et de placement. Donc le motif de base se répète de manière ordonnée et on parle alors de *texture ordonnée*.



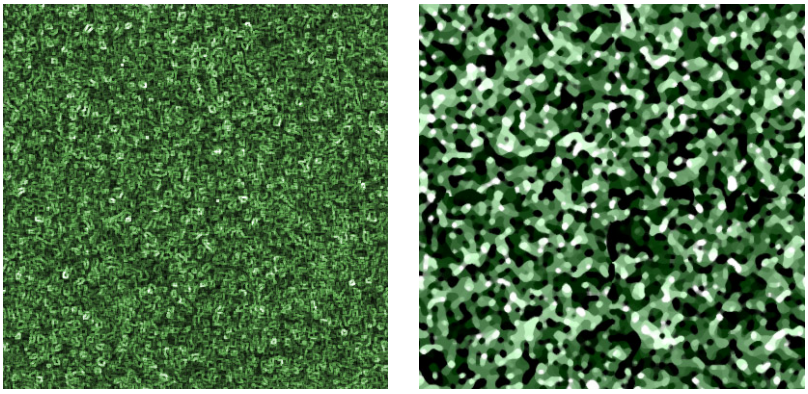
(a) (b) (c)

Figure 6.1. Trois exemples de textures structurelles. (a) et (b) deux textures de métal, (c) une texture brique.

Les méthodes décrivant ce type de texture essaient de découvrir et de caractériser le motif de base (motif générateur).

6.2.2 Les textures aléatoires

Contrairement aux textures structurelles, les textures aléatoires ne contiennent pas de motif de base et chaque pixel semble avoir été tiré aléatoirement. Toutefois ces textures donnent une impression d'homogénéité.



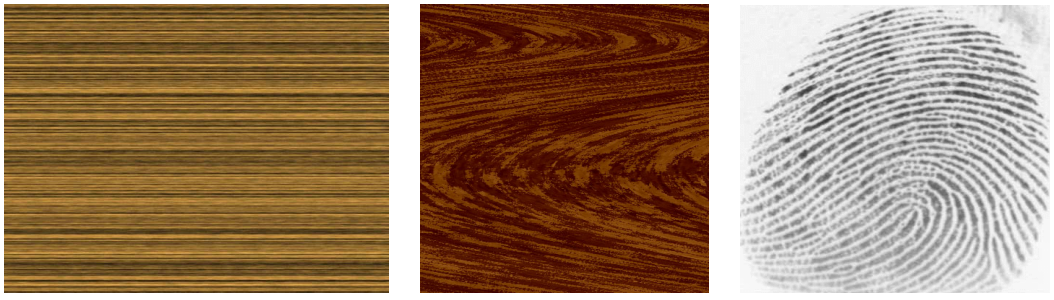
(a) (b)

Figure 6.2. Deux exemples de textures aléatoires.

Une solution pour analyser ces textures est de procéder à une analyse statistique [Haralick et al. 1973; Haralick 1979; Galloway 1975; Albregtsen 1995; Chen et al. 1995].

6.2.3 Les textures directionnelles

Les textures directionnelles n'utilisent pas de motif de base et ne sont pas pour autant aléatoires. Les intensités des pixels qui les composent forment des motifs organisés selon des directions bien précises. Dans [Costa 2001], les auteurs définissent une texture directionnelle comme une texture ayant des primitives pouvant être approchées par une fonction directionnelle.



(a) (b) (c)

Figure 6.3. Trois exemples de textures directionnelles. (a) Une texture de bois dans une direction unique, (b) une autre texture de bois avec différentes directions, (c) une empreinte digitale avec de multiples directions.

Deux travaux en particulier [Mavromatis 2001; Costa 2001] résument et traitent les problèmes de textures directionnelles.

6.2.4 Définition générale

Les définitions de la texture se restreignent à des cas particuliers. Cependant, on retrouve souvent la notion d'arrangement spatial.
Dans le langage courant, la texture est employée pour caractériser la surface d'un objet et on lui associe régulièrement un adjectif : granuleuse, lisse, peau d'orange, etc.

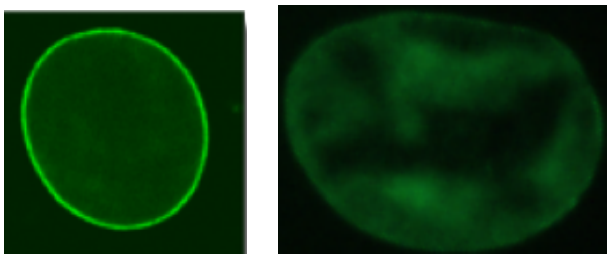
Dans [Mavromatis 2001], l'auteur explique que la texture est une information obtenue à partir d'un ensemble de mesures locales (statistiques, géométriques, sémantiques, etc.) dans une région (appelée *fenêtre de visualisation*) d'une image. Par conséquent une texture à la particularité d'être homogène au regard de la fenêtre de visualisation choisie. Or dans notre problème, la texture d'un noyau est l'ensemble de tous les pixels du noyau (nommés *texels*[2]) et la fenêtre de visualisation englobe le noyau. Donc par définition la texture d'un noyau est homogène, ce qui est en contradiction avec ce que les experts nomment *texture homogène*.
Il faut par conséquent définir la notion de *texture homogène* dans notre contexte.

6.3 Texture homogène

Afin d'apporter une définition à la notion de *texture homogène* relative à notre problème, nous commençons par définir une *région homogène*.

Définition 6.3.1 (Région homogène) – *Une région homogène est définie par une grande ressemblance entre les éléments qui la composent (faible inertie intra-classe) et une faible ressemblance avec les éléments appartenant à d'autres régions (forte inertie interclasse).*

Dans une texture, les éléments sont les caractéristiques extraites des texels, c'est-à-dire leur intensité. Donc une région homogène d'un noyau est une partie de la texture, plus exactement un ensemble de texels d'intensité proche voire égale (faible inertie intra-classe) et formant un ensemble connexe. Donc la texture d'un noyau est un ensemble de régions homogènes (cf. figure 6.4).



(a) (b)

Figure 6.4. Deux exemples de textures de noyaux. (a) une texture homogène issue d'un noyau sain. (b) une texture possédant de grandes régions homogènes mais avec une inertie inter-classe forte, donc une texture non homogène.

[2] Une texture est composée de texels (pour *TEXture ELement*), l'équivalent des pixels d'une image.

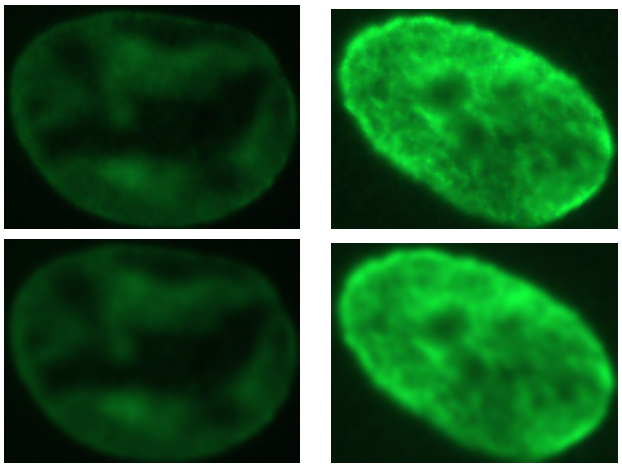
Dans notre problème, les experts utilisent le terme *homogène* pour qualifier des textures ayant de grandes (au sens du nombre de texels) régions homogènes, avec une faible inertie interclasse (ici inter-régions).
Dans une texture, plus les régions sont grandes et leurs niveaux de gris proches, plus la texture est considérée comme homogène.

Cette section a permis de définir l'élément de diagnostic que l'on souhaite caractériser afin de résoudre le sous-problème de classement de la texture.
Dans les chapitres suivants, nous utilisons ou définissons des méthodes permettant de caractériser la présence de telles régions dans une texture.

6.4 Traitement préliminaire

Dans le chapitre 1, il a été expliqué que pour observer les noyaux, les experts utilisent un marqueur fluorescent de type FITC (cf. annexe C.1.1). Ce marqueur réagit à la présence des lamines AC dans le noyau et révèle leur répartition qui doit être uniforme au centre et légèrement plus importante sur la périphérie. Mais en pratique, il est quasiment impossible que le marqueur se répartisse régulièrement sur la totalité du noyau. Une inégalité de répartition du marqueur pourrait laisser croire que les lamines AC ne sont pas réparties de manière uniforme et ainsi engendrer une expertise erronée. Forts de cette information, les experts considèrent comme texture non homogène, uniquement les noyaux ayant une texture *fortement non homogène* (figure 6.5).

Pour contourner ce problème pendant l'analyse, nous employons au préalable un filtre de convolution dont le masque de taille 3×3 est rempli par un noyau Gaussien. Ce filtrage est appliqué sur la totalité de la texture afin d'atténuer les éventuelles variations de marquage.



(a) (b)

Figure 6.5. Deux noyaux de cellules possédant une texture fortement non homogène et les résultats après un filtrage par convolution de type Gaussien.

6.5 Echantillon de travail et validation

Contrairement à l'expertise de la forme, l'expertise de la texture a fourni deux classes, *homogène* et *non homogène*, totalement déséquilibrées (cf. section 3.2) ; parmi les 3000 noyaux dont nous disposons, seulement 135 possèdent une texture non homogène, ce qui fait environ vingt fois plus d'individus dans la classe homogène. Pour construire efficacement le modèle, il faut créer un échantillon de travail contenant des classes équilibrées.
L'échantillon de travail est construit à l'aide de l'algorithme 5 décrit dans la section 2.3.6. Il contient les 135 individus à texture non homogène et un nombre égal de parangons (cf. définition 2.4.5) des noyaux à texture homogène. L'échantillon contient 270 individus.

Le type et le nombre des caractéristiques étant modifiés à chaque nouveau sous-modèle de classement de la texture, l'algorithme de construction de l'échantillon de travail est systématiquement utilisé. Il en est de même pour la sélection des parangons et donc le contenu de l'échantillon de travail est spécifique à chaque étude.

En raison du faible nombre d'individus, la validation est effectuée à l'aide du protocole *Leave-One-Out* (cf. section 2.3.1.2) pour tous les sous-modèles de classement des noyaux en fonction de l'homogénéité de la texture.

CHAPITRE 7

MATRICES DE COOCCURRENCES ET CARACTÉRISTIQUES HARALICK

7.1 Introduction

Les différents types de textures et les propriétés que nous devons caractériser viennent d'être présentés. Les textures que nous traitons sont de type aléatoire pour lesquelles les méthodes de caractérisations statistiques sont des outils de choix.
Dans ce chapitre, nous utilisons une méthode de caractérisation statistique de la texture parmi les plus connues et les plus anciennes : la matrice de cooccurrences analysée par les caractéristiques Haralick.

7.2 Matrices de cooccurrences

La matrice de cooccurrences (ou matrice de dépendance spatiale) est une des approches les plus connues et les plus utilisées pour extraire des caractéristiques de texture. Elle effectue une analyse statistique de second ordre de la texture, par l'étude des relations spatiales des couples de pixels [Haralick et al. 1973; Haralick 1979; Castanon et al. 2007].
La matrice de cooccurrences s'intéresse aux relations qui existent entre les niveaux de gris des pixels de la texture (de l'image) pour un déplacement (translation) $\overrightarrow{d}$ donné. Le résultat est une matrice carrée de taille $N \times N$, où N est le nombre de niveaux de gris de la texture. Pour un déplacement $\overrightarrow{d} = (dx, dy)$, un élément (x,y) de la matrice est défini par le nombre de pixels de la texture de niveau de gris y situés à un déplacement $\overrightarrow{d}$ d'un pixel de niveau de gris x (cf. figure 7.1).
Ce qui peut s'écrire formellement :

$$M_d(x,y) = card\left\{\left((r,s),(r+dx,s+dy)\right)/I(r,s) = x, I(r+dx,s+dy) = y\right\}$$

NOTE - *Dans la pratique, les nuances de gris sont codées le plus souvent sur* 256 *niveaux. La matrice de cooccurrences correspondante (*256×256*) est très sensible à la dynamique des niveaux de gris, puisqu'un changement minime d'intensité peut générer une matrice complètement différente. Donc en pratique on ne construit pas la matrice sur la texture originale, mais sur une version altérée de ses niveaux de gris, en réduisant le nombre de niveaux de gris de la texture. Un codage de l'image sur* 8, 16, 32 *ou* 64 *niveaux de gris constitue en général un bon compromis.*

Généralement le nombre de niveaux de gris est une puissance de 2 afin de permettre un découpage égal et sans ambiguïté des niveaux de gris.

Texture

1	1	0	0
1	1	0	0
0	0	2	2
0	0	2	2

$\Longrightarrow$

Niveaux de gris i \ j	$\overrightarrow{d} = (0,1)$ 0	1	2	$\overrightarrow{d} = (1,1)$ 0	1	2
0	4	0	2	3	1	1
1	2	2	0	1	1	0
2	0	0	2	1	0	1

Figure 7.1. Exemples de résultats de remplissage de la matrice de cooccurrences pour deux déplacements différents ($\overrightarrow{d} = (0,1)$ et $\overrightarrow{d} = (1,1)$) pour une même texture à trois niveaux de gris.

La matrice de cooccurrences met en évidence les relations qui existent entre les pixels à la fois par un aspect local (les niveaux de gris) et un aspect spatial (le déplacement). Cependant, toutes les caractéristiques sont extraites si on calcule un grand nombre de matrices : on utilise un grand nombre de déplacements différents avec des niveaux de gris différents, ce qui génère une quantité importante d'informations. Plusieurs axes de recherche ont été développés, d'une part pour choisir des vecteurs de déplacement pertinents [Davis et al. 1979; Sun and Wee 1983] et d'autre part pour simplifier les matrices [Zucker and Terzopoulos 1980; Lohmann 1995].
Pour un nombre de niveaux de gris N préfixé, nous calculons quatre matrices pour les quatre déplacements $\overrightarrow{d_1} = (1,0)$, $\overrightarrow{d_2} = (1,1)$, $\overrightarrow{d_3} = (0,1)$ et $\overrightarrow{d_4} = (-1,1)$. Ensuite nous effectuons la moyenne des quatre matrices résultats [Yogesan et al. 1996; Wouwer et al. 1999], ce qui permet de fusionner les informations et de s'abstraire de la direction. Toutefois, il est toujours nécessaire de faire varier le nombre de niveaux de gris et de tester des déplacements plus importants (éloignement $\varepsilon\overrightarrow{d_i}$, $\varepsilon \in \mathbb{N}^*$), ce qui augmente le nombre de calculs et d'informations.

$$M_{\varepsilon,N} = \frac{1}{4}\sum_{i=1}^{4} M_{\varepsilon\overrightarrow{d_i},N}$$

A partir de cette matrice réduite, on extrait différents attributs appelés *indices de texture du second ordre* ou *caractéristiques Haralick* [Haralick et al. 1973; Haralick 1979; Parkkinen et al. 1990] du nom de leur concepteur. La liste complète des caractéristiques Haralick utilisées dans ce manuscrit est dans l'annexe B.5.

7.3 Construction du modèle

Après avoir construit l'échantillon de travail, nous avons réalisé une recherche exhaustive afin de déterminer la méthode de classement la plus efficace ((cf. section 2.2.1)) et le meilleur sous-ensemble d'indices de texture parmi les quinze disponibles (cf. annexe B.5). De plus, nous avons effectué les calculs pour 16, 32 et 64 niveaux de gris et pour des éloignements allant de un à cinq ($\varepsilon = 1 \ldots 5$).

Le meilleur sous-modèle de classement de l'homogénéité de la texture est construit par régression logistique, mais il existe un autre sous-modèle avec des résultats comparables basé sur les

réseaux de neurones. Ce sous-modèle est calculé sur une matrice à 32 niveaux de gris, pour un éloignement de 1 et il est constitué de 8 indices :

- La variance.
- La corrélation.
- La moyenne des sommes.
- L'entropie des sommes.
- L'entropie.
- La variance des différences.
- L'homogénéité.
- La dissimilarité.

Le sous-modèle obtient un pourcentage de prédiction de $89,8\%$ ($45,6\%$ VP et $44,2\%$ VN) avec un intervalle de confiance de $[86,9\ldots92,6]$ et une probabilité inférieure à 10^{-4}. Ce pourcentage est bien supérieur au taux de répétabilité des experts (85%, cf. section 3.4). Mais lorsque l'on calcule l'intervalle de confiance sur 63% des individus $[78,2\ldots83,7]$, on remarque que ce pourcentage est nettement plus élevé que la borne supérieure de l'intervalle, ce qui implique le modèle est sensible aux données : supprimer des données perturbe les performances du modèle. On peut remarquer que la distribution des probabilités attribuées par le modèle est moins bonne que celle du sous-modèle de forme (cf. figure 7.2). Malgré une importante répartition sur les extrémités, on peut constater que proportionnellement il existe plus de cas ambigus (40 individus) et d'erreurs graves (8 individus). Ceci permet de conclure que comparativement, ce sous-modèle est moins performant.

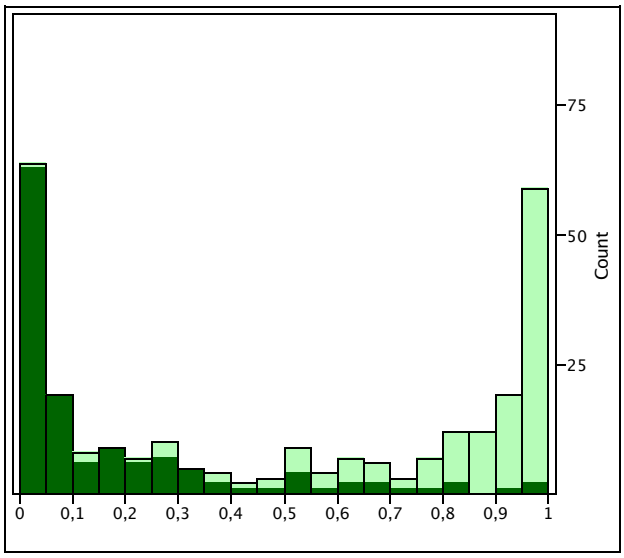


Figure 7.2. Histogramme des distributions des probabilités attribuées par le sous-modèle de classement de l'homogénéité de la texture des noyaux par caractéristiques Haralick. En vert clair (resp. vert foncé) les individus à texture homogène (resp. non homogène).

Il y a plusieurs explications à ce résultat moins performant : la principale raison est le biais introduit par les défauts de marquage qui diminuent la fiabilité du diagnostic. Ensuite, les caractéristiques Haralick constituent une méthode de caractérisation générale de la texture et donc absolument pas spécifique à notre problème.
Pour cette dernière raison, il nous faut trouver de nouvelles méthodes qui répondent mieux au sous-problème de classement de la texture que nous souhaitons résoudre.

CHAPITRE 8

NOUVELLE MÉTHODE DE CARACTÉRISATION DE LA TEXTURE : GRAY LEVELS SIZE ZONE MATRIX

8.1 Introduction

Les matrices de cooccurrences sont une des plus anciennes et des plus efficaces méthodes de caractérisations statistiques d'une texture. Elles ont permis de construire un sous-modèle de classement atteignant juste 90% sur l'échantillon d'apprentissage dans notre contexte. Bien que ce taux soit supérieur au taux de répétabilité des experts, nous souhaitons l'améliorer.
Nous avons vu au chapitre 6 qu'une texture homogène se caractérise par le nombre, la taille et les intensités de ses régions. Donc il vient intuitivement l'idée de dénombrer les tailles des zones de même niveau de gris.
Pour cette raison ce chapitre commence par décrire et utiliser une ancienne méthode de caractérisation basée non plus sur des couples de texels, mais sur des segments (cf. section 8.2). Par la suite cette méthode est modifiée (cf. section 8.3) afin de répondre spécifiquement au problème de caractérisation de l'homogénéité d'une texture.

8.2 Run length matrix

La matrice des longueurs de segments (*Gray Level Run Length Matrix, GLRLM*) est une méthode statistique de caractérisation de la texture [Galloway 1975; Haralick 1979; Chu et al. 1990].
Cette méthode effectue le dénombrement des longueurs des segments de pixels de même niveau d'intensité dans une direction donnée et représente les résultats dans une matrice (cf. figure 8.1). On fixe préalablement un nombre de niveaux de gris qui détermine la hauteur de la matrice résultat (même principe que la matrice de cooccurrences, cf. chapitre 7) et une direction θ (0°, 45°, 90° ou 135°). La case (i, j) de la matrice contient le nombre de segments de longueur i et de niveau d'intensité j dans la direction θ. Elle permet de caractériser la grossièreté d'une texture [Haralick 1979], ce qui correspond davantage à la définition d'une texture homogène. Cette méthode paraît mieux adaptée au problème que l'on souhaite résoudre.
Le calcul de la longueur des segments est symétrique (la longueur d'un segment est la même dans

les directions θ et $\theta + \pi$) et il est par conséquent inutile de calculer la matrice dans les quatre directions opposées (180°, 225°, 270° ou 315°). Cette méthode nécessite donc moins de calculs que les matrices de cooccurrences. La figure 8.1 montre un exemple du calcul de la matrice.

Texture

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

$\Longrightarrow$

Gray level	*Run length* (j)			
i	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Figure 8.1. Exemple de remplissage de la matrice de longueur de segments pour une texture 4×4 à quatre niveaux de gris, dans la direction 0°.

Dans [Xu et al. 2004] les auteurs définissent onze indices spécialement conçus pour cette matrice pour construire le vecteur caractéristique de la texture. La liste de ces indices est donnée dans l'annexe B.6.

Pour chacun des trois niveaux de gris 16, 32 et 64 retenus, nous calculons la matrice dans les quatre directions, puis nous faisons la moyenne des quatre matrices résultantes (même principe de fusion des informations que pour les matrices de cooccurrences).
Une recherche exhaustive sur l'ensemble des onze indices de texture utilisables avec les matrices *run length* a permis de déterminer un sous-ensemble composé de sept indices en 32 niveaux de gris et qui obtient un pourcentage de prédiction de 84,81% par régression logistique.
Bien que nous pensions que cette méthode était mieux adaptée à notre problème, elle obtient un résultat nettement inférieur à celui obtenu avec les matrices de cooccurrences et les caractéristiques Haralick (90%).

8.3 Gray Levels Size Zone Matrix (GLSZM)

8.3.1 Présentation

Donc pour répondre de manière vraiment satisfaisante, il nous faut caractériser les textures des noyaux par une description statistique directement sur les régions.
Nous avons élaboré une matrice qui représente le dénombrement des tailles de toutes les régions de texels de même niveau d'intensité. Elle se construit sur le même principe que la *run length matrix*. Elle dénombre non plus des segments d'une longueur donnée, mais les tailles des régions de mêmes niveaux d'intensités.
Pour la remplir, il faut tout d'abord isoler chaque région et cela s'effectue à l'aide d'un algorithme récursif de complexité linéaire. Cet algorithme parcourt l'image et à chaque découverte d'une région non étiquetée, il dépose un germe qui se répand dans la région en l'étiquetant. Une *étiquette* de valeur différente est affectée à chaque région (cf. figure 8.2).

Figure 8.2. Résultat de l'algorithme utilisé pour étiqueter les régions des textures. Les régions sont caractérisées par leur niveau d'intensité, puis étiquetées (chaque étiquette est représentée à l'aide d'une couleur spécifique).

REMARQUE - *L'avantage de cet algorithme est sa complexité en $O(n)$, avec n le nombre de texels. Mais dans la pratique cet algorithme provoque des débordements de la pile d'exécution pour des textures de grandes tailles nécessitant donc un grand nombre d'appels récursifs. Il peut alors être remplacé par un algorithme de type Union-Find [Galler and Fischer 1964] qui opère de manière itérative, mais qui en revanche possède une complexité non linéaire en $O(n\log(n))$.*

Après l'étiquetage, pour chaque région, on incrémente la case de coordonnées [niveau de gris de la région, taille de la région] de la matrice.

Données : La texture T
Résultat : La *size zone matrix GLSZM*
Début
  $Z \leftarrow$ Etiquetage des composantes connexes de T ;
  Pour tous les $z_i \in Z$ **Faire**
    $Tailles[z_i] \leftarrow$ Trouver la taille de z_i ;
    $Niveaux[z_i] \leftarrow$ Trouver l'intensité de z_i ;
  Initialiser la matrice *GLSZM* à 0 ;
  Pour tous les $z_i \in Z$ **Faire**
    Incrémenter$(GLSZM\,[Tailles\,[z_i]\,,Niveaux\,[z_i]])$;
Fin

Algorithme 11 : Remplissage de la GLSZM.

La première dimension de la matrice correspond au nombre de niveaux de gris de l'image (cf. GLCM et GLRLM). Mais la deuxième dimension est dynamique car elle dépend de la taille de la plus grande région.
Plus la texture est homogène, plus la deuxième dimension de la matrice est élevée car les régions de même niveau de gris sont grandes et par conséquent la matrice est du type *matrice creuse*. La figure 8.3 montre deux exemples de remplissage de cette matrice, baptisée *Size Zone Matrix*, pour 4 niveaux de gris.

Texture

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

$\Longrightarrow$

Gray level i	*Size zone* (j) 1	2	3
1	2	1	0
2	1	0	1
3	0	0	1
4	2	0	1

Texture

1	1	3	4
1	3	4	4
3	2	4	4
3	2	1	1

$\Longrightarrow$

Gray level i	*Size zone* (j) 1	2	3	4	5
1	0	1	1	0	0
2	0	1	0	0	0
3	0	0	0	1	0
4	0	0	0	0	1

Figure 8.3. Exemples de remplissage de la GLSZM pour deux textures 4×4 à 4 niveaux de gris.

Cette matrice conçue pour caractériser l'homogénéité possède l'avantage de ne pas nécessiter de calcul dans plusieurs directions ou distances.
En revanche, il est toujours nécessaire de spécifier le nombre de niveaux de gris. Mais réduire le nombre de niveaux de gris permet de réduire un éventuel bruit sur la texture. De plus, il faut effectuer un calcul relativement coûteux pour réaliser l'étiquetage : soit par l'algorithme récursif qui est linéaire, soit par l'algorithme *Union-Find* qui a une complexité plus élevée. Pour l'implémentation récursive, le temps de calcul est comparable à celui des matrices de cooccurrences dans les quatre directions. Pour l'approche *Union-Find* le temps de calcul est environ deux fois plus important.

8.3.2 Modèle de classement de la texture à l'aide des GLSZM

Après avoir construit l'échantillon de travail, nous effectuons une recherche exhaustive pour déterminer le meilleur sous-ensemble d'indices de texture parmi les onze indices utilisés pour la run length matrix (cf. annexe B.6) pour 16, 32 et 64 niveaux de gris.
La meilleure solution est d'utiliser la totalité des indices en 32 niveaux de gris ce qui permet d'obtenir un taux de classement de $91,11\%$ ($45,3\%$ VP et $45,8\%$ VN) avec la régression logistique (un résultat proche est obtenu à l'aide des réseaux de neurones). L'intervalle de confiance est de $[89,1 \ldots 93,1]$ et la probabilité est inférieure à 10^{-4}. Ces résultats montrent que cette méthode de caractérisation de la texture est la meilleure parmi toutes celles testées jusqu'alors.

La figure 8.4 montre la distribution des probabilités engendrées par le modèle. On peut remarquer la forte concentration des probabilités vers les extrémités de l'histogramme et la présence de seulement 29 cas ambigus. Cette répartition illustre la puissance de classement du modèle. On peut cependant remarquer que le modèle commet 6 erreurs graves.

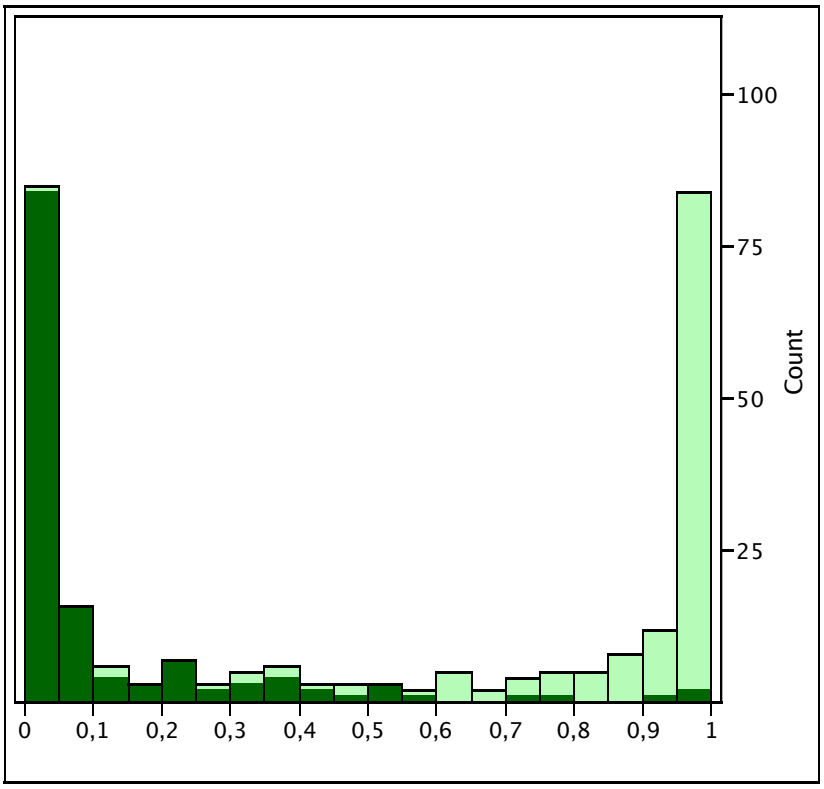


Figure 8.4. Histogramme des distributions des probabilités attribuées par le modèle de classement utilisant la GLSZM et onze indices de texture. En vert clair (resp. vert foncé) les individus à texture homogène (resp. non homogène).

8.3.3 Deux nouveaux indices

Dans notre contexte, les faux négatifs sont des individus à texture homogène avec une probabilité inférieure à $0,5$ (individus en vert clair à gauche de $0,5$) et les faux positifs des individus à texture non homogène avec une probabilité supérieure à $0,5$ (individus en vert foncé à droite de $0,5$).
Nous avons remarqué qu'une majorité de faux positifs ont les mêmes caractéristiques : ils possèdent de grandes zones homogènes, mais avec de grandes variations d'intensité entre les régions (cf. figure 6.4*b*), donc une forte inertie inter-classes. En effet, une faible inertie inter-régions constitue une des propriétés des textures homogènes (cf. section 6.3).
Mais on peut constater qu'aucun des onze indices de texture ne caractérise l'inertie inter-régions. Au mieux, certains indices caractérisent la présence de grandes zones ou de zones de forte intensité.

L'inertie inter-régions peut être évaluée par l'écart type des niveaux d'intensité des régions. On doit également prendre en compte la taille des régions. Pour cela, nous introduisons un nouvel indice qui est basé sur l'écart type des niveaux de gris pondéré par les tailles des zones :

$$\psi_N = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}(n*M(n,s)-\mu_N)^2} \text{ avec } \mu_N = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} n*M(n,s)$$

avec N la première dimension de la matrice (nombre de niveaux de gris), S la deuxième dimen-

sion (taille de la plus grande région), $M(n,s)$ l'élément de la matrice de coordonnée (n,s) et μ_N la moyenne pondérée des éléments de la matrice.
Plus les régions sont grandes, plus elles influent sur la valeur de l'indice et plus l'inertie inter-régions est importante, plus la valeur de cet indice est élevée.

Par analogie, nous proposons également un deuxième indice qui est conçu de la même manière mais pour caractériser les variations entre les tailles des régions.

$$\psi_S = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}(s*M(n,s)-\mu_S)^2} \text{ avec } \mu_S = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} s*M(n,s)$$

En revanche, aucune structure particulière n'a pu être révélée lors de l'étude des faux négatifs.

8.4 Analyse mono-variable

Nous avons effectué les mêmes étapes d'analyse que celles effectuées pour la forme (cf. section 5.3) en commençant par : la distribution, les *outliers*, le classement. Mais aucun élément, individu ou phénomène particulier n'est apparu durant ces analyses.

Seule l'analyse des corrélations a révélé certaines propriétés (cf. figure 8.5) :
- Corrélation entre LGLRE et SRLGLE, avec un coefficient de corrélation de $0,9965$ et une probabilité inférieure à 10^{-4}. Ce sont deux indices permettant de caractériser la présence importante de zones de faibles niveaux de gris.
- HGLRE et SRHGLE qui sont deux indices caractérisant les textures ayant des zones de haute intensité. Le coefficient de corrélation est de $0,9959$ avec une probabilité inférieure à 10^{-4}.

8.5 Modèle final de caractérisation de la texture

L'échantillon de travail est reconstruit à partir d'un ensemble composé maintenant de treize indices. Une nouvelle recherche exhaustive est effectuée pour 16, 32 et 64 niveaux de gris. Comme cela a été effectué pour la forme des noyaux (cf. section 5.4), nous testons différents nombres de neurones ($\nu = 2\ldots6$) de la couche cachée du réseau de neurones ainsi que divers paramètres de k (fixe ou dépendant du nombre d'indices N_i) pour les k-plus proches voisins. La figure 8.6 et le tableau 8.1 illustrent les meilleurs résultats obtenus.

La meilleure configuration du réseau de neurone est d'utiliser $\nu = 2$. Les résultats sont supérieurs lorsque le nombre d'indices est faible, mais sont inférieurs lorsque l'on utilise plus de 8 indices. Toutefois, les écarts de prédiciton sont de l'ordre d'un à deux noyaux au maximum. Cela provient de la difficulté à ajuster les poids du réseau avec le peu d'individus dont nous disposons, ce qui engendre un apprentissage par cœur des données. Les k-plus proches voisins apportent leur meilleur résultat pour k égal au nombre d'indices additionné de 13, mais les résultats sont nettement inférieurs à ceux du réseau de neurones.

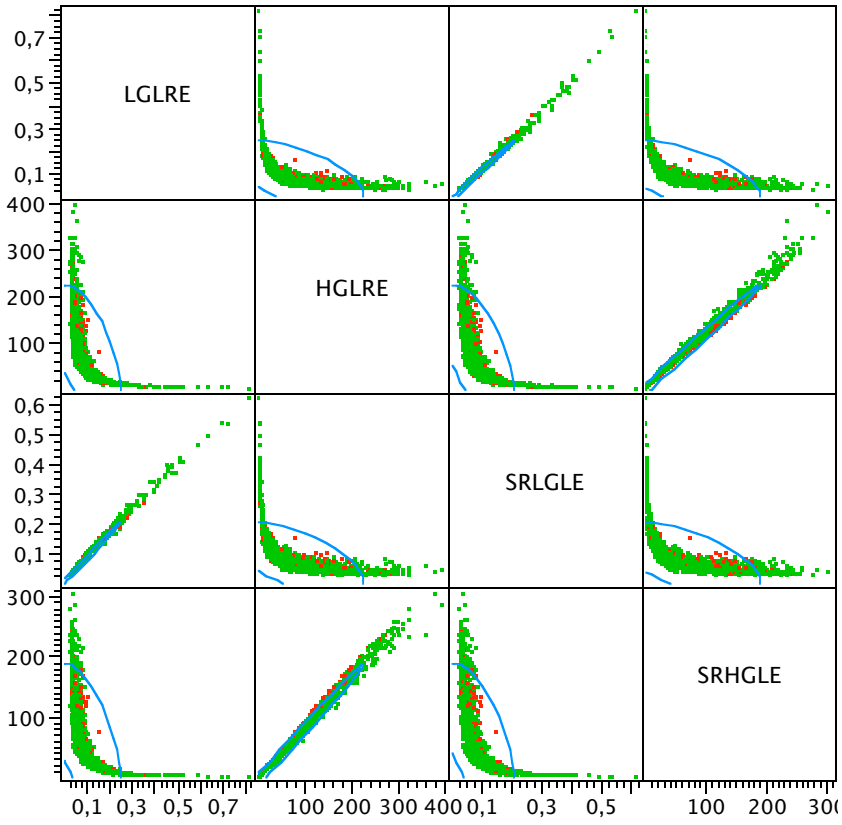


Figure 8.5. Illustration des corrélations entre quatre indices de texture. Les noyaux à texture homogène (resp. non homogène) sont en vert (resp. rouge). L'ellipse (en bleu) englobe 95% des noyaux sous l'hypothèse de binormalité.

Méthodes / Nombre d'indices	$N_i + 13$-PPV	RL	FA	PMC / 2
1	66,66	67,03	66,74	68,14
2	72,96	80,74	80	75,55
3	75,18	85,55	80,74	87,4
4	75,55	89,25	82,96	88,88
5	76,66	89,25	82,59	91,48
6	76,66	89,25	84,07	91,48
7	78,51	91,11	84,44	91,48
8	77,4	91,11	84,04	92,22
9	77,03	92,59	82,96	91,11
10	76,66	92,22	81,48	91,48
11	76,66	91,85	81,48	91,85
12	74,81	94,07	81,48	90,37
13	73,7	92,96	78,14	86,66

Table 8.1. Pourcentages de prédiction obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement de la texture.

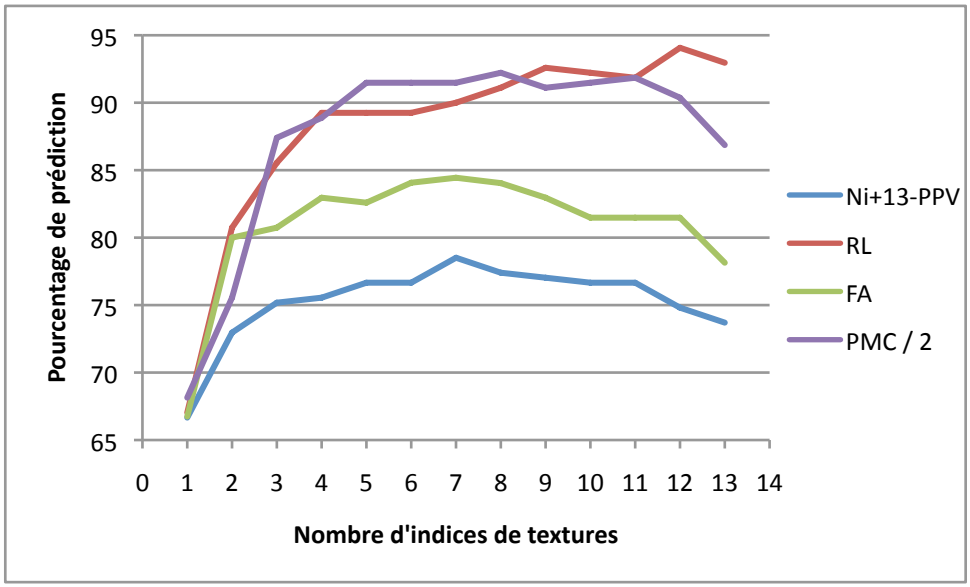


Figure 8.6. Comparaison graphique des performances des différents classifieurs appliqués au problème de la texture. En abscisse le nombre d'indices de texture utilisés et en ordonnée le pourcentage de prédiction obtenu.

La régression logistique obtient le meilleur résultat avec un sous-ensemble composé de douze indices[1] et apportant un pourcentage de prédiction de la texture des noyaux de $94,07\%$ (47% VP et 47% VN) sur l'échantillon de travail (cf. tableau 8.1). L'intervalle de confiance est de $[92,2 \ldots 96,2]$ et la probabilité du sous-modèle est inférieure à 10^{-4}.
L'utilisation des deux indices que nous avons construits permet d'améliorer la prédiction du sous-modèle de près de 3%. Mais il existe une intersection entre l'intervalle de confiance de ce dernier sous-modèle et celui du précédent n'utilisant pas nos deux indices. Une étude de rang en fonction de la médiane révèle que la probabilité du sous-modèle précédent d'apporter des résultats équivalents au nouveau sous-modèle est de $0,087$. De même, une analyse de la variance (par le test de Wilcoxon [Tufféry 2007; Wonnacott and Wonnacott 1998]) des pourcentage de prédiction utilisés pour le calcul des intervalles de confiance de chaque sous-modèle, révèle que la probabilité de ces deux distributions d'être issues d'une même loi est inférieure à 10^{-4}. Ces trois informations montrent la nécessité d'utiliser nos deux indices.
Le gain apporté au sous-modèle par nos indices peut être observé sur la figure 8.7 qui montre la distribution des probabilités attribuées par le modèle. La forte répartition sur les extrémités de l'histogramme montre l'efficacité de classement et la pertinence dans le choix des indices. On peut remarquer une nette amélioration de l'augmentation des distributions des probabilités vers les valeurs extrêmes et la faible présence de cas ambigus (seulement 16 individus).

REMARQUE - *Les deux couples d'indices fortement corrélés sont utilisés dans le modèle. Mais supprimer un indice de chaque couple ne fait pas chuter de manière significative les performances du sous-modèle.*

[1] Seul l'indice LRHGLE n'est pas utilisé dans le sous-modèle.

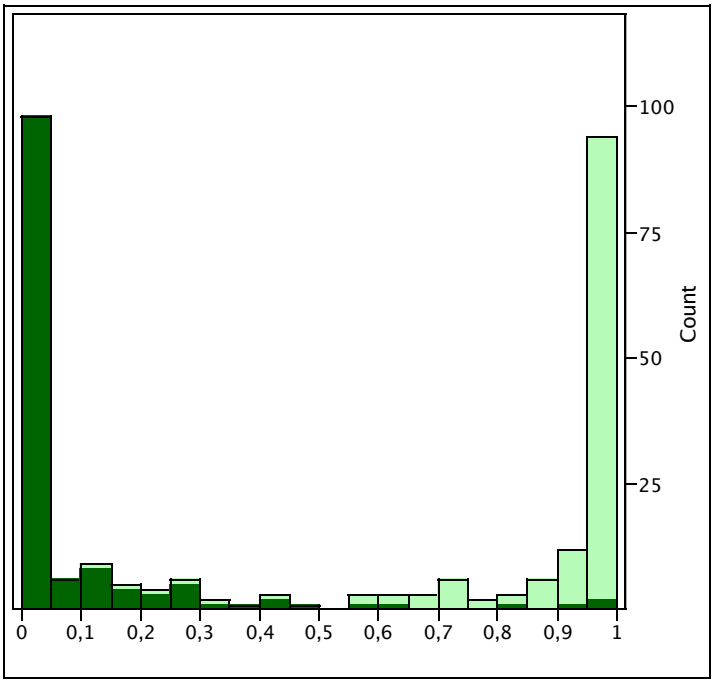


Figure 8.7. Distribution des probabilités de classement attribuées par le modèle utilisant la GLSZM avec douze indices. En vert foncé (resp. vert clair) les noyaux à texture non homogène (resp. homogène). La très grande majorité des noyaux sont classés avec des probabilités proches des extrêmes et il existe seulement 16 cas ambigus.

8.6 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode de caractérisation statistique de l'homogénéité de la texture. Cette méthode effectue un dénombrement des régions en fonction de leur taille, construit une matrice représentative de la texture, puis extrait les caractéristiques à l'aide d'indices. Parmi ces indices, on peut maintenant en compter deux nouveaux qui ont apporté une amélioration significative du pourcentage de prédiction d'environ trois points. Cette contribution spécifiquement créée pour répondre au sous problème de classement de la texture apporte un pourcentage de prédiction de 94% sur l'échantillon de travail, ce qui est un taux bien supérieur à toutes les méthodes qui ont été essayées. Ces résultats ont donné lieu à deux publications [Thibault et al. 2008b; Thibault et al. 2009].

Maintenant que nous possédons un modèle de classement de la texture (cf. figure 8.8), il est nécessaire de l'introduire dans le modèle de classement des noyaux afin de répondre au problème final qui est de classer les noyaux dans les classes *Sains* ou *Pathologiques*.
En combinant le sous-modèle de classement de la forme et celui que nous venons de créer, nous obtenons un taux de classement des noyaux de $87,6\%$. Ce taux est bien inférieur à ceux obtenus pour les sous-problèmes[2]. De plus la forme seule permet de répondre à $86,9\%$ au problème final. Donc comment expliquer cette faible amélioration de l'ordre du pourcent apportée par le modèle de classement de la texture ?

[2] $95,4\%$ pour la forme et $94,07\%$ pour l'homogénéité de la texture, sur leurs échantillons de travail respectifs.

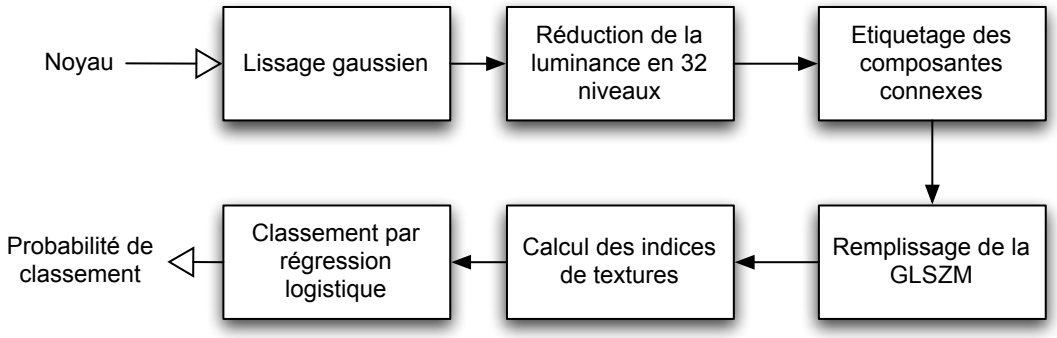


Figure 8.8. Schéma récapitulatif des différentes étapes nécessaires à la construction du sous-modèle de classement de l'homogénéité de la texture.

Les réponses se trouvent une fois de plus dans l'étude des individus et ce résultat de classement décevant s'explique par deux raisons :

1. La première est l'intersection qui existe entre l'ensemble des noyaux à texture non homogène et les noyaux ayant une forme boursouflée. Cette intersection avait été constatée lors de l'étude de la distribution des individus (cf. section 3.2). En effet, parmi les 135 noyaux à texture non homogène, plus d'une centaine ont une forme anormale et ont déjà été classés comme pathologiques lors de l'analyse de leur forme. Donc même si le sous-modèle de classement de la texture permettait de classer parfaitement les noyaux par l'homogénéité de leur texture, il ne serait utile que pour une trentaine d'individus. Etant donné que nous disposons d'environ 3300 noyaux, cette amélioration de $0,7\%$ correspond bien à une vingtaine d'individus bien classés.
2. La deuxième explication s'obtient en analysant l'expertise des noyaux. Parmi tous les noyaux dont nous disposons, seulement 94% peuvent être classés en fonction de leurs formes et/ou de leurs textures. Donc les $87,6\%$ obtenus le sont sur les 94% que nous pouvons caractériser. Ainsi dans sa forme actuelle, le modèle classe $93,2\%$ des noyaux qui peuvent l'être. Ce pourcentage est en fait très satisfaisant.

Malgré le faible gain apporté par le sous-modèle de classement de la texture, ce dernier ne pouvait améliorer davantage le classement final. Donc le gain obtenu est significatif et important.
Pour améliorer de manière encore plus significative le modèle, il nous faut désormais caractériser les éléments représentant les 6% des individus qui peuvent être caractérisés uniquement par la détection de trous ou de focis dans leur texture (cf. chapitre 1).

CHAPITRE 9

CARACTÉRISATION DE TEXTURE PAR INDICES DE FORME 3D

9.1 Introduction

Dans le chapitre 3, il est démontré qu'une analyse des trous et des focis présents dans la texture est nécessaire pour améliorer le classement des noyaux : ces éléments interviennent dans le diagnostic d'un ensemble de noyaux qui ne peuvent en aucun cas être classés par leur forme ou l'homogénéité de leur texture. Donc ces noyaux ne peuvent être classés par les précédents sous-modèles malgré les bons résultats de ces derniers.
Ce chapitre présente les différentes étapes de la construction de plusieurs nouveaux sous-modèles pour parvenir au classement des noyaux en fonction de la présence des trous et des focis qu'ils contiennent. Dans un premier temps nous présentons des travaux originaux d'analyse de texture basés sur les indices de forme. L'analyse et la résolution des problèmes détectés dans ces travaux nous permettent de proposer une succession d'étapes qui effectuent la caractérisation d'éléments de texture. Pour cela nous détaillons le type de représentation de la texture que nous utilisons ainsi que la procédure d'extraction des éléments de la texture que nous souhaitons classer. Ces éléments sont ensuite caractérisés à l'aide d'indices de forme 3D issus de la 2D ou créés pour répondre de manière plus spécifique au problème.

9.2 Indices de forme 2D pour la caractérisation de texture

Les indices de forme ont montré leur efficacité dans la construction du sous-modèle de classement de la forme (chapitre 5). Cependant dans [Chen et al. 1995] les auteurs ont développé une approche qui permet de les utiliser pour caractériser des textures. Pour ce faire, ils considèrent une image I comme la somme des résultats de tous les seuillages binaires :

$$I(x,y) = \sum_{\alpha=1}^{N} f_b(x,y,\alpha) \text{ , avec } f_b(x,y,\alpha) = \begin{cases} 1 \text{ si } I(x,y) \geqslant \alpha \\ 0 \text{ sinon} \end{cases}$$

avec N le nombre de niveaux de gris de l'image.
Soit E_α, l'ensemble des composantes connexes issues du seuillage de I pour un seuil $\alpha \in [1,N]$.

Pour chaque valeur de α et pour un indice de forme χ, on peut calculer :

$$\chi_\alpha = \frac{\sum_{x \in E_\alpha} [S(x) * \chi(x)]}{\sum_{x \in E_\alpha} S(x)} \text{ , avec } S(x) \text{ la surface de } x$$

On obtient ainsi N valeurs par indice et les auteurs calculent alors quatre indices globaux qui constituent les caractéristiques de la texture pour l'indice χ :

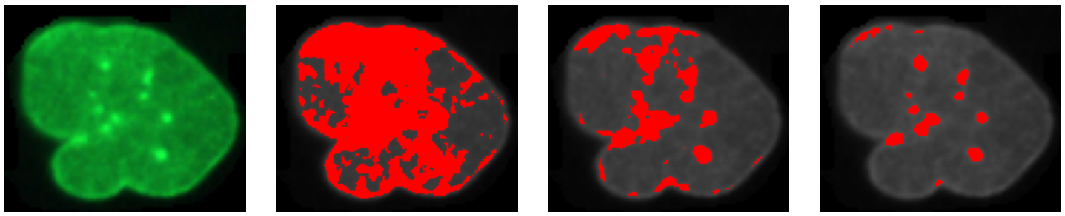
- max
$$\max_\alpha \chi_\alpha$$
- moyenne
$$\frac{1}{N} \sum_\alpha \chi_\alpha$$
- moyenne pondérée par les valeurs de α
$$sm = \frac{1}{\sum_\alpha \chi_\alpha} \sum_\alpha \alpha \chi_\alpha$$
- écart type de l'échantillon
$$\sqrt{\frac{1}{\sum_\alpha \chi_\alpha} \sum_\alpha (\alpha - sm)^2 \chi_\alpha}$$

Nous avons utilisé cette technique avec les onze indices de forme employés dans le modèle de caractérisation de la forme des noyaux. Dans cette approche chaque indice de forme apporte quatre valeurs lors du calcul des indices globaux. Donc chaque noyau possède un vecteur caractéristique de dimension 44 ce qui engendre un problème d'apprentissage par cœur.
Cette technique a tout d'abord été testée pour caractériser l'homogénéité de la texture. Mais déterminer le meilleur sous-ensemble d'indices par recherche exhaustive nécessite de tester 2^{44} ($\simeq 8.10^{12}$) combinaisons. Ceci n'est pas réalisable dans un temps raisonnable. Une méta-heuristique de type *tabou* [Glover 1986; Glover 1990; Michel and Hentenryck 2004] a été utilisée pour déterminer le meilleur sous-ensemble d'indices à utiliser. Le résultat est un sous-ensemble composé de 28 indices apportant un taux de bon classement de 85% par régression logistique.
Nous avons également testé cette méthode pour déterminer si un noyau contient un nombre de focis (resp. de trous) suffisamment importants pour être pathologique. Pour cela nous construisons un échantillon de travail contenant les 123 (resp 111) noyaux rendus pathologiques par leur nombre de focis (resp. trous), auxquels nous avons ajouté un nombre égal de noyaux sains sélectionnés par k-moyennes initialisées par formes fortes (cf. algorithme 5). Des recherches avec tabou ont été réalisées, mais aucun sous-ensemble d'indices n'a permis de classer correctement les noyaux et ce pour les quatre méthodes de classement utilisées.

Cette technique examine la texture selon N niveaux de gris. Il en résulte N textures différentes. Mais lorsque l'on observe le résultat du seuillage pour un niveau donné (cf. figure 9.1), on s'aperçoit que la ou les formes à caractériser sont extrêmement complexes :

- présence de trous
- aucune connaissance a priori sur les formes à caractériser
- formes totalement irrégulières
- formes très petites, parfois constituées de seulement quelques pixels

Cette vision 2.5D d'une texture est difficilement utilisable même avec la puissance et la souplesse des indices de forme. Toutefois, nous avons pensé pouvoir généraliser cette idée, non plus



(a) (b) (c) (d)

Figure 9.1. Résultats des seuillages (en rouge) pour un noyau de cellule (a) avec différents seuils : 60 (b), 70 (c) et 80 (d).

en considérant une texture comme une somme de textures obtenues par seuillages, mais comme un volume constitué de l'empilement de ces seuillages : représenter une texture par son volume sous la nappe.

9.3 Indices de forme 3D pour la caractérisation de texture

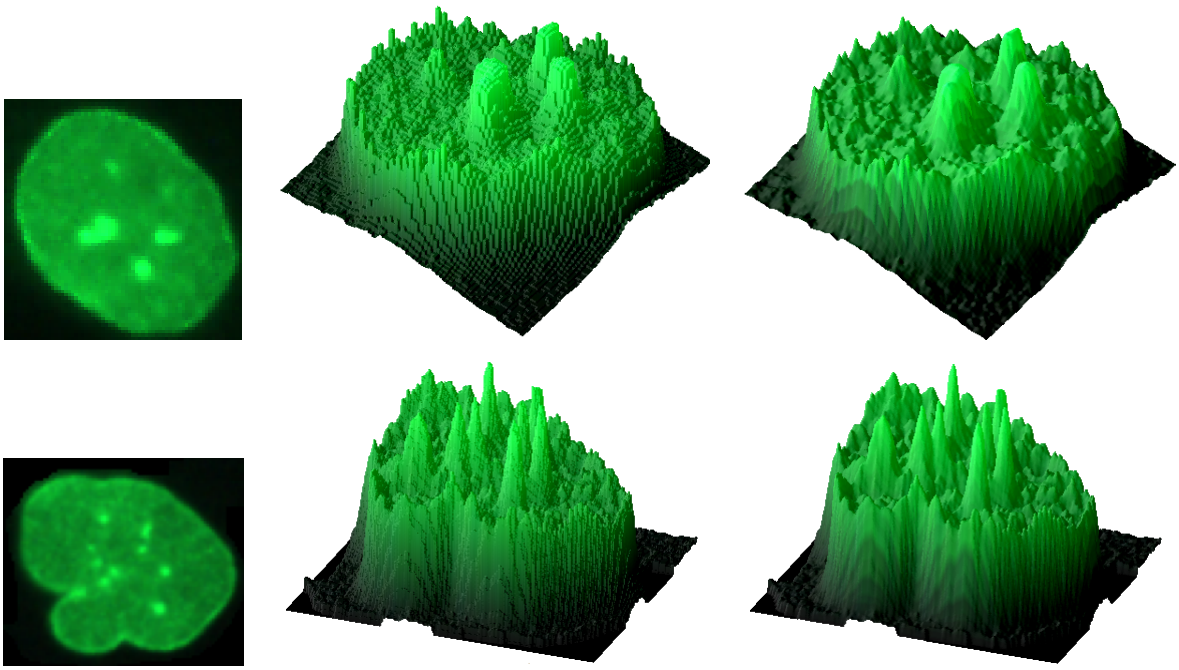
9.3.1 Le volume sous la nappe

Une image peut être considérée comme un tableau à deux dimensions dont chaque case contient la valeur de l'intensité d'un pixel. Mais on peut également considérer une image comme une carte d'élévation : la valeur de chaque case ne représente plus l'intensité d'un pixel, mais une altitude (cf. figure 9.2).

Cette représentation des images permet de transformer les variations d'intensité en variations d'altitude qui génèrent une multitude d'éléments : des *pics* de différentes formes (cylindrique, conique, etc.), des *lacs*, des *crêtes*, des *failles*, etc.
Il est alors nécessaire de disposer d'une méthode d'extraction capable de découper les éléments souhaités. Les capacités de cette méthode dépendent des éléments extraits et donc des caractéristiques de la texture que l'on souhaite analyser (texture directionnelle, structurelle, etc.).
Une fois les éléments isolés du reste du volume sous la nappe, il est alors possible de les caractériser. Mais la grande diversité des éléments que l'on peut extraire nécessite l'utilisation d'une méthode de caractérisation capable de s'adapter et d'être utilisable dans un classifieur.

9.3.2 Indices de forme : de la 2D vers la 3D

Pour caractériser les éléments (volumes) extraits, nous utilisons des indices de forme, mais dans une nouvelle version spécifique aux espaces à trois dimensions. On pourrait ainsi appeler les indices de forme 3D, des *indices de volume*.
Dans le chapitre 5, il est expliqué que les indices de forme sont des fonctions multi-variables. Les variables utilisées pour construire les indices sont appelées des *mesures*. Donc l'étape préliminaire dans la construction d'un indice de forme 3D consiste à extraire des mesures spécifiques à la 3D et



(a) (b) (c)

Figure 9.2. Exemples de deux noyaux de cellules (a), de leur volume sous la nappe (b) et de leur volume sous la nappe lissé (c) par un filtrage Gaussien (qui améliore la visibilité du relief).

à étudier le comportement de certaines mesures 2D lors du passage vers la 3D.

Toutes les mesures ne se comportent pas de la même façon lors du changement de dimension et on peut les classer en trois catégories, selon qu'elles :

- sont différentes car elles changent de dimension.
- offrent de nouvelles informations.
- ne sont pas modifiées.

Seulement deux mesures parmi celles utilisées dans la construction des indices de forme 2D connaissent un changement de dimension :

- le périmètre devient la surface.
- la surface devient le volume.

En revanche, l'axe principal possède désormais deux axes secondaires dont les longueurs sont notées $L_{AP\perp 1}$ et $L_{AP\perp 2}$.

Certaines familles de mesures ne subissent aucune modification lors du changement de dimension :

- Les mesures basées sur la distance (dimension 1) entre deux points : les diamètres, les épaisseurs, les rayons, etc. Dans ce cas, seul le calcul est modifié ; par exemple le calcul du plus grand rayon ne s'effectue plus entre le barycentre et un point du périmètre, mais entre le barycentre et un point de la surface. Ou encore le rayon du plus petit (resp. grand) disque circonscrit (resp. inscrit) à la forme est désormais le rayon de la plus petite (resp. grande)

boule circonscrite (resp. inscrite) au volume.
– Les mesures sans dimension qui sont déjà, par définition, des indices de forme (le nombre de trous N_{Trous}, le nombre de composantes connexes d'écart N_{Cce}, etc.).

Par définition les indices de forme doivent être sans dimension (cf. définition 5.1.1). Or le changement de dimension de certaines mesures lors de leur extension en 3D, fait perdre cette propriété. Un exemple est donné pour le déficit isopérimétrique dans le cas d'une boule :

$$\text{Déficit isopérimétrique} = 4\pi\frac{A}{P^2}$$

$$\Rightarrow \text{Déficit isosurfacique} = 4\pi\frac{V}{S^2}$$

$$\text{Déficit isosurfacique}(boule) = 4\pi\frac{\frac{4}{3}\pi R^3}{(4\pi R^2)^2} = \frac{1}{\pi R}$$

Il est donc nécessaire de modifier les puissances afin de respecter la définition et les coefficients pour conserver les propriétés de l'indice :

$$\text{Déficit isosurfacique} = 36\pi\frac{V^2}{S^3}$$

$$\text{Déficit isosurfacique}(boule) = 36\pi\frac{(\frac{4}{3}\pi R^3)^2}{(4\pi R^2)^3} = 1$$

La liste complète des versions 3D des indices de forme 2D utilisés dans ce manuscrit est en annexe B.4 de ce manuscrit.
Nous avons vu dans le chapitre 5 que la souplesse des indices de forme les rend particulièrement efficaces pour la caractérisation. En effet il est possible de créer des indices pour répondre spécifiquement au problème de caractérisation traité. La capacité d'adaptation de cette méthode en fait un choix judicieux pour une utilisation dans l'étape de caractérisation du procédé que nous présentons.

9.3.3 Les étapes du procédé utilisé pour l'analyse de la texture

Les sections précédentes ont décrit deux étapes importantes du travail que nous présentons : le mode de représentation de la texture et la technique de caractérisation des éléments. Cette section présente la totalité des étapes que nous mettons en œuvre pour réaliser l'analyse de la texture. Ces étapes sont représentées dans la figure 9.3.

Décrivons les différentes étapes dans l'ordre de réalisation :
– Traitement(s) préliminaire(s) : élimination des informations qui ne sont pas en rapport avec la problématique et/ou tous les traitements nécessaires pour faciliter l'extraction des éléments (suppression du bruit, étalement d'histogramme, modification du contraste, gradients, etc.). Il est possible qu'aucun traitement ne soit nécessaire, donc cette étape est optionnelle. Durant ce travail préliminaire, il est tout à fait envisageable d'utiliser une méthode existante de segmentation. Cette méthode permettrait de détecter et filtrer les éléments que l'on souhaite caractériser avant leur transformation en volume sous la nappe.
– 1. Transformation de la texture en volume sous la nappe.
– 2. Extraction des éléments à caractériser : dans cette étape le choix de la méthode dépend des éléments que l'on souhaite extraire. Le résultat de cette étape est fortement lié aux traitements préliminaires.

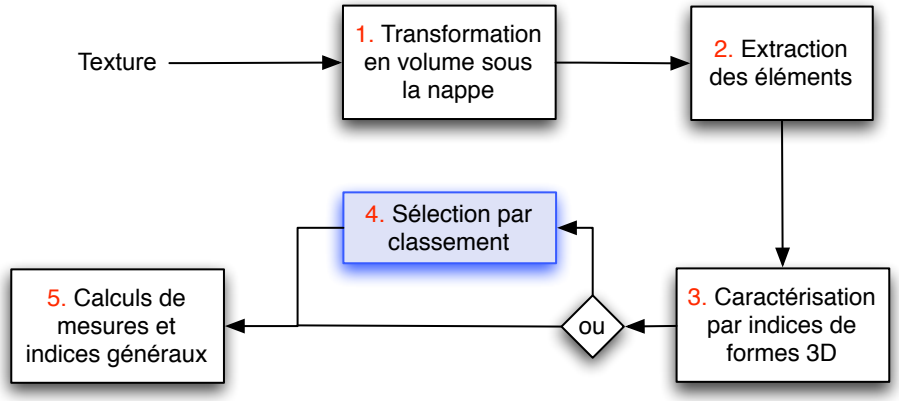


Figure 9.3. Schéma récapitulatif de la description statistique de texture par transformation en volume sous la nappe et caractérisation par indices de forme 3D. L'étape 4 (en bleu) est facultative.

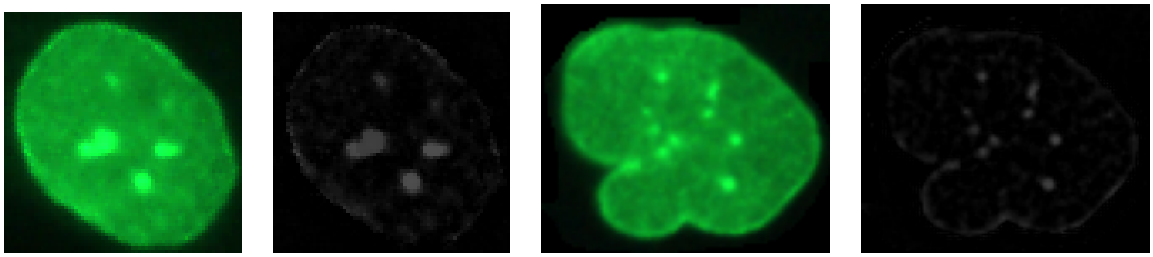
- 3. Caractérisation par indices de forme 3D de tous les éléments extraits. On vient de voir que cette technique est très bien adaptée au problème présent grâce notamment à sa souplesse : il est possible de créer un ou plusieurs indices de forme 3D pour caractériser les éléments de texture que l'on extrait.
- 4. Sélection par classement : cette étape est optionnelle dans la méthode, mais elle peut améliorer le résultat final. Les éléments extraits ne sont pas nécessairement ceux souhaités (erreurs d'extraction). Mais tous les éléments ont été caractérisés par indices de forme et il est alors possible de construire un modèle de classement des éléments extraits. Ce classement permet de sélectionner les éléments afin de corriger les erreurs de la méthode d'extraction.
- 5. Tous les éléments extraits ont été caractérisés par indices de forme. Malheureusement, on ne peut construire le vecteur caractéristique de la texture étudiée avec toutes les caractéristiques des éléments extraits. En effet, le nombre d'éléments varie entre chaque noyau, donc la taille du vecteur caractéristique est dynamique, ce qui n'est généralement pas supporté par les classifieurs. Ainsi il est nécessaire de construire des variables globales qui permettent de donner des informations pertinentes et représentatives sur la population des éléments étudiés. Ces informations peuvent être le nombre d'éléments, la moyenne et/ou l'écart type de leurs caractéristiques, etc. Toutes ces informations statistiques permettent de décrire la texture d'intérêt.

9.4 Application au classement des noyaux

Nous utilisons le procédé qui vient d'être présenté afin de classer les noyaux en fonction de la présence des trous et des focis dans la texture des noyaux. Cette section décrit l'application des différentes étapes et en particulier celles laissées au choix de l'utilisateur : la méthode d'extraction et le modèle de filtrage des éléments. Lorsque nous explicitons ces choix, nous donnons les raisons qui font que ce procédé est particulièrement bien adapté au problème.

9.4.1 Etape 1 : représentation par volume sous la nappe

Les trous (resp. les focis) sont des zones plus sombres (resp. claires) dans la texture des noyaux. Lorsque l'on souhaite détecter de telles zones, on utilise généralement des méthodes de type morphologie mathématique [Serra 1982; Matheron and Serra 2002] et plus particulièrement les *chapeaux haut de forme* (*top hat*) [Netten et al. 1996; Jalba et al. 2004; Jalba et al. 2006].
Mais le résultat de toute opération de morphologie mathématique dépend du choix de l'élément structurant qui se compose de deux paramètres : le type (forme de l'élément) et la taille[1]. Dans le cas des focis, un élément structurant de taille trop petite ne détecte aucun foci, mais en revanche un élément de taille trop importante détecte la moindre variation d'intensité dans la texture. Le choix de l'élément structurant représente un inconvénient majeur dans notre problème (cf. figure 9.4) car les focis sont de taille et d'intensité différentes. La figure 9.4 montre le résultat de la détection de focis à l'aide d'outils de morphologie mathématique.



(a) (b) (c) (d)

Figure 9.4. (a) et (c) deux noyaux contenant des focis, (b) et (d) le résultat de la transformation de type *White Top Hat* avec un élément structurant d'ordre 5 et de type disque.

On peut observer sur les images *(c)* et *(d)* de la figure 9.4 que certains focis ne sont composés que de quelques pixels. Ce faible nombre de pixels rend extrêmement difficile l'utilisation des indices de forme ou de tout autre méthode de caractérisation. En effet, plus les formes sont petites, plus l'erreur de discrétisation est importante et donc plus la caractérisation est hasardeuse. Par exemple un disque de rayon 1 est une croix dans un espace discret (figure A.1*a*) et la surface est égale au périmètre, ce qui n'est pas le cas pour un disque dans un espace continu (excepté si $R = 2$). Ce problème peut être observé en annexe sur la figure B.3 relative aux mesures des erreurs entre les valeurs théoriques et réelles du déficit isopérimétrique : plus le rayon est petit, plus l'erreur est importante. Utiliser le volume sous la nappe d'un foci permet d'ajouter une dimension et ainsi d'augmenter la taille (le volume en trois dimensions) afin d'utiliser les indices de forme 3D. Cela permet de représenter l'intensité de tous les pixels constituant le foci.

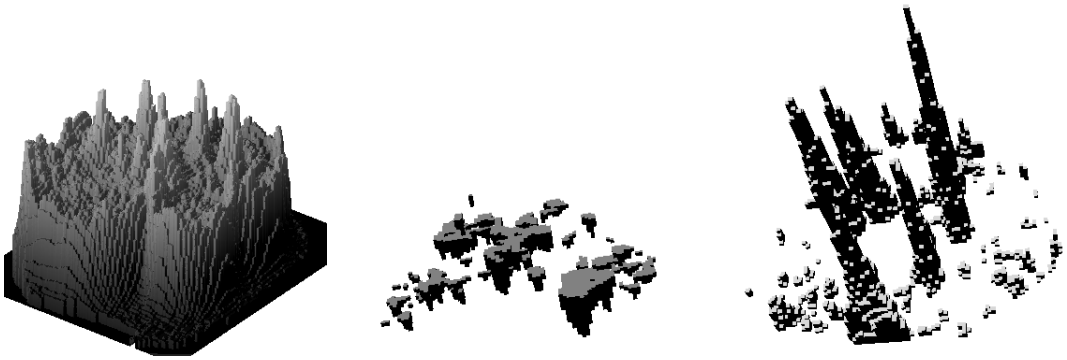
9.4.2 Etape 2 : extraction des focis et des trous

Les deux éléments que l'on souhaite caractériser sont les trous et les focis. Grâce à la représentation par volume sous la nappe, il est facile de les isoler : les focis sont des pics et les trous des lacs. Pour cela nous utilisons une procédure d'extraction de tous les pics et les lacs (cf. figure 9.5) qui opère de la façon suivante :

- détection de tous les maximums (resp. minimums) locaux.
- propagation récursive aux couches inférieures (resp. supérieures) pour trouver tous les niveaux du pic (resp. lac).

[1] Dans la littérature on trouve également le terme *ordre* pour parler de la taille

– arrêt lorsque l'on rencontre un autre pic (resp. lac).

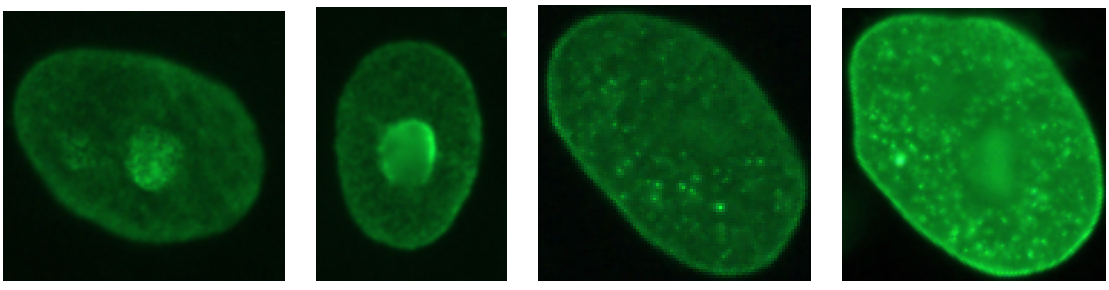


(a) (b) (c)

Figure 9.5. Extraction des lacs (b) et des pics (c) pour un volume sous la nappe (a) issu du noyau de la figure 9.4*c*. Tous les lacs et les pics ont été ramenés à un même niveau d'altitude lors de l'affichage.

La figure 9.5 montre le résultat de l'extraction des pics et des lacs pour le noyau *(c)* de la figure 9.4. Or on peut remarquer que ce noyau ne contient aucun trou (cf. figure 9.4*c*). Mais la procédure détecte des lacs (cf. figure 9.5*b*). Donc parmi tous les lacs extraits aucun n'est un trou. Il est par conséquent nécessaire de construire un modèle pour classer les lacs dans les classes : *trou* et *non trou*. Ce modèle de classement constitue le cœur de l'étape 5 relative au filtrage des éléments (cf. section 9.3.3). Le problème se pose également pour les pics extraits qui ne sont pas tous des focis. Mais dans le cas des focis, certains éléments de la texture posent des problèmes supplémentaires :

– les précipités (cf. figure 9.6 *c* et *d*) : ils sont dus à un défaut du marqueur fluorescent qui précipite et forme des amas. Ce sont des tâches lumineuses comparables aux focis car aussi intenses mais toutefois plus petites.
– la périphérie : dans le cas des noyaux à périphérie marquée mais non régulière, les variations de la périphérie forment des pics. Toutefois, la forme est plus allongée et la variation d'intensité moins importante. Cette notion d'allongement nécessite l'utilisation d'au moins un des indices d'allongement dans le modèle de filtrage.
– les artefacts de marquage (cf. figure 9.6 *a* et *b*) : ils sont dus à une mauvaise répartition du marqueur.



(a) (b) (c) (d)

Figure 9.6. (a) et (b) deux noyaux contenant un artefact de marquage de grande taille, (c) et (d) deux noyaux contenant des artefacts de marquage de taille inférieure (des précipités).

Tous ces éléments sont extraits parmi les pics et ressemblent fortement à des focis. Il est par conséquent nécessaire de trouver des indices de forme 3D pour caractériser les différents éléments extraits du volume sous la nappe et ainsi permettre la construction d'un modèle de classement des pics pour filtrer et corriger les erreurs d'extraction.

9.4.3 Etape 3 : nouveaux indices de forme 3D

Une méthode d'extraction vient d'être présentée et nous disposons désormais de deux ensembles de volumes : les pics et les lacs. Mais tous les lacs et les pics extraits du volume sous la nappe ne sont pas nécessairement des trous et des focis. Pour cette raison il faut construire deux modèles de classement qui permettent de filtrer et corriger une partie des erreurs. Pour parvenir à la construction de ces modèles, il faut utiliser des indices de forme 3D qui décrivent les lacs et les focis afin de les isoler des autres éléments dans l'espace des caractéristiques.

En observant la forme des focis dans leur représentation à l'aide des volumes sous la nappe, on peut constater qu'elle est proche de celle d'un cylindre. De même pour les précipités, la forme est proche de celle d'un cône. Il est par conséquent nécessaire de construire des indices de forme 3D qui permettent de caractériser ces deux primitives afin de pouvoir caractériser efficacement les volumes que l'on souhaite classer.

Afin de construire des indices caractérisant des cylindres, on utilise les égalités inhérentes à cette forme (procédure déjà utilisée lors de la construction des indices caractérisant des ellipses dans la section 5.2.1).
Dans un cylindre, le volume est égal à $V = B \times H$ avec B l'aire de la base qui est un disque (donc $B = \pi R^2$ avec R le rayon) et H la hauteur. L'idée est de remarquer qu'un cylindre possède un axe de symétrie confondu avec l'axe principal, dont la longueur est égale à la hauteur du cylindre. De plus, les axes secondaires (orthogonaux à l'axe principal) sont de longueur égale au diamètre de la base du cylindre. On peut également montrer que le plus petit rayon est égal au rayon de la base. Cette dernière égalité permet d'utiliser le théorème de Pythagore pour calculer la longueur du plus grand rayon en fonction du plus petit rayon et de la moitié de la hauteur. Tout ceci peut s'observer sur la figure 9.7. On obtient les égalités suivantes :

- $H = L_{AP}$
- $L_{AP\perp 1} = L_{AP\perp 2} = R_{min} = 2R$
- $H = 2\sqrt{R_{max}^2 - R_{min}^2}$

Un cylindre doit vérifier les égalités suivantes :

$$V = \frac{\pi}{4} L_{AP} L_{AP\perp 1} = \frac{\pi}{4} L_{AP} L_{AP\perp 2} = 2\pi R_{min}^2 \sqrt{R_{max}^2 - R_{min}^2}$$

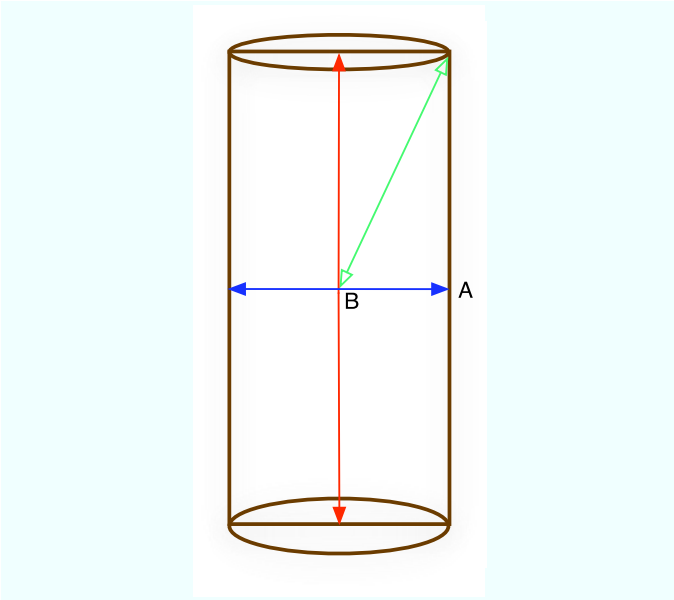


Figure 9.7. Illustration de la position des mesures sur un cylindre : en rouge l'axe principal, en bleu les axes orthogonaux à l'axe principal ainsi que le plus petit rayon (de *A* à *B*) et en vert le plus grand rayon.

On peut ainsi construire trois indices de volume permettant de caractériser des cylindres, par :

Les rayons

$$\Psi_{CylindreR} = 2\pi \frac{R_{min}^2}{V} \sqrt{R_{max}^2 - R_{min}^2} \in [0,1]$$

L'axe principal 1

$$\Psi_{CylindreAP1} = \frac{\pi}{4} \frac{L_{AP} L_{AP\perp 1}^2}{V} \in [0,1]$$

L'axe principal 2

$$\Psi_{CylindreAP2} = \frac{\pi}{4} \frac{L_{AP} L_{AP\perp 2}^2}{V} \in [0,1]$$

Tous ces indices valent 1 dans le cas d'un cylindre.

Mais on peut remarquer sur la figure 9.4 que la base des focis n'est pas circulaire, mais plutôt elliptique. On effectue alors le même raisonnement que précédemment en se basant sur le volume des cylindres à base elliptique donné par : $V = H\pi ab$ avec a la longueur du demi grand axe de l'ellipse et b la longueur du demi petit axe. On obtient de même les égalités suivantes :

- $H = L_{AP}$
- $a = \frac{1}{2} L_{AP\perp 1}$
- $b = \frac{1}{2} L_{AP\perp 2}$

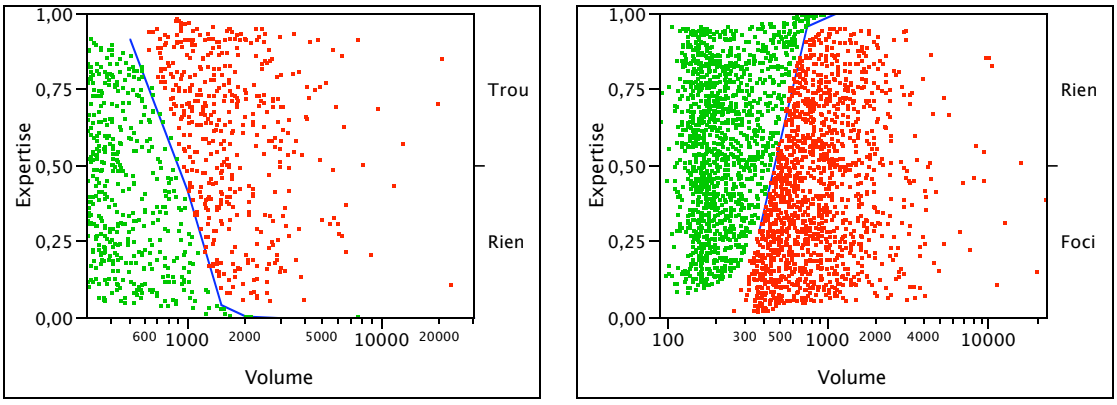
Ce qui permet de construire le dernier indice de caractérisation des cylindres qui vaut 1 pour tous les cylindres à base elliptique et donc aussi les cylindres à base circulaire :

$$\psi_{CylindreAP} = \frac{\pi}{4} \frac{L_{AP} L_{AP\perp 1} L_{AP\perp 2}}{V} \in [0,1]$$

REMARQUE - *Si un cône et un cylindre ont tous les deux une base et une hauteur égales, alors le volume d'un cône est le tiers de celui d'un cylindre. Donc construire des indices de caractérisation d'un cône à partir du volume engendre des indices linéairement dépendants des indices de caractérisation d'un cylindre. Il est par conséquent inutile d'utiliser des indices de caractérisation d'un cône dans nos modèles.*

9.4.4 Etape 4 : sélection des lacs et des pics

Nous avons utilisé la procédure d'extraction sur les 3300 noyaux dont nous disposions. Le résultat est deux ensembles constitués d'environ 6300 lacs et 15500 pics. Mais tous ces volumes ne sont pas nécessairement des trous ou des focis. Il faut donc construire deux modèles pour distinguer d'une part les volumes engendrés par de simples variations d'intensité et d'autre part ceux qui sont les trous ou les focis que nous souhaitons détecter.
Pour cela, chaque volume est caractérisé par 21 indices de forme : les versions 3D des indices de forme 2D auxquels s'ajoutent les nouveaux indices de forme 3D que l'on vient de présenter. Mais l'analyse des mesures effectuées sur les deux ensembles a mis en exergue l'importance d'une mesure qui se révèle discriminante : le volume. En effet, il existe un écart de volume significatif entre ceux les trous (resp. focis) et les autres (variations d'intensité de la texture, précipités, artefacts, etc.). La figure 9.8 montre la capacité de classement du volume. En mono-variable et par régression logistique, le volume classe correctement 79% des focis et 80% des trous. Cette mesure est donc ajoutée aux caractéristiques. Chaque volume dispose désormais d'un vecteur caractéristique de dimension 22.



(a) (b)

Figure 9.8. Illustration de la capacité discriminante du volume pour le classement : des lacs (a) dans les classes "Trous" et "Rien" et des pics (b) dans les classes "Focis" et "Rien".

L'expertise a révélé le nombre et la position des trous et des focis dans les noyaux. En utilisant ces informations lors de l'extraction des lacs et des pics, on obtient une expertise de ces derniers :
- 454 lacs sont des trous, soit environ $7,25\%$ de la population totale.
- 1424 pics sont classés parmi les "Focis", soit environ $9,2\%$.

Tout comme pour l'homogénéité de la texture (cf. section 6.5), ces déséquilibres nous contraignent une fois de plus à construire des échantillons de travail équilibrés en sélectionnant les indi-

vidus les plus représentatifs à l'aide des formes fortes. Une fois les échantillons construits, nous testons les différentes méthodes de classement.

9.4.4.1 Les trous

Nous souhaitons classer les lacs dans les classes *Trou* et *Rien*. Pour cela tous les modèles sont construits à partir d'un échantillon de travail composé des 454 volumes classés comme "Trous" auxquels sont ajoutés les 454 volumes les plus représentatifs de la classe "Rien" (cf. algorithme 5). Ce qui fait un échantillon composé de 908 individus. Une recherche exhaustive est lancée pour la régression logistique et les forêts aléatoires. En revanche, il n'est pas possible d'appliquer une telle recherche dans un temps raisonnable pour les k-plus proches voisins et le réseau de neurones. Nous avons utilisé une méta-heuristique de type *tabou* [Glover 1986] qui est une méthode recherche locale très utilisée dans la résolution de problèmes difficiles tels que le voyageur de commerce. La taille importante de cet échantillon permet de valider les sous-modèles par un protocole de type *k-fold* avec $k = 10$, car il apporte un bon compromis entre validation robuste (sur 10% des données) et ménagement de l'apprentissage par cœur (défaut du protocole *Leave-One-Out*), tout en utilisant la totalité des données. Le tableau 9.1 et la figure 9.9 montrent les résultats des recherches.

Méthodes / Nombre d'indices	$N_i + 19$-PPV	RL	FA	PMC / 2
1	90,85	89,76	86,88	90,74
2	90,85	91,95	90,97	92,07
3	91,07	93,17	92,93	92,84
4	91,4	93,72	93,5	93,51
5	91,85	94,38	93,49	93,5
6	91,62	94,82	93,71	94,31
7	91,74	94,6	93,72	94,66
8	91,51	94,71	93,94	94,69
9	91,51	94,82	93.94	94,34
10	91,62	94,82	93,94	94,72
11	91,85	94,82	93,94	94,27
12	91,74	94,6	93,72	94,61
13	91,96	94,6	93,83	94,28
14	91,74	94,39	94,04	93,92
15	91,07	94,38	93,59	94,07
16	91,51	94,27	93,49	93,9
17	91,5	94,16	93,17	93,18
18	91,18	93,83	92,94	93,07
19	91,29	93,82	92,72	92,96
20	91,07	93,72	92,63	92,4
21	90,96	93,38	92,51	92,19
22	90,74	92,83	91,84	91,84

Table 9.1. Taux de prédictions obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement des trous. Les abréviations correspondent aux méthodes suivantes : $N_i + 19$-PPV les k-plus proches voisins (avec k égal 19 additionné du nombre d'indices utilisés), RL la régression logistique, FA les forêts aléatoires et PMC / 2 le réseau de neurones avec $v = 2$.

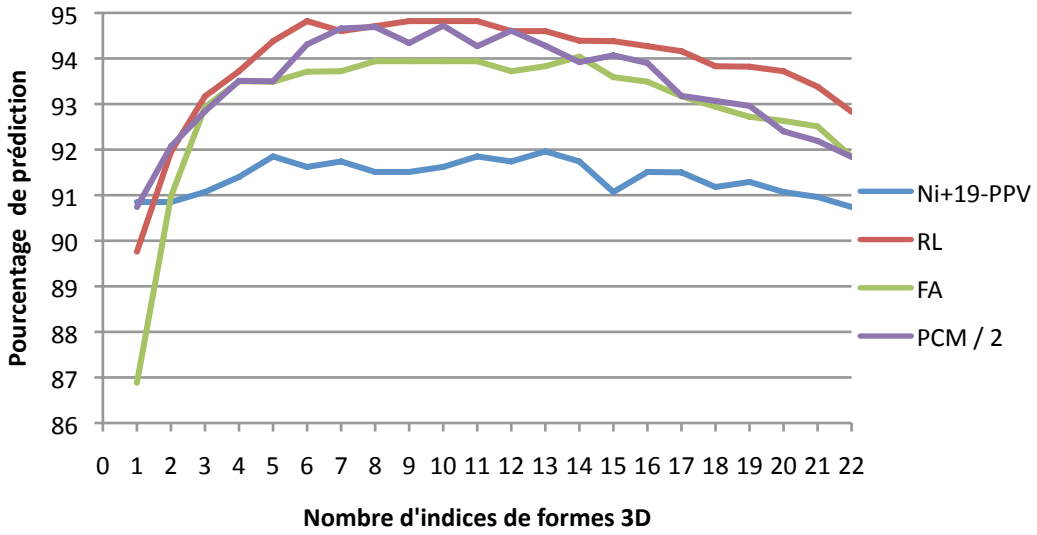


Figure 9.9. Comparaison graphique des performances des différentes méthodes de classement appliquées au classement des trous. En abscisse le nombre d'indices de volume utilisés et en ordonnée le pourcentage de classement obtenu.

La régression logistique apporte le meilleur résultat pour une combinaison de 9 indices, avec un taux de classement de $94,82\%$ ($46,7\%$ VP et $48,1\%$ VN), un intervalle de confiance de $[93,9\ldots95,7]$ et une probabilité inférieure à 10^{-4}. Le réseau de neurones (avec $\nu = 2$) apporte des résultats comparables malgré une courbe de résultats irrégulière (cf. figure 9.9). Ces irrégularités proviennent de la difficulté rencontrée par la méthode tabou à trouver les meilleures combinaisons d'indices. On peut toutefois remarquer des résultats très proches pour 7, 8, 10 et 12 indices. Les 9 indices utilisés par le modèle sont les suivants (classés par ordre décroissant d'importance en fonction du χ^2) :

1. Convexité volumétrique, $\chi^2 = 52$.
2. Convexité surfacique, $\chi^2 = 43$.
3. Volume, $\chi^2 = 20$.
4. Symétrie de Besicovitch, $\chi^2 = 14$.
5. $\psi_{Courbure}$, $\chi^2 = 13$.
6. Variance surfacique, $\chi^2 = 7,5$.
7. Sphéricité, $\chi^2 = 1,2$.
8. $\psi_{CylindreR}$, $\chi^2 = 0,08$.
9. Ecart à la sphère inscrite, $\chi^2 = 0,03$.

Il apparaît que la convexité et la taille sont les deux éléments les plus importants dans le classement des lacs. La figure 9.10 montre la distribution des probabilités qui est nettement bimodale (forte répartition sur les extrémités). On peut remarquer le faible nombre de cas ambigus et d'erreurs graves. Les résultats obtenus ainsi que ces observations nous permettent de conclure que le modèle classe efficacement les lacs.

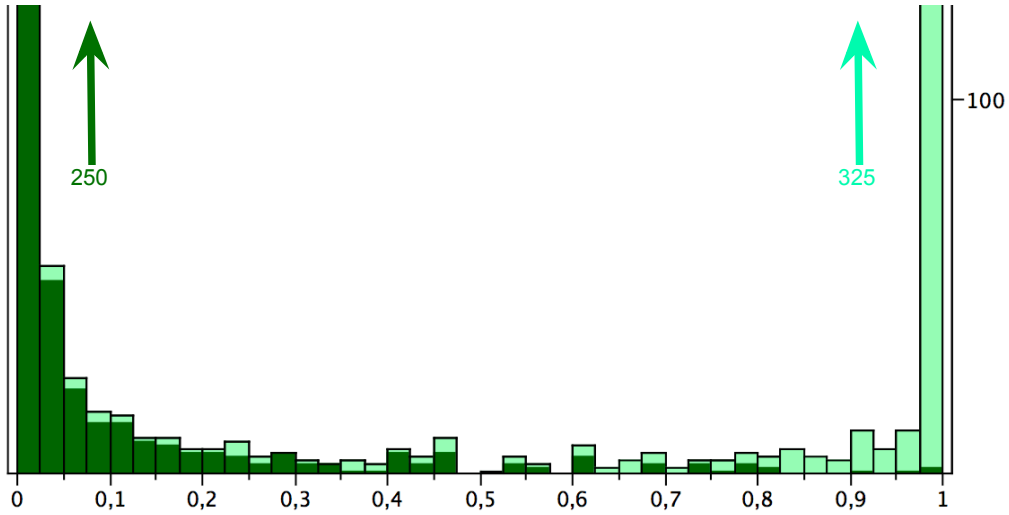


Figure 9.10. Histogramme des distributions des probabilités attribuées par le modèle de classement des lacs par indices de forme 3D. En vert foncé (resp. vert clair) les trous (resp. les lacs qui ne sont pas des trous).

9.4.4.2 Les focis

Avec les pics, nous réalisons les mêmes recherches du meilleur modèle de classement, mais cette fois-ci pour répartir dans les classes *Foci* ou *Rien*. Nous disposons de 1424 volumes classés comme focis auxquels ont été ajoutés les 1424 individus les plus représentatifs de la classe "Rien" sélectionnés à l'aide de l'algorithme 5. Nous avons donc construit un échantillon de travail composé de 2848 individus. Le tableau 9.2 et la figure 9.11 montrent les résultats des recherches.

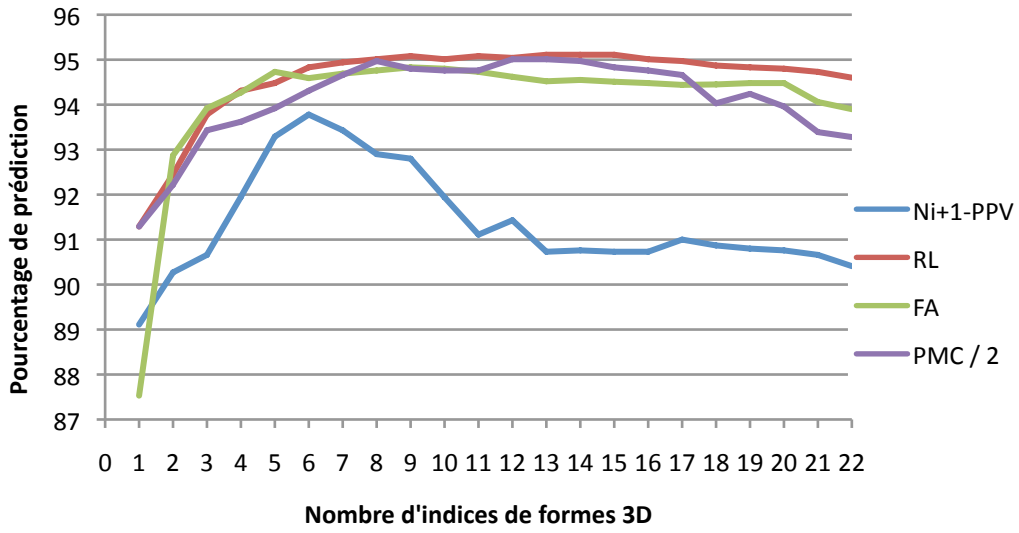


Figure 9.11. Performances des différentes méthodes de classement appliquées au classement des Focis. En abscisse le nombre d'indices de forme 3D utilisés et en ordonnée le pourcentage de classement obtenu.

Méthodes / Nombre d'indices	N_i+1-PPV	RL	FA	PMC / 2
1	89,11	91,29	87,53	91,29
2	90,27	92,45	92,87	92,22
3	90,66	93,78	93,92	93,43
4	91,95	94,31	94,27	93,62
5	93,29	94,48	94,73	93,92
6	93,78	94,83	94,59	94,31
7	93,43	94,94	94,69	94,66
8	92,9	95,01	94,76	94,97
9	92,8	95,08	94,83	94,8
10	91,94	95,01	94,8	94,76
11	91,11	95,08	94,73	94,76
12	91,43	95,04	94,62	95,01
13	90,73	95,11	94,52	95,01
14	90,76	95,11	94,55	94,97
15	90,73	95,11	94,51	94,83
16	90,73	95,01	94,48	94,76
17	91	94,97	94,44	94,66
18	90,87	94,87	94,45	94,03
19	90,8	94,83	94,48	94,24
20	90,76	94,8	94,48	93,96
21	90,66	94,73	94,06	93,39
22	90,41	94,6	93,9	93,28

Table 9.2. Taux de prédictions obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement des focis. Les abréviations correspondent aux méthodes suivantes : N_i+1-PPV les k-plus proches voisins (avec k égal au nombre d'indices utilisés additionné de 1), RL la régression logistique, FA les forêts aléatoires et PMC / 2 le réseau de neurones avec $\nu = 2$.

La régression logistique apporte le meilleur résultat avec un taux de classement de $95,11\%$ ($48,3\%$ VP et $46,8\%$ VN) pour une combinaison de 13 indices, mais le réseau de neurones avec $\nu = 2$ apporte des résultats comparables. L'intervalle de confiance est égal à $[94,8\ldots95,4]$ et la probabilité est inférieure à 10^{-4}. Mais lors de cette recherche, les forêts aléatoires obtiennent des résultats très proches de la régression logistique, voire comparables lorsque le nombre d'indice utilisé est inférieur à 10. Les 13 indices utilisés sont les suivants (classés par ordre décroissant d'importance en fonction du test du χ^2) :

1. Allongement par les rayons, $\chi^2 = 98$.
2. $\psi_{CylindreR}$, $\chi^2 = 95,5$.
3. Symétrie de Besicovitch, $\chi^2 = 95$.
4. Déficit iso-surfacique, $\chi^2 = 66$.
5. Volume, $\chi^2 = 20$.
6. Déficit, $\chi^2 = 12$.
7. Sphéricité, $\chi^2 = 10,5$.
8. $\psi_{CylindreAP2}$, $\chi^2 = 9$.

9. Ecart à la sphère inscrite, $\chi^2 = 5$.
10. $\psi_{Ellipsoide}$, $\chi^2 = 4,7$.
11. Etalement de Morton, $\chi^2 = 4,35$.
12. Convexité volumétrique, $\chi^2 = 0,05$.
13. Convexité surfacique, $\chi^2 = 0,01$.

L'importance des trois premiers indices est comparable. Dans leur représentation par des volumes binaires, les focis sont des formes allongées et par conséquent en première position on retrouve un indice caractérisant l'allongement. Mais la forme des focis est très proche de celle d'un cylindre (cf. section 9.4.3), il est donc logique d'avoir un de nos indices de caractérisation des cylindres en deuxième position. On retrouve également un autre de nos indices en huitième position (donc avec un rôle moins important), mais bien que cet indice emploie d'autres mesures son utilisation apporte peu d'informations supplémentaires au modèle. Contrairement aux lacs, la convexité joue un rôle mineur dans le classement des pics. Supprimer ces deux indices du modèle ne fait perdre que $0,2\%$ d'efficacité.
La figure 9.12 montre la distribution des probabilités attribuées par le modèle. Elle est nettement bimodale (forte répartition sur les extrémités). On peut remarquer le faible nombre de cas ambigus et d'erreurs graves. Les résultats obtenus ainsi que ces observations nous permettent de conclure que le modèle classe efficacement les pics.

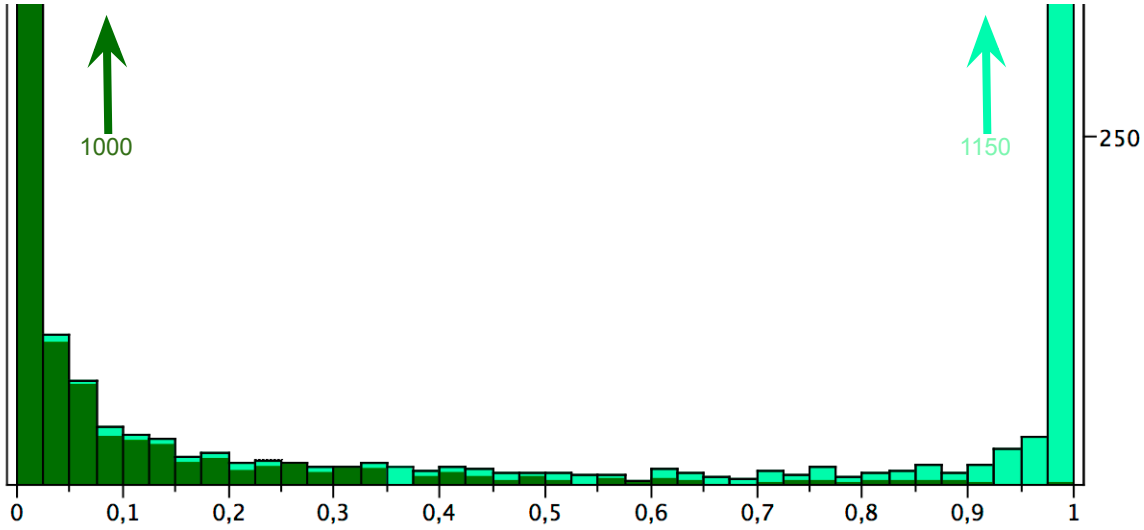


Figure 9.12. Distribution des probabilités attribuées par le modèle de classement des pics dans les classes "Focis" et "Rien". En vert foncé (resp. vert clair) les pics classés parmi les focis (resp. les non focis).

9.4.5 Etape 5 : classement des noyaux par l'analyse des trous et des focis

Nous disposons désormais de deux modèles permettant de classer efficacement les lacs (resp. les pics) dans les classes *Trou* (resp. *Foci*) ou *Rien*. Ces modèles permettent de déterminer le nombre de trous et de focis présents dans chaque noyau avec environ 5% d'erreur. Mais il est expliqué dans le chapitre 1, que le nombre de trous ou de focis présents dans un noyau n'est pas une information

suffisante (cf. section 1.5.2.2) : la taille est aussi à prendre en considération.
Malheureusement, on ne peut construire le vecteur caractéristique de chaque noyau avec le nombre des focis (ou trous) présents et leur volume. En effet, le nombre de focis (resp. trous) varie entre chaque noyau, donc la taille du vecteur caractéristique est dynamique, ce qui n'est généralement pas supporté par les classifieurs. Ainsi il nous faut choisir des variables globales qui permettent de donner des informations pertinentes et représentatives sur la population de focis (resp. trous) présente dans le noyau. Pour cela nous avons sélectionné six indices et mesures qui font intervenir aussi bien le nombre que le volume des focis (resp. trous) :

- le nombre de focis (resp. trous).
- la moyenne des volumes.
- la variance des volumes.
- la somme des volumes.
- le volume du plus petit foci (resp. trou).
- le volume du plus grand foci (resp. trou).

Chaque noyau est représenté par un vecteur caractéristique de dimension six afin de déterminer si les focis (resp. trous) qu'il contient le rendent pathologique. Nous construisons ainsi deux sous-modèles de classement des noyaux par l'étude de la présence des trous et des focis.

9.4.5.1 Classement des noyaux par l'analyse des trous

Parmi les 3300 noyaux dont nous disposons pour construire les modèles, 473 contiennent des trous et parmi ceux-ci 111 peuvent être classés comme pathologiques grâce à l'analyse des trous qu'ils contiennent. Nous construisons un échantillon de travail (cf. algorithme 5) constitué des 111 noyaux pathologiques et des 111 noyaux les plus représentatifs parmi les noyaux possédant des trous mais qui ne sont pas pathologiques. L'échantillon de travail est donc formé de 222 noyaux. C'est un échantillon spécialisé car il n'est constitué que de noyaux dont la texture contient au moins un trou. Compte tenu des individus, il est inutile de fournir à l'échantillon d'apprentissage des noyaux sans trou car il existe une relation d'ordre : noyaux sans trous $\prec$ noyaux non pathologiques avec trous $\prec$ noyaux pathologiques avec trous (en raison des trous qu'ils contiennent).

Nous testons de manière exhaustive toutes les combinaisons d'indices afin de déterminer le meilleur sous-ensemble et ce pour tous les classifieurs (cf. tableau 9.3). La validation est effectuée avec le protocole *Leave-One-Out* en raison de la faible taille de l'échantillon.

Méthodes / Nombre d'indices	$N_i + 1$-PPV	RL	FA	PMC / 3
1	85,13	86,93	83,65	87,38
2	85,19	88,73	85,61	88,9
3	88,32	89,63	85,65	90,09
4	86,56	90,09	85,11	89,78
5	85,57	89,63	84.62	89,66
6	83,75	90,54	81.46	89,18

Table 9.3. Taux de prédictions obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement des noyaux en fonction des trous présents.

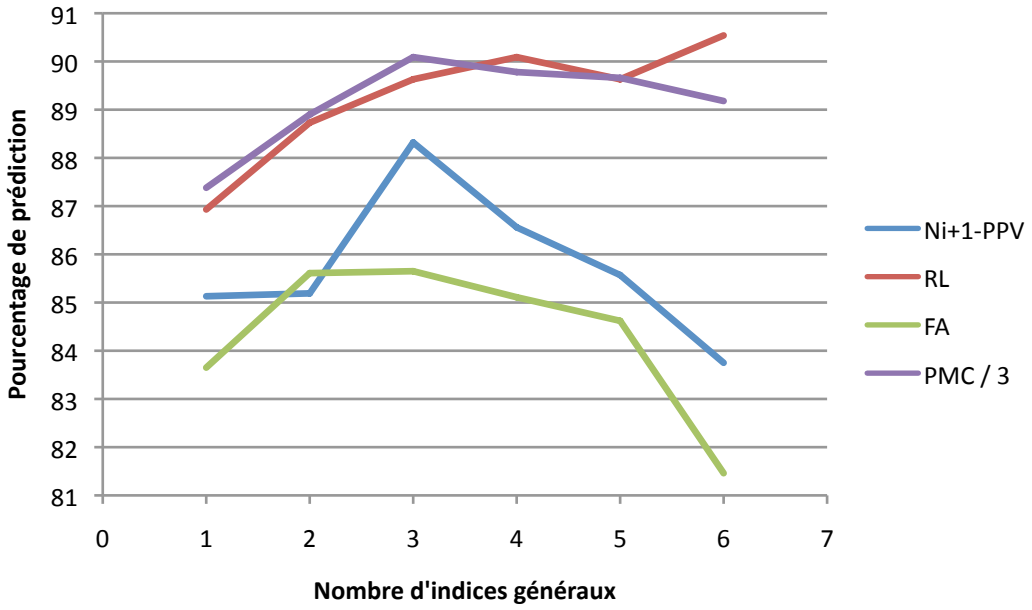


Figure 9.13. Comparaison graphique des performances des différentes méthodes de classement appliquées au classement des noyaux en fonction de l'analyse des trous. En abscisse le nombre d'indices et mesures généraux sur les trous présents dans les noyaux et en ordonnée le pourcentage de prédiction obtenu.

La figure 9.13 montre les performances des différents classifieurs. On peut remarquer immédiatement les performances nettement plus élevées de la régression logistique et du réseau de neurones (avec $v = 3$) par rapport aux *k*-plus proches voisins (avec *k* égal 1 additionné du nombre d'indice utilisés) et aux forêts aléatoires. La meilleure méthode de classement est la régression logistique avec tous les indices. Elle apporte un pourcentage de prédiction satisfaisant, supérieur à 90% (47,4% VP et 43,1% VN). Les erreurs du modèle s'expliquent par la difficulté à différencier les trous contenus dans le noyau avec les variations d'intensité de la texture du noyau. Le niveau variable du contraste des noyaux est un défaut dû à l'acquisition. Ce défaut d'intensité influe directement sur la méthode d'extraction des lacs qui commet alors des erreurs en ne détectant pas tous les lacs. L'absence de détection de certains lacs qui s'avèrent être des trous, entraîne presque systématiquement une erreur de diagnostic des noyaux.

REMARQUES - *Ce problème est également constaté sur l'analyse de l'histogramme des distributions des probabilités (figure 9.14). On observe une concentration importante des noyaux pathologiques pour des probabilités inférieures à* 0,05*, mais on peut remarquer des erreurs graves. Ce sont les noyaux dont l'extraction des lacs n'a pas été optimale et dont certains lacs non détectés étaient des trous. On peut également remarquer une concentration moins marquée sur les extrémités des probabilités des noyaux sains par rapport aux précédents modèles.*

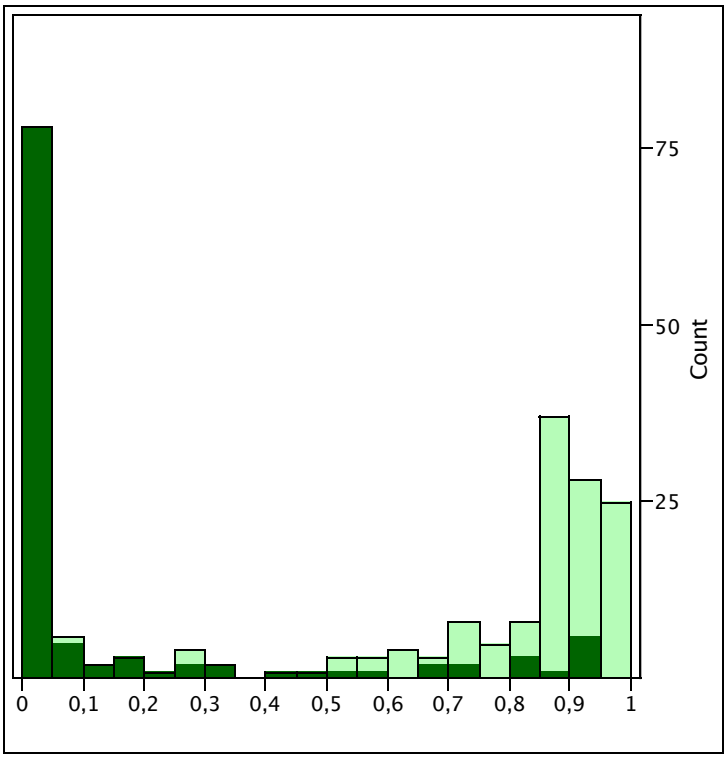


Figure 9.14. Distribution des probabilités attribuées par le modèle de classement des noyaux par l'analyse des trous. En vert foncé (resp. vert clair) les noyaux pathologiques (resp. sains).

9.4.5.2 Classement des noyaux par l'analyse des focis

Pour le classement des noyaux par l'étude des focis, un travail identique à celui effectué précédemment avec les trous est réalisé : construction de l'échantillon de travail équilibré (nous disposons de 123 noyaux contenant des focis et pouvant être classés comme pathologiques), suivi d'une recherche exhaustive.

Méthodes / Nombre d'indices	$N_i + 1$-PPV	RL	FA	PMC / 3
1	90,17	92,27	90,32	92,68
2	92,24	93,49	91,44	93,92
3	93,92	94,3	91,44	94,72
4	92,17	94,3	91,37	94,71
5	91,37	93,08	90,97	94,61
6	89,37	92,27	89,44	94,3

Table 9.4. Taux de prédictions obtenus par chaque classifieur en fonction du meilleur sous-ensemble d'indices pour le classement des noyaux par l'étude des focis.

On peut noter un écart important entre l'efficacité des méthodes : la régression logistique et le réseau de neurones (avec $v = 3$) sont nettement plus performants que les deux autres méthodes. Mais dans ce sous-modèle, les performances du réseau de neurones sont supérieures à la régression logistique. Toutefois, on observe un pic de performances des k-plus proches voisins (avec k égal

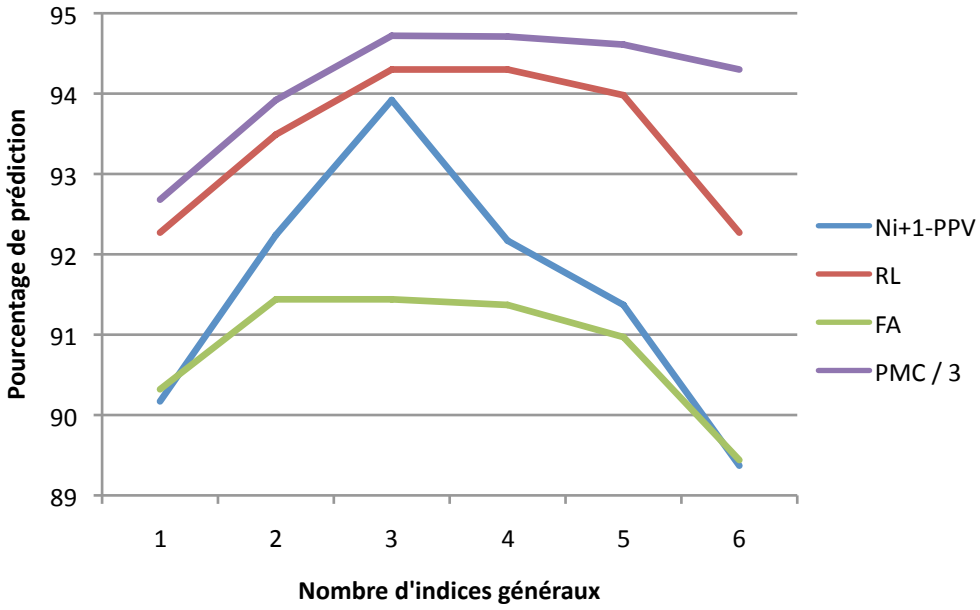


Figure 9.15. Performances des différentes méthodes de classement des noyaux par l'analyse des focis. En abscisse le nombre d'indices et de mesures utilisés et en ordonnée le pourcentage de prédiction obtenu.

au nombre d'indices utilisés additionnés de 1) pour trois indices (écart inférieur au pourcent avec le modèle optimal). Ce pic témoigne d'une bonne séparation des individus dans l'espace des caractéristiques ce qui est en accord avec les performances de ce sous-modèle. La meilleure configuration est l'utilisation de trois indices (nombre de focis, volume minimum et maximum) avec le réseau de neurones et elle obtient $94,72\%$ de prédiction ($46,8\%$ VP et $47,9\%$ VN).

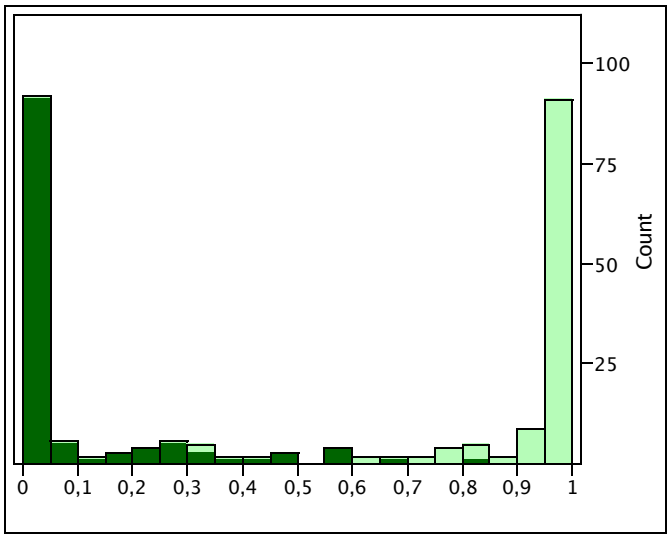


Figure 9.16. Distribution des probabilités attribuées par le sous-modèle de classement des noyaux utilisant la présence des focis. En vert foncé (resp. vert clair) les noyaux pathologiques (resp. sains).

La distribution des probabilités attribuées par le sous-modèle de classement des noyaux en fonction de la présence des focis (cf. figure 9.16) est très importante sur les extrémités. On observe aussi le très faible pourcentage de cas ambigus. De plus, on dénombre un seul et unique faux positif (noyau pathologique classé comme étant sain). Donc le sous-modèle est capable d'apprendre de manière robuste le classement des noyaux pathologiques, mais en revanche pas pour les noyaux sains qui représentent la quasi totalité des erreurs.

9.5 Conclusion

Dans ce chapitre, nous avons décrit une succession de procédés et de transformations de la texture, dont le résultat apporte une description statistique. La première étape importante est basée sur une représentation originale de la texture par son volume sous la nappe. Elle permet de caractériser de manière différente par l'utilisation d'indices de forme 3D les éléments extraits. Une description statistique des éléments ou de leurs caractéristiques apporte alors la description finale de la texture. Dans tout ce chapitre et cette conclusion, nous n'employons pas les mots "méthode" ou "technique" pour citer le travail que nous présentons. Ceci est motivé par la généricité et les étapes laissées au choix de l'utilisateur en fonction du problème qu'il souhaite traiter. En effet, la méthode d'extraction et le choix des indices généraux qui sont ceux caractérisant la texture, ne sont pas décrits alors qu'ils conditionnent le résultat. Mais en laissant ces choix à l'utilisateur, cela ne restreint pas les possibilités d'analyse de la texture, mais permet au contraire une utilisation pour tous les types de textures existants.

Les différentes étapes ont été appliquées avec succès sur les noyaux de cellules afin d'extraire et d'analyser la présence des trous et des focis. Pour améliorer la caractérisation des éléments extraits, nous avons proposé de nouveaux indices de forme 3D caractérisant des cylindres par l'emploi de différentes mesures. Ces indices ont permis d'améliorer les sous-modèles de classement permettant de différencier les trous et les focis parmi les éléments extraits. Ensuite les indices généraux utilisés ont apporté des informations pertinentes qui ont permis la construction de deux modèles de classement efficaces des noyaux contenant des trous et des focis.
Ces deux modèles fournissent les dernières informations nécessaires sur la caractérisation des noyaux et offrent la possibilité de construire le modèle final de classement des noyaux.

TROISIÈME PARTIE

MODÈLE FINAL, CONCLUSION ET PERSPECTIVES

CHAPITRE 10

MODÈLE FINAL DE CLASSEMENT DES NOYAUX

10.1 Introduction

Dans les premiers chapitres de ce manuscrit, nous avons présenté et étudié l'importance des différents éléments de diagnostic utilisés par les experts pour classer les noyaux de cellules dans les classes "sain" et "pathologique". Pour chacun de ces éléments, nous avons apporté une solution de caractérisation permettant de décrire de manière pertinente les sous-problèmes. Ces caractérisations ont permis de construire des sous-modèles de classement pour chaque élément de diagnostic. Chaque sous-modèle apporte une probabilité de classement des noyaux pour le sous-problème qu'il résout, ainsi que les indices et la façon de les combiner pour obtenir cette probabilité. Nous sommes donc maintenant en possession de sous-modèles de classement permettant de résoudre tous les sous-problèmes de diagnostic. Nous pouvons utiliser tous ces sous-modèles afin de résoudre le problème présenté dans le premier chapitre. L'étape finale de ce travail consiste donc à fusionner toutes les informations apportées par les sous-modèles créés, dans un ultime modèle de classement répondant au problème qui a motivé cette thèse.
Pour cela, nous proposons différentes combinaisons logiques de ces informations afin de pouvoir sélectionner la meilleure. La validation est effectuée par validation croisée *k-fold* pour $k = 10$, car elle permet un bon compromis entre validation robuste (sur 10% des données) et ménagement de l'apprentissage par cœur (défaut du *Leave-One-Out*), tout en utilisant la totalité des individus.

10.2 Les différentes approches de la construction du modèle de classement final

Cette section présente les différentes approches de construction du modèle final à partir de toutes les informations apportées par chaque sous-modèle de classement.

10.2.1 Classement par arbre binaire

Lorsqu'un expert réalise le classement des noyaux, il évalue chaque élément de diagnostic. Si parmi ces éléments il en existe au moins un qui est anormal, alors le noyau analysé est considéré comme pathologique. Sinon, l'expert le classe parmi les noyaux sains.
Cette approche peut se modéliser sous forme d'un arbre binaire dans lequel chaque nœud est constitué du résultat d'un classifieur. Si la probabilité générée par un classifieur est inférieure à $0,5$ alors le résultat du classement est *pathologique* (cf. figure 10.1), sinon le résultat est *sain*.

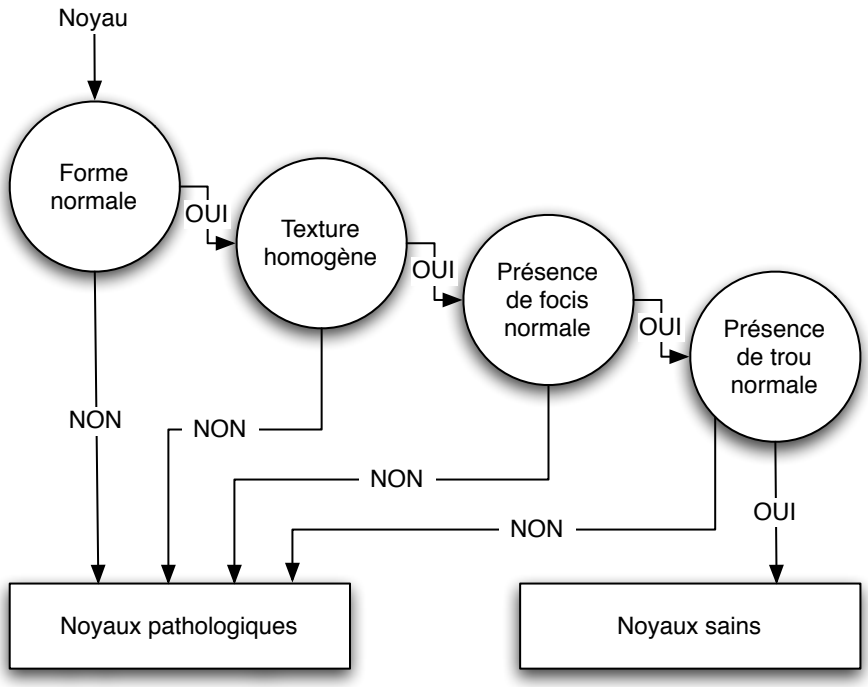


Figure 10.1. Arbre binaire de décision dans lequel chaque nœud contient le résultat d'un classifieur.

Combiner les probabilités dans un arbre binaire construit en fonction de l'importance des éléments de diagnostic permet d'obtenir $84,02\%$ de bon classement des noyaux.
Mais cette approche est extrêmement sensible aux faux négatifs. En effet, travailler de façon binaire engendre le cumul de tous les faux négatifs de chaque sous-modèle de classement. De plus, il n'y a aucune correction des faux positifs. Donc cette approche additionne les erreurs de chaque modèle ce qui la rend particulièrement sensible aux faux engendrés par tous les sous-modèles.

10.2.2 Classement par combinaison des probabilités

Les décisions trop strictes de la méthode précédente ne permettent pas d'obtenir un taux satisfaisant de classement des noyaux. La décision d'un expert n'est parfois pas aussi tranchée et un *doute de diagnostic* sur plusieurs éléments peut entraîner un classement *pathologique*.
Afin de diminuer l'accumulation des erreurs, il est envisageable de combiner les probabilités en utilisant les méthodes de classement (cf. figure 10.2). Toutefois, la représentation du diagnostic sous forme d'arbre de décision est intuitive. Pour cette raison, nous utilisons en plus des méthodes déjà utilisées dans les chapitres précédents un arbre de classement de type *Classification And Regres-*

sion Trees (CART) [Breiman et al. 1984]. L'ordre des nœuds dans l'arbre apporte une information importante pour le classement. Plus un nœud est proche de la racine de l'arbre, donc parmi les premiers interrogés, plus celui-ci est discriminant et donc important dans le modèle. Cette information permet de déterminer l'importance de chaque élément de diagnostic dans la pratique.

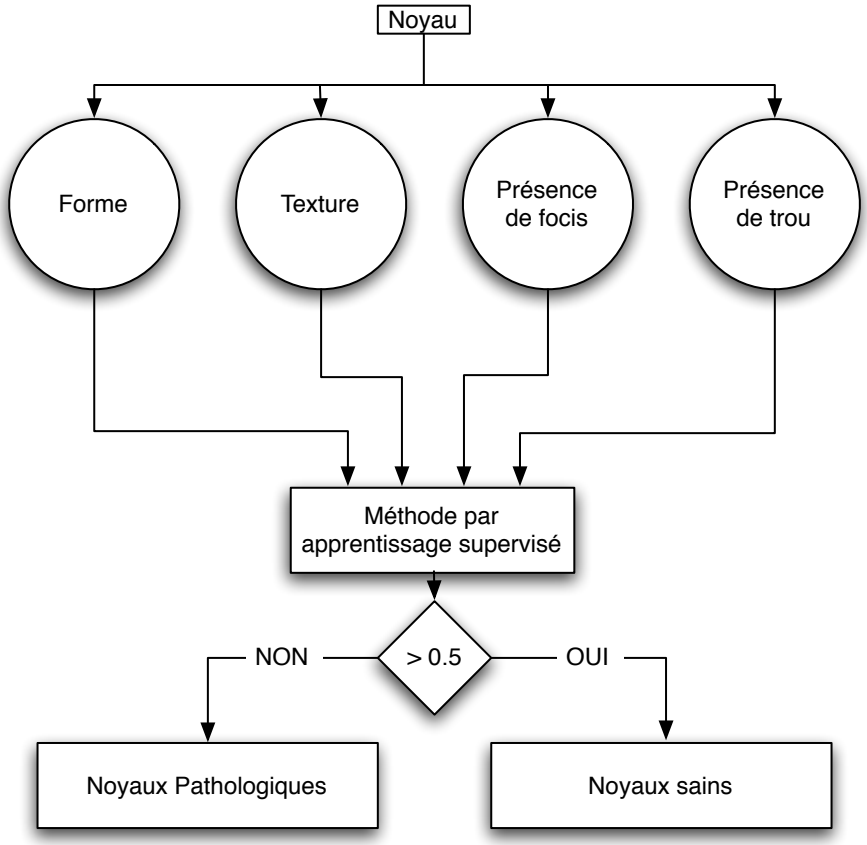


Figure 10.2. Combinaison des probabilités générées par chaque sous-modèle.

REMARQUE - *Bien que chaque sous-modèle construit soit performant dans le sous-problème qu'il résout, il existe des inégalités de performance. Par exemple, pour deux sous-modèles modélisant deux éléments de diagnostic d'importance strictement équivalente, le sous-modèle qui a le meilleur taux de prédiction est considéré comme étant plus pertinent pour le classement. Au contraire, si un des deux éléments de diagnostic est légèrement moins pertinent mais que son sous-modèle associé est nettement plus efficace, il est positionné haut dans l'arbre. Donc ces différences de performance peuvent influer sur l'ordre de positionnement de chaque sous-modèle dans la construction de l'arbre de décision. Mais dans notre travail, on sait que la forme est l'élément le plus important, que son sous-modèle associé est celui qui obtient les meilleures performances et que les noyaux boursouflés sont majoritaires parmi les noyaux pathologiques. Donc cette discrimination n'influe pas dans l'interprétation de l'arbre. De plus, les trois autres sous-modèles ont des performances relativement équivalentes et des tailles de classes proches. Ceci facilite l'interprétation de l'arbre qui est construit.*

Les résultats obtenus avec les différentes méthodes sont les suivants :

1. Régression logistique, $91,33\%$.
2. Réseau de neurones, $91,22\%$ ($v = 2$).
3. Arbre de décision CART, $90,5\%$.
4. K-plus proches voisins, $89,46\%$ ($k = 5$).
5. Forêts aléatoires, $89,22\%$.

On peut constater une amélioration significative des performances par rapport à l'arbre binaire précédemment construit. Donc l'utilisation d'un classifieur sur les probabilités obtenues des sous-modèles permet d'améliorer le classement. La combinaison des résultats corrige une partie des erreurs commises par les sous-modèles.

REMARQUES - *Le classement des noyaux par l'arbre de décision CART a mis en exergue l'ordre d'importance des sous-modèles (donc des éléments de diagnostic, cf. figure 10.3) et confirme l'étude réalisée dans la section 3.2 relative à l'analyse du diagnostic des noyaux par les experts.*

- *Le nœud le plus proche de la racine est composé du sous-modèle de classement de la forme, puis vient le sous-modèle inhérent à la présence de trous et enfin celui relatif aux focis. mais on constate que le sous-modèle de texture n'est pas utilisé dans cet arbre de décision.*
- *Donc la forme est bien l'élément de diagnostic le plus important. Ne pas l'utiliser dans le modèle ne permet pas de dépasser* 73% *de prédiction, ce qui est nettement inférieur au taux de prédiction obtenu en utilisant uniquement la forme (environ* 87%*).*
- *La forte intersection des noyaux classables par leur texture avec ceux classables par les autres éléments de diagnostic rend l'analyse de la texture moins significative pour le classement final. Supprimer la texture du modèle final fait chuter les performances à* $90,99\%$*, soit environ* $0,3\%$ *ce qui n'est pas une baisse significative. C'est la raison pour laquelle ce sous-modèle a été supprimé lors de la phase d'élagage de l'arbre de décision CART.*
- *En revanche, ne pas utiliser les probabilités fournies par les sous-modèles construits à partir des trous et des focis engendre une chute des performances d'environ* 2%*. Il est par conséquent plus important d'analyser la présence des trous et des focis lors de l'expertise que l'homogénéité de la texture.*

10.2.3 Utilisation des indices

Les sous-modèles donnent chacun une probabilité de classement pour chaque noyau afin de répondre aux sous-problèmes qu'ils modélisent. Lors de la construction, une recherche exhaustive a été systématiquement réalisée afin de déterminer les meilleures combinaisons d'indices (de forme, de texture ou d'indices généraux). Donc nous disposons de la liste de tous ces indices qui permettent de construire les sous-modèles. Il est alors possible d'utiliser la totalité de ces indices dans la construction du modèle final. Mais cela représente au total 32 indices de forme, de texture et généraux. Ce nombre élevé risque d'engendrer un apprentissage par cœur qui nuirait au classement des noyaux.

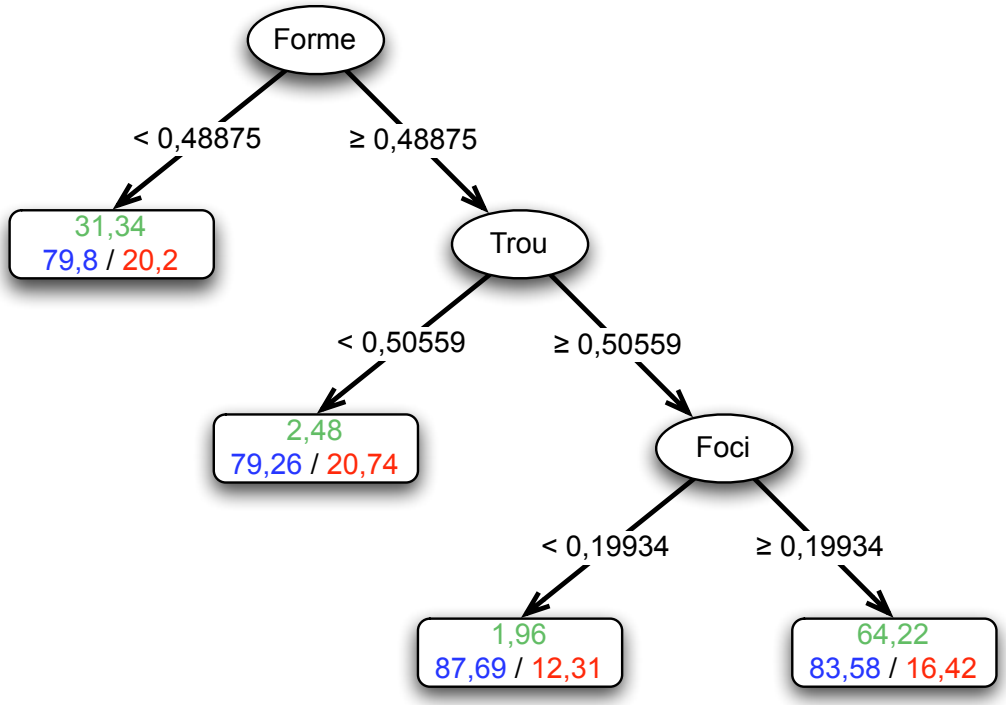


Figure 10.3. Illustration de l'arbre CART construit à l'aide des résultats des sous-modèles. En vert le pourcentage de noyaux arrivant dans la feuille et en bleu (resp. rouge) le pourcentage de noyaux bien (resp. mal) classés.

Pour construire ce dernier modèle, nous disposons de :
- 11 indices de forme.
- 12 indices de texture.
- 6 indices généraux issus de l'étude des trous.
- 3 indices généraux issus de l'étude des focis.

La construction des différents modèles apporte les résultats suivants :

1. La régression logistique, $90,57\%$.
2. Les forêts aléatoires, $89,57\%$.
3. Le réseau de neurones, $88,84\%$ ($\nu = 2$).
4. L'arbre de décision CART, $88,6\%$.
5. Les k-plus proches voisins, $88,39\%$ ($k = 33$).

On constate que ces résultats sont moins performants que précédemment car le nombre important d'indices utilisés a engendré un apprentissage par cœur.

10.3 Modèle final

Plusieurs solutions de construction du modèle final viennent d'être testées. Il apparaît que la meilleure est de combiner le résultat de chaque sous-modèle par régression logistique (cf. figure 10.5). Cette solution permet d'obtenir 91,3% de classement des noyaux (60,9% VP soit 3,9% d'erreur et 30,4% VN soit 4,8% d'erreur). Ce taux est supérieur au taux de répétabilité des experts qui est inférieur à 90%, ce qui confirme la pertinence du modèle. De plus, aucun modèle de classement des noyaux par analyse de la périphérie n'a été créé. Or environ 1% des noyaux pathologiques ne peuvent être classés comme tels sans effectuer une étude de leur périphérie. Ainsi le modèle final apporte 91,3% de prédiction sur 99% possible.

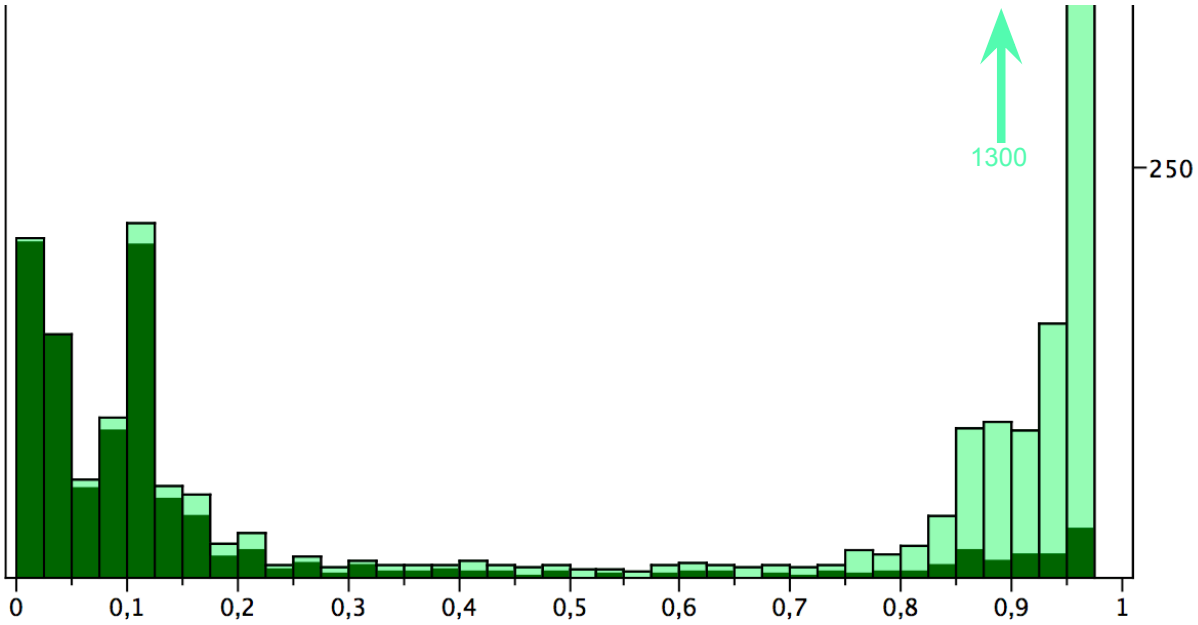


Figure 10.4. Distribution des probabilités attribuées par le modèle final. En vert foncé (resp. vert clair) les noyaux pathologiques (resp. sains).

La figure 10.4 montre la distribution des probabilités attribuées par ce modèle. On remarque une fois de plus une très forte répartition des probabilités sur les extrémités de l'histogramme et le faible nombre de cas ambigus. Ces deux informations démontrent l'efficacité du modèle et sa puissance de séparation des données. Mais on peut également observer la présence de quelques erreurs graves constituées par les faux positifs.

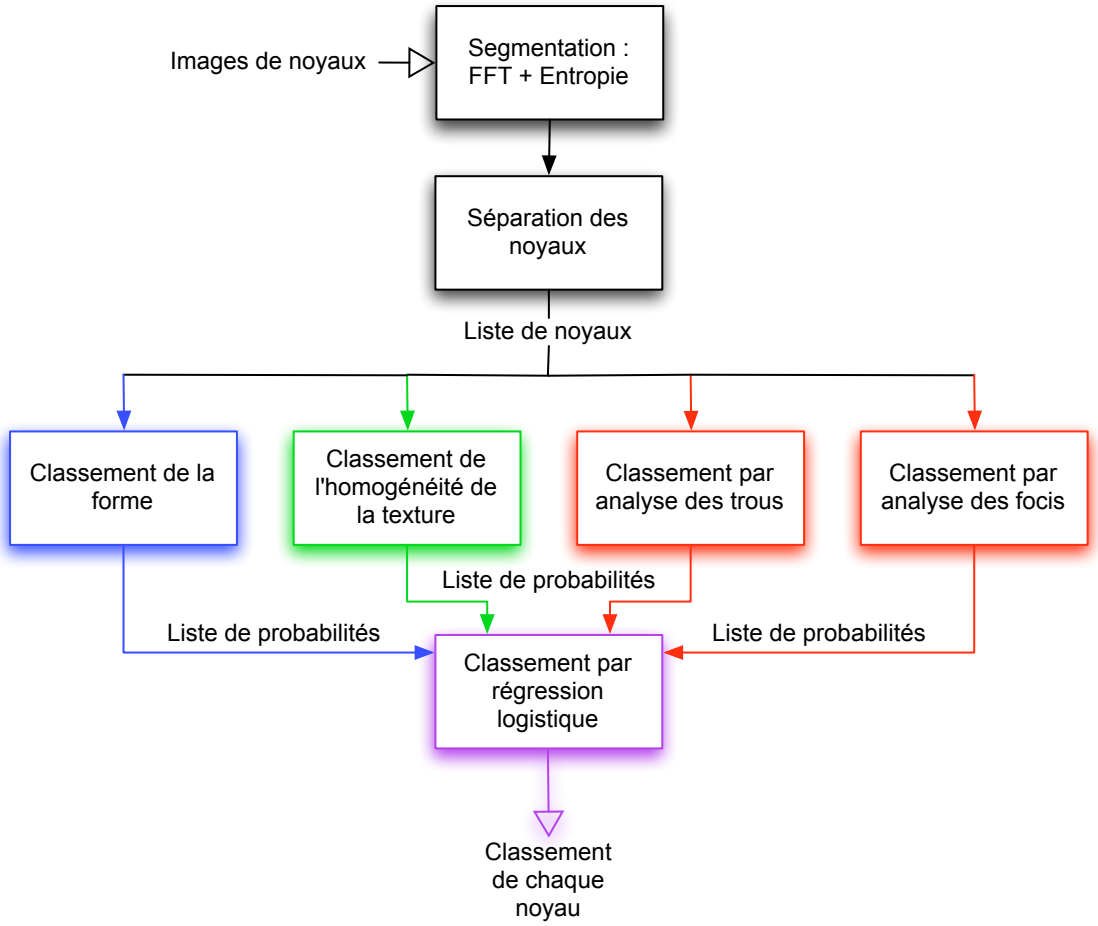


Figure 10.5. Schéma récapitulatif des étapes nécessaires au classement des noyaux de cellules.

CONCLUSIONS

Nous avons présenté une étude de la caractérisation et du classement des noyaux de cellules prélevés chez des patients atteints par la maladie de la Progéria. Dans un premier temps une analyse du problème a été réalisée afin d'appréhender les connaissances des experts du domaine dans le diagnostic (cf. chapitre 1). Pour chaque élément de diagnostic, les noyaux ont été expertisés et le taux de répétabilité mesuré (variation entre 85% et 92% selon les éléments) ce qui a constitué notre base de travail. Ce premier travail a permis d'établir un fil conducteur dans le processus de classement des noyaux de cellules dans les classes *sain* et *pathologique*, ainsi que l'importance de chaque élément de diagnostic (cf. section 3.2). Il est aussi apparu que les experts utilisaient certains critères qui faisaient intervenir des seuils approximatifs. Une partie de ce travail a donc consisté a déterminer ces seuils. Toutes ces remarques et observations ont été confirmées dans la suite du travail.

L'étape prédominante a été la réalisation d'un sous-modèle de classement de la forme des noyaux (dans les classes *forme normale* ou *forme boursouflée*) par indices de forme 2D (cf. chapitre 5). Une analyse des noyaux sains et boursouflés a révélé la nécessité de construire quatre nouveaux indices qui caractérisent les individus appartenant aux deux classes. La construction de ces indices (en particulier l'indice ψ_{Ncce} caractérisant la convexité) a été déterminante. Grâce à leur utilisation, ce sous-modèle a obtenu $95,4\%$ de classement de la forme. Il a permis de modéliser de manière très satisfaisante le sous-problème et son très bon résultat n'incitait pas à rechercher une amélioration des performances.

Mais le sous-modèle de classement de la forme n'a pas pu résoudre seul le problème final (seulement $86,9\%$ de prédiction). Une étude de l'homogénéité de la texture des noyaux a été nécessaire. Pour caractériser l'homogénéité, nous avons utilisé plusieurs méthodes statistiques connues (cf. chapitre 8) qui ont apporté des résultats peu satisfaisants. Une nouvelle méthode spécifique a alors été proposée : les *Gray Levels Size Zone Matrix* (GLSZM). Elle s'est révélée performante pour le problème traité, mais une analyse de ses erreurs a montré une forte corrélation entre les textures des faux positifs. Deux nouveaux indices de texture ont alors été proposés afin de corriger le classement de ces noyaux. Grâce à cette nouvelle méthode et aux deux nouveaux indices, ce sous-modèle a obtenu un taux de prédiction de plus 94%. Son utilisation combinée au sous-modèle de forme a amélioré le classement des noyaux pour le problème final de $0,7\%$. Cette faible amélioration était hautement significative vis-à-vis des possibilités d'améliorations théoriques (cf. section 3.2).

Certains noyaux ne pouvaient être caractérisés par leur forme et/ou leur texture. Il a alors été nécessaire de pratiquer une analyse des *focis* et des *trous* présents dans la texture des noyaux. Afin de caractériser ses éléments parfois composés de quelques pixels, la texture a tout d'abord été représentée par son volume sous la nappe (cf. chapitre 9). Cette représentation originale a permis d'extraire les éléments sous forme de volumes, puis de les caractériser par indices de forme 3D. Pour ce faire, une étude de l'extension des indices de forme de la 2D vers la 3D a été réalisée (cf.

section 9.3.2) et de nouveaux indices 3D (cf. section 9.4.3) ont aussi été élaborés. La caractérisation de ces volumes a apporté une description statistique de la texture. Grâce à ce procédé, il a été possible de construire deux sous-modèles de classement de la texture en fonction de la présence des trous et des focis. Ces sous-modèles ont permis respectivement de caractériser correctement $90,5\%$ et $94,7\%$ des noyaux concernés.

Après avoir construit un sous-modèle de classement pour chacun des éléments de diagnostic[1], il a été nécessaire de combiner les informations obtenues pour construire le modèle final. Pour cela nous avons étudié différentes approches : arbre de décision, combinaison des résultats et combinaison des indices utilisés.
L'emploi d'un arbre de décision a révélé que le classement des noyaux par la présence des trous et des focis est plus décisif que l'homogénéité de la texture. En revanche, utiliser les informations sous forme d'arbre n'apporte pas le meilleur taux de prédiction. La meilleure solution a été de combiner le résultat de chaque sous-modèle par régression logistique. Cette solution a apporté un taux de classement des noyaux de $91,3\%$, résultat supérieur au taux de répétabilité des experts.

Dans ce manuscrit, nous avons décrit les différentes étapes nécessaires à la réalisation d'un modèle de classement des noyaux de cellules prélevées chez des patients atteints par le syndrome de Hutchinson-Gilford. Grâce à notre modèle, il est désormais possible de classer de manière automatique, fiable et plus rapide les noyaux de cellules. Cette automatisation et ce gain de temps constituaient l'objectif de notre travail qui a donc été atteint. De plus, notre travail a montré la faisabilité du classement automatique de noyaux de cellules par l'étude de la répartition des lamines A et C.

[1] A l'exception de la périphérie qui ne concernait qu'une minorité de noyaux (environ 1%).

PERSPECTIVES

Dans ce manuscrit, nous avons décrit les différentes étapes nécessaires à la réalisation d'un modèle de classement de noyaux de cellules sanguines prélevées chez des patients atteints par le syndrome de Hutchinson-Gilford. Notre travail a montré la faisabilité du projet : il est désormais possible de classer de manière automatique, fiable et rapide les noyaux de cellules. Bien que nous ayons atteint l'objectif de notre travail, les perspectives sont nombreuses.

Le modèle que nous avons construit apporte un taux de classement des noyaux de l'ordre de 90%. Bien que ce résultat soit satisfaisant, une perspective immédiate et naturelle est de souhaiter améliorer le taux de classement des noyaux. Il est envisageable de travailler suivant deux approches.

La première approche est d'augmenter les capacités de prédiction de chaque sous-modèle. Pour cela, on peut souhaiter améliorer les méthodes utilisées (par exemple, élaborer de nouveaux indices de forme et de texture plus discriminants ou apportant des informations complémentaires) ou créer de nouvelles techniques dédiées plus efficaces. En particulier, nous avons vu que la méthode d'extraction des lacs utilisée dans la section 9.4.2 n'était pas totalement satisfaisante et elle était responsable de la présence d'une partie des faux positifs.
Nous avons élaboré des sous-modèles de classement des noyaux en fonction de l'analyse de la texture : homogénéité, présence des focis et des trous. Tous ces modèles ont été construits à l'aide d'indices qui sont des primitives extraites des noyaux. Les modèles ont nécessité une phase de pré-traitement afin d'extraire les caractéristiques de la texture. Nous souhaiterions ne plus effectuer cette phase et construire un classifieur qui travaille directement sur l'image des noyaux [Orlov et al. 2008]. Ce type de classifieur extrairait les caractéristiques au niveau du pixel afin de s'abstraire de tout type de pré-traitement. Pour cela, il serait nécessaire de recaler les noyaux (orientation et sans doute mise à l'échelle, cf. sections 4.2.2 et 4.3.2).
Cette mise à l'échelle pourrait d'ailleurs être directement implémentée sur le modèle actuel afin d'améliorer l'analyse de la texture.
Dans la plupart des modèles que nous avons construits, le réseau de neurones et la régression logistique ont apporté des résultats comparables. Lorsque plusieurs classifieurs obtiennent des résultats proches (pour un même problème) il est souvent intéressant d'utiliser la technique du *Bagging* [Breiman 1996] : les modèles sont validés par protocole *bootstrap* et leur prédiction sert de vote au classement final de l'individu étudié (principe du vote majoritaire par exemple). Ce système de vote dans lequel chaque classifieur a le même poids permet fréquemment d'améliorer les performances. Par la suite, si lors de l'utilisation du *bagging* les classifieurs commettent les mêmes erreurs, il est conseillé d'utiliser la technique du *Boosting* [Freund and Schapire 1999; Buhlmann and Hothorn 2007] : contrairement au *bagging*, le *boosting* pondère les individus mal classés par une majorité des classifieurs afin de les forcer à prendre en compte ces individus. Dans le cas du *boosting*, il n'est pas nécessaire de valider par *boostrap*, on préfère même utiliser un protocole de type *k-fold* afin de prendre en compte tous les individus.

La deuxième approche consiste à utiliser de nouvelles informations non traitées dans notre travail. Tous les sous-modèles obtiennent au minimum 90% de prédiction, voire même certains dépassent les 95%. Essayer d'améliorer ces sous-modèles reviendrait à tenter de corriger le classement de quelques noyaux.
Tout d'abord il y a la périphérie qui est le seul élément de diagnostic que nous n'avons pas utilisé dans ce document. Classer les noyaux ayant un défaut de périphérie peut théoriquement améliorer le classement de 1%.
De plus, les marqueurs DAPI et TRITC (cf. chapitres 1 et C) sont également utilisés lors de l'acquisition des noyaux. Par exemple, le DAPI (resp. le TRITC) marque la répartition de l'ADN inactif (resp. des lamines B) qui apporte des informations complémentaires sur l'état des noyaux. Disposer d'une expertise des noyaux à partir de la répartition de l'ADN (resp. des lamines B), permettrait de compléter le diagnostic et ainsi construire un nouveau sous-modèle pour accroître les performances.

Grâce à notre modèle, il est désormais possible de classer de manière automatique, fiable et rapide les noyaux de cellules par l'étude de la répartition des lamines A et C. Nous disposons désormais d'outils et de certaines connaissances pour la caractérisation et le classement des noyaux, il serait intéressant d'appliquer notre travail sur d'autres laminopathies, voire étendre à tout type d'analyse de noyaux.

Pour réaliser notre travail, nous ne disposions que d'une seule expertise ainsi que d'un nombre relativement limité de noyaux appartenant à certaines classes minoritaires. A l'exception de la forme, les autres sous-modèles ont été construits sur de petits échantillons et validés par protocole *Leave-One-Out*. Disposer d'un plus grand nombre de noyaux expertisés permettrait d'améliorer (en termes de performances et/ou de robustesse) tous nos sous-modèles. De même, nous souhaiterions disposer d'autres expertises car nous savons qu'il y a des divergences d'expertise entre les experts. Construire nos modèles à partir de noyaux pour lesquels tous les experts sont en accord permettrait d'obtenir des modèles plus robustes car les cas ambigus sont éliminés par la forte indécision des experts qui engendre des classements différents.
Par la suite, il serait intéressant de se concerter avec les experts à propos des noyaux sur lesquels ils sont en désaccord. Le modèle permettrait d'apporter un vote supplémentaire issu des connaissances de tous. Se concerter sur les résultats et les opinions de chacun pourrait faire avancer les connaissances en diagnostic. De plus, lorsque notre modèle est utilisé pour résoudre un désaccord, une étude des probabilités attribuées par chaque sous-modèle permettrait de comprendre quels éléments de diagnostics sont déterminants (probabilités extrêmes) lors du classement des cas litigieux.

Pour caractériser la convexité des noyaux, nous avons élaboré un nouvel indice de forme (cf. section 5.2) qui s'est révélé particulièrement pertinent. Mais pour obtenir le maximum d'efficacité de cet indice, il est nécessaire de réaliser une phase de calibrage. Nous avons effectué ce calibrage à l'aide d'une étude systématique bi-dimension mais qui a nécessité un temps de calcul important. Bien que ce calibrage rende cet indice souplesse et efficace, elle représente un inconvénient pratique. Afin de supprimer cet inconvénient, nous proposons de construire un modèle de calibrage automatique de l'indice pour chaque type de problème posé. A partir d'un échantillon expertisé, le modèle de calibrage déterminerait les seuils (en nombre de pixels et de composantes) à partir desquels on peut comptabiliser les composantes connexes d'écart.
De même, nous avons systématiquement effectué des recherches exhaustives afin de trouver le meilleur sous-ensemble d'indice pour chaque sous-modèle. Ces recherches nécessitent un temps de calcul considérable, voire dans certains cas un temps si important que nous avons dû utiliser une

méta-heuristique de type *tabou*. Nous souhaiterions utiliser des méthodes de sélections automatiques du meilleur sous-ensemble d'indices [Bruno et al. 1998; Jain et al. 2000; Wirjadi et al. 2006] afin d'éviter des recherches lourdes ou locales.
Mais l'étude des résultats des recherches exhaustives a révélé que des combinaisons d'indices différentes apportent exactement le même taux de classement, voire parfois commettent les mêmes erreurs. Donc il apparaît des équivalences de performances lors de l'utilisation de certains groupes d'indices. Une étude statistique de ces groupes permettrait de nous renseigner sur la nature même des indices. De plus, il serait envisageable de créer une *notice* d'utilisation des indices de forme/texture et ainsi déterminer automatiquement le meilleur choix d'indices en fonction du contexte d'application.

Dans la partie II de ce manuscrit, nous avons présenté une nouvelle technique de caractérisation de l'homogénéité d'une texture qui s'est révélée performante pour le sous-problème traité. Malheureusement, cette technique n'a pas apporté de résultat concluant lors de tests effectués sur des bases de textures telles que Brodatz [Brodatz 1966] et Ponce's Group [Lazebnik et al. 2005] (résultats non présentés dans ce document). Nous souhaitons améliorer notre méthode afin qu'elle caractérise tout type de texture :

- soit en pondérant les valeurs contenues dans la matrice à l'aide d'indices de forme. Parvenir à utiliser efficacement des indices de forme avec notre méthode permettrait de prendre en compte la forme des régions lors de l'analyse de la texture.
- soit en effectuant une fusion avec d'autres techniques telles que la matrice de cooccurrences ou la matrice des longueurs de segments. Fusionner notre contribution avec d'autres techniques caractérisant d'autres types de textures permettrait d'améliorer la caractérisation d'un plus grand nombre de textures.

La première étape de notre travail a consisté à segmenter puis à extraire les noyaux des images acquises au microscope. Mais lors de cette phase nous avons été confrontés à des problèmes de noyaux en contact, voire superposés. La méthode de segmentation ne permettait pas de séparer judicieusement les noyaux. Cette séparation n'est pas un problème trivial et mérite des recherches : la zone de chevauchement des noyaux est plus intense (plus grande quantité de marqueur). Si nous parvenons à séparer les deux noyaux sans déformer leur contour (car sinon cela pourrait perturber la caractérisation de la forme), il faudrait également séparer les intensités des deux textures. L'aboutissement de cette phase permettrait de travailler avec davantage de noyaux.

Nous savons qu'un individu est atteint par la Progéria si son taux de noyaux pathologiques dépasse un certain seuil. Or il existe différents degrés de sévérité dans la maladie. Il serait alors intéressant d'utiliser des méthodes de régression afin de construire un modèle qui permettrait d'estimer la sévérité de la maladie en fonction du pourcentage de noyaux pathologiques, voire d'autres critères.

QUATRIÈME PARTIE

ANNEXE

ANNEXE

A

NOTIONS MATHÉMATIQUES, DÉFINITIONS ET PROPRIÉTÉS DANS UN ESPACE DISCRET

A.1 Introduction

Tout au long de ce manuscrit, nous manipulons et étudions des noyaux de cellule dont les représentations sont présentes dans des images ou des volumes. Ces images et volumes sont des espaces discrets dont les spécificités engendrent certains inconvénients qui sont décrits dans ce chapitre. L'étude des noyaux fait également intervenir certaines définitions comme la distance et d'autres notions mathématiques qui sont également exposées.

A.2 Image et volume

Dans la section 1.3, il a été expliqué le type d'acquisition utilisé de manière majoritaire par les experts pour la visualisation des noyaux de cellule : le microscope à fluorescence. Ce microscope fournit des images qui contiennent un résultat visuel des acquisitions (cf. annexe C.2). Une image I peut être considérée comme un tableau à deux dimensions dans lequel $I(x,y)$ $(x,y \in \mathbb{N})$ représente la valeur du pixel d'abscisse x et d'ordonnée y. Il existe plusieurs façons de coder la valeur d'un pixel :

- Si l'image est *binaire* (ne contient que deux couleurs), $I(x,y) \in \{0,1\}$. Les pixels noirs représentent un pixel vide (absence de matière $I(x,y) = 0$) et les pixels blancs représentent de la matière ($I(x,y) = 1$).
- Si l'image est en *niveaux de gris*, $I(x,y) \in [0,N]$ avec $N > 1$ et en général $N = 2^n - 1$, $n \in \mathbb{N}^*$. Plus n est grand, plus le niveau de détail de la luminance des pixels de l'image est élevé. Dans ce manuscrit, les images en niveaux de gris sont codées sur $[0,255]$.
- Si l'image est en *couleurs*, $I(x,y) \in [0,N]^3$ avec $N > 1$ et en général $N = 2^n - 1$, $n \in \mathbb{N}^*$. Plus n est grand, plus le niveau de détail des couleurs de l'image est élevé. Dans notre travail, les images en couleur sont codées sur $[0,255]^3$ dans le mode *RGB*[1].

[1] http://fr.wikipedia.org/wiki/Codage_informatique_des_couleurs, codage additif des couleurs en rouge, vert et bleu.

A partir de ces définitions, on peut construire un volume en *empilant* une succession d'images de même dimension. Le volume ainsi construit est *binaire*, en *niveaux de gris* ou encore en *couleurs*.

Ces représentations faciles à appréhender et à manipuler, constituent des espaces discrets 2D et 3D qui posent quelques problèmes pour certaines transformations.

A.2.1 Rotation et homothétie

Dans un espace discret, certaines transformations connaissent des problèmes.
Le meilleur exemple est la rotation d'angle $\theta \in [0,\pi]$ dans $\mathbb{Z}^2$. Les rotations d'un objet dans un espace discret sont définies pour des valeurs $\theta = \frac{k\pi}{2}$, $k \in \mathbb{Z}$. Pour toute autre valeur de θ, il existe des pixels dont les coordonnées (x,y) après rotation n'appartiennent pas à $\mathbb{Z}^2$: $\exists\ p = I(x,y), \theta \neq \frac{k\pi}{2},\ k \in \mathbb{Z}\ /\ Rotation(p,\theta) = q(x_\theta, y_\theta)\ et\ (x_\theta, y_\theta) \notin \mathbb{Z}^2$. La figure A.1 illustre ce problème.

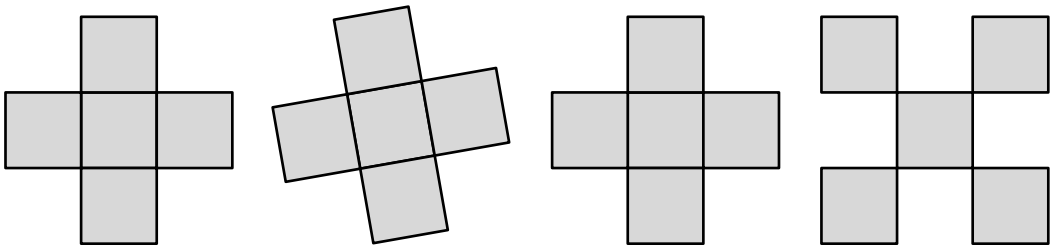


Figure A.1. Illustration de différentes rotations dans différents espaces. (a) Forme initiale, (b) rotation d'angle $\theta = 10^\circ$ dans $\mathbb{R}^2$, (c) rotation d'angle $\theta = 10^\circ$ dans $\mathbb{Z}^2$ et (d) rotation d'angle $\theta = 45^\circ$ dans $\mathbb{Z}^2$.

Par analogie, l'homothétie connaît exactement le même type de problème que la rotation lorsque le coefficient n'est pas un entier.

Dans ce manuscrit, certaines méthodes sont dites *invariantes* par rotation ou/et homothétie. Dans ces cas là, il est sous entendu que ceci est vrai malgré les problèmes liés au travail dans un espace discret qui viennent d'être cités.

A.3 Distance, voisinage, voisin, point adjacent, chemin et connexité

A.3.1 Distance

Définition A.3.1 (Distance) – *Une distance est une fonction* $d : A \longmapsto \mathbb{R}$ *qui doit avoir les propriétés suivantes :*

- *Etre positive :* $\forall x,y \in A, d(x,y) \geqslant 0$.
- *Etre réflexive :* $\forall x \in A, d(x,x) = 0$.

– *Etre symétrique* $\forall x,y \in A, d(x,y) = d(y,x)$.
– *Vérifier l'inégalité triangulaire :* $\forall x,y,z \in A, d(x,z) \leqslant d(x,y) + d(y,z)$.

A partir de cette définition, on définit un ensemble de distances 2D pour les images, 3D pour les volumes et dans $\mathbb{R}^n$ applicable à tous les supports :

– Distances 2D
Soit $x(x_1,x_2)$ et $y(y_1,y_2) \in \mathbb{Z}^2$

$$d_4(x,y) = \sum_{i=1}^{2} |y_i - x_i| \text{ , City block ou Manhathan}$$

$$d_8(x,y) = \max_{i=1,2} |y_i - x_i| \text{ , Diamond ou Chessboard}$$

On trouve également deux autres distances qui sont définies sur des maillages particuliers (hexagonaux et octogonaux) [Chassery and Montanvert 1991], où $[x]$ désigne la partie entière de x.

$$d_h(x,y) = \max(|x_1 - y_1|, \tfrac{1}{2}(|x_1 - y_1| + (x_1 - y_1)) - ([\tfrac{x_1}{2}] - [\tfrac{y_1}{2}]) + y_2 - x_2, \tfrac{1}{2}(|x_1 - y_1| + (x_1 - y_1)) - ([\tfrac{x_1}{2}] - [\tfrac{y_1}{2}]) + x_2 - y_2) \text{ , Hexagonale}$$

$$d_o(x,y) = \max\left(d_8(x,y), \left[\frac{2}{3}(|x_1 - y_1| + |x_2 - y_2| + 1)\right]\right) \text{ , Octogonale}$$

– Distances 3D
Soit $x(x_1,x_2,x_3)$ et $y(y_1,y_2,y_3) \in \mathbb{Z}^3$

$$d_6(x,y) = \sum_{i=1}^{3} |y_i - x_i|$$

$$d_{26}(x,y) = \max_{i=1,2,3} |y_i - x_i|$$

– Distances dans $\mathbb{R}^n$
Soit $x(x_1,x_2,...x_n)$ et $y(y_1,y_2,...y_n) \in \mathbb{R}^n$

$$d_1(x,y) = \sum_{i=1}^{n} |y_i - x_i|$$

$$d_2(x,y) = \sum_{i=1}^{n} \sqrt{(y_i - x_i)^2} \text{ , Euclidienne}$$

$$d_\infty(x,y) = \max_{i=1...n} |y_i - x_i|$$

REMARQUE - *Si les attributs possèdent des écarts types différents, alors ils interviennent de manière différente dans la distance. Ce problème peut parfois être intéressant lorsque l'on souhaite privilégier certains attributs par rapport à d'autres. Mais si ce n'est pas le cas, il convient alors de centrer et réduire*[2] *les attributs.*

[2] http://www.bibmath.net/dico/index.php3?action=affiche&quoi=./v/varcent.html.
http ://fr.wikipedia.org/wiki/Variable_centrée_réduite.

A.3.2 Voisinage, voisin et point adjacent

A l'aide des distances précédentes, on définit des voisinages N_i dans $\mathbb{Z}^2$ et $\mathbb{Z}^3$ (figure A.2) :

– Voisinages 2D
Soit $x \in \mathbb{Z}^2$,

$$N_4(x) = \{y \in \mathbb{Z}^2 / d_4(x,y) \leqslant 1\}$$
$$N_h(x) = \{y \in \mathbb{Z}^2 / d_h(x,y) \leqslant 1\}$$
$$N_8(x) = \{y \in \mathbb{Z}^2 / d_8(x,y) \leqslant 1\}$$
$$N_o(x) = \{y \in \mathbb{Z}^2 / d_o(x,y) \leqslant 1\}$$

– Voisinages 3D
Soit $x \in \mathbb{Z}^3$,

$$N_6(x) = \{y \in \mathbb{Z}^3 / d_6(x,y) \leqslant 1\}$$
$$N_{18}(x) = \{y \in \mathbb{Z}^3 / d_6(x,y) \leqslant 2 \text{ et } d_{26}(x,y) \leqslant 1\}$$
$$N_{26}(x) = \{y \in \mathbb{Z}^3 / d_{26}(x,y) \leqslant 1\}$$

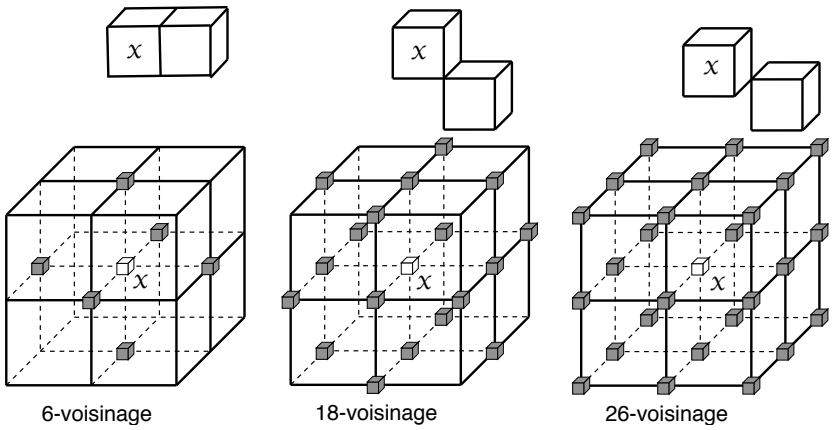


Figure A.2. Illustration des différents types de voisinage (adjacence) en 3D.

NOTE - *On note* $N_n^*(x) = N_n(x) \setminus \{x\}$.

Définition A.3.2 (*n*-voisinage) – *Les points de* $N_n^*(x)$ *forment le n-voisinage de x et card* $(N_n^*(x)) = n$, *avec* $n = 4$ *ou* 8 *pour* $\mathbb{Z}^2$ *et* $n = 6$, 18 *ou* 26 *pour* $\mathbb{Z}^3$.

Définition A.3.3 (Points *n*-adjacents ou *n*-voisins) – *Deux points x et y de* $\mathbb{Z}^2$ *ou* $\mathbb{Z}^3$ *sont n-adjacents si* $y \in N_n^*(x)$, *avec* $n = 4$ *ou* 8 *pour* $\mathbb{Z}^2$ *et* $n = 6$, 18 *ou* 26 *pour* $\mathbb{Z}^3$.

REMARQUES -
– *Deux points de* $\mathbb{Z}^2$ 4*-adjacents sont* 8*-adjacents.*
– *Deux points de* $\mathbb{Z}^3$ 6*-adjacents sont* 18*-adjacents.*
– *Deux points de* $\mathbb{Z}^3$ 18*-adjacents sont* 26*-adjacents.*

A.3.3 Chemin et connexité

Maintenant que l'on a défini les notions de points adjacents (ou voisins), on peut définir les notions de chemin et d'ensemble connexe.

Définition A.3.4 (*n*-Chemin) – *Pour tout ensemble S de points, un n-chemin dans S est une suite* $< p_i \in S, 0 \leqslant i \leqslant n >$ *de points telle que* p_i *est n-adjacent à* p_{i+1}, $0 \leqslant i < n$.

Définition A.3.5 (Connexité) – *Un ensemble de points S est n-connexe si et seulement si pour tout couple de points* $(x,y) \in S$ *il existe un n-chemin qui relie x et y.*

REMARQUES -
- *Un* 4*-chemin est un* 8*-chemin. Donc un ensemble* 4*-connexe est* 8*-connexe.*
- *Un* 6*-chemin est un* 18*-chemin. Donc un ensemble* 6*-connexe est* 18*-connexe.*
- *Un* 18*-chemin est un* 26*-chemin. Donc un ensemble* 18*-connexe est* 26*-connexe.*

Dans ce manuscrit, tous les calculs de mesures s'effectuent en 8*-connexité en 2D et en* 26*-connexité en 3D.*

A.3.4 Distance géodésique

Dans le chapitre 5, la notion de *diamètre géodésique* est évoquée. Elle est basée sur la distance géodésique, mais pour la mesurer, il faut au préalable définir la longueur d'un chemin.

Définition A.3.6 (Longueur d'un *n*-chemin) – *La longueur d'un n-chemin* $C =< c_i, 0 \leqslant i \leqslant n >$ *pour une distance d est la somme des distances entre les* c_i.
Ce qui s'écrit : $Longueur_d(C) = \sum_{i=0}^{n-1} d(c_i, c_{i+1})$.

Définition A.3.7 (Distance géodésique) – *La distance géodésique entre deux points x et y pour une distance d est la longueur du plus court n-chemin reliant x et y.*
S'il n'existe pas de chemin reliant x et y, la distance est infinie.

NOTE - *On note* d_G *la distance géodésique.*

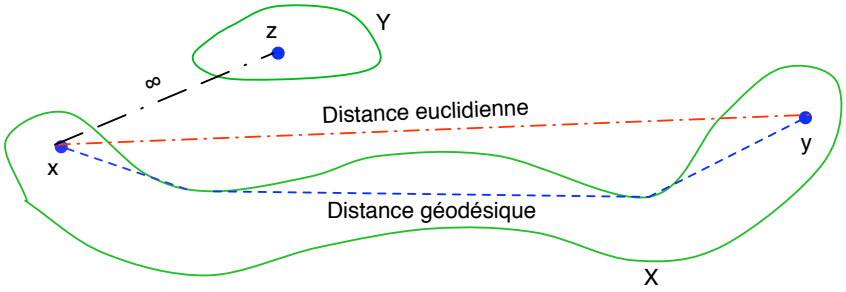


Figure A.3. Illustration de la différence entre distance géodésique et distance euclidienne dans une forme X. Illustration de la distance géodésique infinie entre deux points *x* et *z*.

A.4 Bijection

Les indices qui sont définis dans le chapitre 5 sont des fonctions.

Définition A.4.1 (Application) – *Une application (fonction) f est la donnée d'un ensemble de départ A, d'un ensemble d'arrivée B et d'une relation associant à chaque élément x de A un unique élément de B.*
Ce qui peut s'écrire : $f : A \longmapsto B \Leftrightarrow \forall x \in A, \exists! y \in B / y = f(x)$.
On dit alors que $f(x)$ *est l'image de x par* f.

On souhaite alors associer les propriétés classiques des fonctions à ces indices et notamment la propriété de bijection. Les trois définitions suivantes permettent de définir cette propriété.

Définition A.4.2 (Injection) – *Une application* $f : A \longmapsto B$ *est injective si et seulement si* $\forall x, y \in A, f(x) = f(y) \Rightarrow x = y$

Définition A.4.3 (Surjection) – *Une application* $f : A \longmapsto B$ *est surjective si et seulement si* $\forall y \in B\,, \exists x \in A / f(x) = y$

Définition A.4.4 (Bijection) – *Une application* $f : A \longmapsto B$ *est bijective si et seulement si elle est injective et surjective.*
Ce qui peut s'écrire : $\forall y \in B\,, \exists! x \in A / f(x) = y$

Une autre définition possible de la bijection est la suivante :

Définition A.4.5 (Bijection) – *Une application* $f : A \longmapsto B$ *est bijective si et seulement si il existe une application réciproque* f^{-1} *telle que* $f \circ f^{-1} = f^{-1} \circ f = I$, *avec I la fonction identité.*
Ce qui peut s'écrire : f bijective $\Leftrightarrow \exists f^{-1} / f \circ f^{-1} = f^{-1} \circ f = I$.

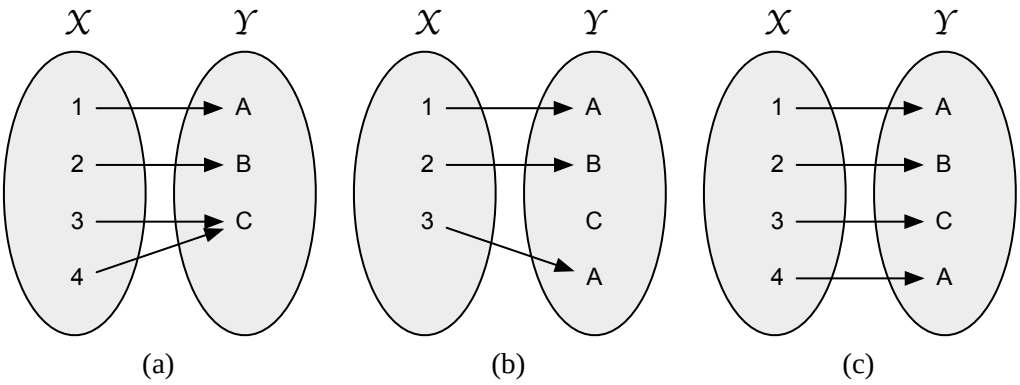


Figure A.4. Illustration des différentes propriétés des fonctions : (a) une fonction injective, mais non surjective, (b) une fonction surjective mais non injective, (c) une fonction injective et surjective, donc bijective.

REMARQUE - *La propriété de bijection permet d'affirmer que la fonction effectue une correspondance unique entre les éléments des deux ensembles de départ et d'arrivée. Elle est une propriété fondamentale qui est utilisée dans le chapitre 5.*

A.5 Convexité

Définition A.5.1 (Convexité) – *Une forme F est convexe si et seulement si pour tout couple de points $(x,y) \in F^2$, tous les points du segment $[x,y]$ sont dans F.*
Ce qui peut s'écrire : F convexe ssi $\forall (x,y) \in F^2, [x,y] \subset F$.

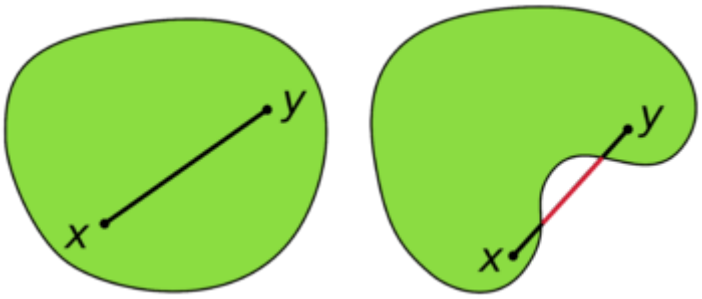


Figure A.5. A gauche une forme convexe, à droite une forme non convexe.

A.6 Axe principal

L'*Analyse en Composantes Principales* (ACP, en anglais PCA pour *Principal Component Analysis*) a été introduite en 1901 par K. Pearson, puis développée par H. Hotelling en 1933 [Hotelling 1933] : c'est une méthode d'*ordination classique*. A partir d'un ensemble de N objets dans un espace de P descripteurs, son but est de trouver une représentation dans un espace réduit de K dimensions ($K \ll P$) qui conserve *le meilleur résumé* au sens du maximum de la variance projetée.
L'ACP connaît de nombreuses applications en reconnaissance de forme (caractérisation de visage) et surtout en classification car elle permet de réduire le nombre de dimensions. Pour une forme binaire l'ACP permet de calculer l'*axe principal* (noté *AP*, figure A.6).

Définition A.6.1 (Axe principal) – *L'axe principal d'une forme F est l'axe le plus représentatif de la forme au sens de la variance projetée.*

REMARQUE - *Dans le cas d'une forme 2D, l'axe principal est l'hyperplan de la forme.*

Pour calculer l'axe principal, il faut tout d'abord calculer la matrice d'inertie du nuage de points :

$$M = \begin{bmatrix} m_X & m_{XY} \\ m_{XY} & m_Y \end{bmatrix}$$

$$\text{avec } m_X = \sum_{x,y \in F} (x-\widehat{x})^2 \text{ , } m_Y = \sum_{x,y \in F} (y-\widehat{y})^2 \text{ , } m_{XY} = \sum_{x,y \in F} (x-\widehat{x})(y-\widehat{y})$$

et $(\widehat{x},\widehat{y})$ le barycentre du nuage de points.
On calcule ensuite les valeurs propres de la matrice. Mais par construction cette matrice est symé-

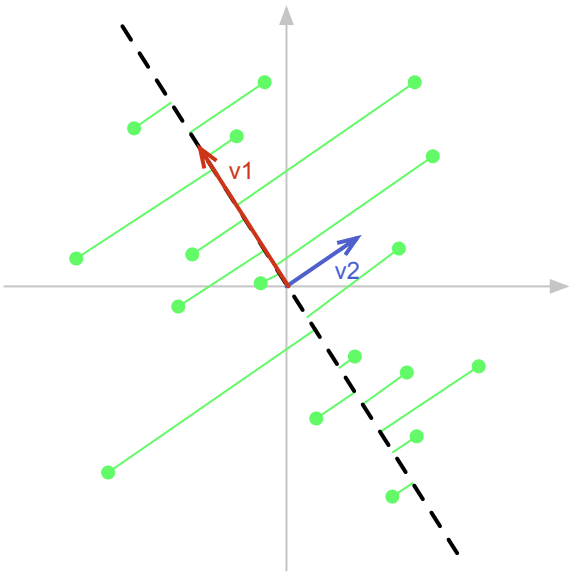


Figure A.6. Axe principal d'un nuage de points 2D et illustration de la projection sur l'axe. L'axe principal est celui qui minimise la variance des distances de projection, ici en pointillés noirs.

trique et on peut alors utiliser l'algorithme de Jacobi[3] [Smith et al. 1976; Golub and Loan 1989] pour le calcul des valeurs propres. Le vecteur propre avec la norme la plus grande est le vecteur directeur de l'axe principal.
L'axe principal et sa longueur constituent deux caractéristiques de la forme qui sont utilisées pour le calcul d'indices de forme dans les chapitres 5, 9 et B.

NOTE - *Dans ce manuscrit, on note AP⊥i le i^{ième} axe orthogonal à l'axe principal, qui est porté par le i^{ième} vecteur propre de la matrice. Cet axe est illustré en bleu sur la figure A.6.*

[3] *Numerical recipes* : http://www.nrbook.com/a/bookcpdf/c11-1.pdf

ANNEXE **B**

LISTES DES MESURES ET INDICES DE FORME, DE TEXTURE ET DE VOLUME

B.1 Liste des mesures

Le chapitre 5 a défini les notions de mesures et d'indices de forme. Les indices de forme sont des fonctions à plusieurs variables et ces variables sont les mesures. Donc pour calculer un indice sur une forme, il faut au préalable extraire des mesures sur la forme. Cette annexe dresse la liste et définit l'ensemble des mesures utilisées dans ce document.

B le barycentre. Bien que le barycentre ne soit pas une mesure à proprement parler, ce dernier est souvent utilisé dans le calcul des mesures.

P le périmètre.

A l'aire, également noté S pour "surface".

C_H pour *Convex Hull* l'enveloppe convexe.

L_{AP} la longueur de l'axe principal (cf. section A.6). On note également $L_{AP\perp i}$ la longueur du i^{ieme} orthogonal à l'axe principal.

N_{Trous} le nombre de trous de la forme.

ρ_e le plus petit disque circonscrit à la forme. Calculé à l'aide de l'algorithme décrit dans [Gärtner 1999].

ρ_i le plus grand disque inscrit dans la forme. Calculé à l'aide des cartes de distances [Thiel 1994; Remy 2001].

Définition B.1.1 (Diamètre) – *On appelle "diamètre" d'une forme (noté **D**) la plus grande distance entre deux points de la forme :*

$$D = \max_{x,y \in F} d(x,y)$$

Dans ce manuscrit, nous avons utilisé la distance Euclidienne pour calculer les diamètres.

Définition B.1.2 (Diamètre géodésique) – *Le "diamètre géodésique" (noté D_G) est la plus grande distance géodésique (cf. définition A.3.7) entre deux points de la forme :*

$$D_G = \max_{x,y \in F} d_G(x,y)$$

Définition B.1.3 (Epaisseur) – *L'épaisseur E_C issue d'une courbe C (en général le diamètre ou l'axe principal) est la plus grande distance entre un point de la forme et son projeté orthogonal sur la courbe :*

$$E_C = \max_{p \in F} d(p, Projeté_{\perp}(p, C))$$

Définition B.1.4 (Plus petit et plus grand rayons) – *Le "plus petit rayon" (resp. "plus grand") est la distance minimale (resp. maximale) entre le barycentre et un point du contour de la forme :*

$$R_{min} = \min_{p \in Contour(F)} d(B,p) \text{ et } R_{max} = \max_{p \in Contour(F)} d(B,p)$$

Définition B.1.5 (Rayon moyen) – *Le "rayon moyen" (noté μ_R) d'une forme F est la moyenne des distances entre le barycentre et les points du contour :*

$$\mu_R = \frac{1}{|Contour(F)|} \sum_{p \in Contour(F)} d(B,p)$$

Définition B.1.6 (Nombre de composantes connexes d'écart) – *Le "nombre de composantes connexes d'écart" (noté N_{Cce}) d'une forme F est le nombre de composantes connexes issues de la soustraction de la forme à son enveloppe convexe :*

$$N_{Cce} = |C_H(F) \backslash F|$$

Les notations D, D_G, E_C, R_{min}, R_{max}, μ_R désignent aussi bien les diamètres, rayons et épaisseurs que les longueurs de ces éléments lorsqu'il n'existe aucune ambiguïté. Notamment dans les indices de forme.

B.2 Liste des indices de forme 2D

Maintenant que nous avons défini l'ensemble des mesures, voici la liste des indices de forme 2D existants dans la littérature et utilisés dans cette thèse. Les intervalles de valeurs pour les indices sont calculés pour des formes convexes variant du segment au disque dans un espace continu.

Allongement par le diamètre [Santalo 1976; Coster and Chermant 1985]

$$\frac{E_D}{D} \in [0,1]$$

Allongement par les rayons [Coster and Chermant 1985]

$$\frac{\rho_i}{\rho_e} \in [0,1]$$

Allongement géodésique [Coster and Chermant 1985]

$$\frac{4}{\pi}\frac{A}{D_G^2} \in [0, \frac{\pi}{4}]$$

Circularité [Coster and Chermant 1985]

$$\frac{R_{min}}{R_{max}} \in [0,1]$$

Convexité périmétrique [Coster and Chermant 1985]

$$\frac{P(C_H)}{P} \in]0,1]$$

Convexité surfacique [Coster and Chermant 1985]

$$\frac{A}{A(C_H)} \in]0,1]$$

Déficit [Coster and Chermant 1985]

$$\pi \frac{(\rho_e - \rho_i)^2}{P^2} \in [0, \frac{\pi^2}{16}]$$

Déficit iso-périmétrique [Santalo 1976; Coster and Chermant 1985; Tuset et al. 2003]

$$4\pi \frac{A}{P^2} \in [0,1]$$

Ecart au disque inscrit [Coster and Chermant 1985]

$$\frac{\pi \rho_i^2}{A} \in [0,1]$$

Etalement de Morton [Ung et al. 2002]

$$\frac{4A}{\pi L_{AP}^2} \in [0,1]$$

Irrégularité [Chen et al. 1995]

$$\frac{A + \sqrt{\pi} \max_{p \in F} d(p,B)}{\sqrt{A}}$$

Symétrique de Bezicovitch [Fillère 1995]

$$\max_{p \in F} A\left(F \bigcap Symétrique_p(F)\right)$$

Variance circulaire [Iivarinen and Peura 1997]

$$\frac{1}{|Contour(F)|\mu_R^2} \sum_{p \in Contour(F)} (\|p - B\| - \mu_r)^2$$

Pour répondre de manière satisfaisante aux différents problèmes, nous avons dû construire de nouveaux indices qui caractérisent certains aspects des noyaux.

Ellipse par les rayons

$$\psi_{EllipseR} = \frac{\pi R_{min} R_{max}}{A} \in [0,1]$$

Ellipse par l'axe principal

$$\psi_{EllipseAP} = \frac{\pi}{4}\frac{L_{AP}L_{AP\perp}}{A} \in [0,1]$$

Ellipse par le diamètre

$$\psi_{EllipseD} = \frac{\pi}{2}\frac{E_D D}{A} \in [0,1]$$

Convexité

$$\psi_{N_{Cce}} = \frac{1}{1+N_{Cce}} \in]0,1]$$

B.3 Deux nouveaux indices de forme 2D

Lors de notre travail, nous avons également élaboré d'autres indices mais qui n'ont pas été utilisés dans le classement des noyaux.

B.3.1 Indice de courbure

Cet indice permet de déterminer la courbure générale d'une forme, mais il était impossible de l'utiliser pour les noyaux car ces derniers possédaient des déformations sur les bords et non dans leur forme générale :

$$\Psi_{Courbure} = \frac{D}{D_G} \in]0,1]$$

Il sera particulièrement intéressant pour des formes allongées qui ont subi une déformation générale et non ponctuelle comme cela est le cas dans la figure B.1. En revanche, il vaut 1 pour toutes les formes convexes et certaines formes non convexes possédant de légères déformations sur le contour.

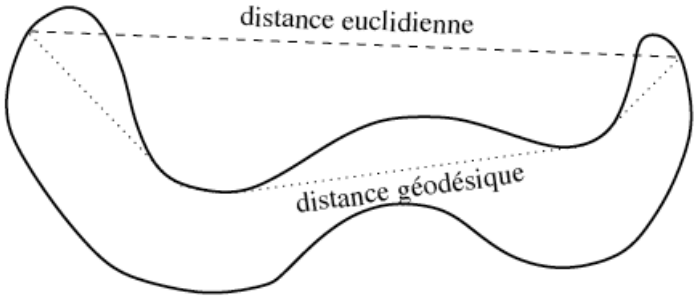


Figure B.1. Extraction des mesures pour le calcul de l'indice de courbure sur une forme allongée et déformée dans sa globalité.

B.3.2 Nouveau déficit iso-périmétrique

Lorsque nous avons utilisé le déficit iso-périmétrique dans sa forme classique (annexe B.2) nous avons constaté que les valeurs obtenues pour des formes très proches du disque étaient éloignées de 1. Pour confirmer ce problème, nous avons calculé la valeur du déficit iso-périmétrique pour

des disques de rayon allant de 1 à 250. La figure B.3 montre le résultat de cette étude et l'on peut constater que l'erreur de valeur de l'indice converge vers $0,22$.
Cette erreur est due à l'erreur entre périmètre théorique et périmètre mesuré en huit connexités sur la forme (cf. figure B.2).

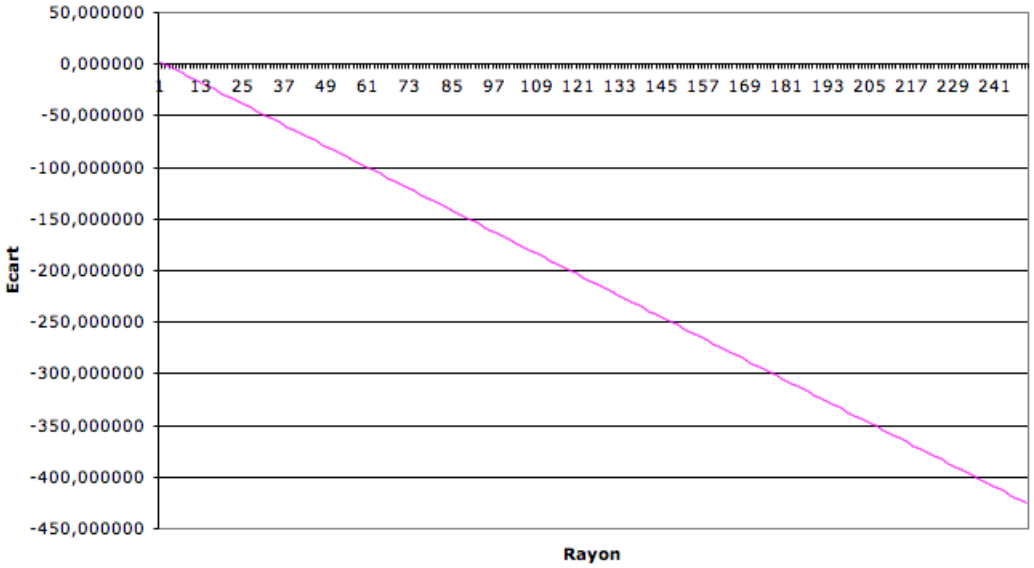


Figure B.2. Mesure de l'erreur périmètre théorique et périmètre mesuré. On peut constater que l'erreur est systématiquement proche d'une droite affine proportionnelle au rayon.

On remarque immédiatement que l'erreur de périmètre mesurée est proche d'une droite affine fonction du rayon. Pour cela, nous avons modifié la forme du déficit iso-périmétrique afin de corriger cette erreur de mesure :

$$4\pi \frac{A}{(P+\alpha R_{max})^2}$$

avec α une constante. On peut constater sur la figure B.3 que l'erreur est maintenant quasiment nulle.

B.4 Indices de forme 3D et extensions des indices de forme 2D vers la 3D

Allongement par le diamètre

$$\frac{E_D}{D}$$

Allongement par les rayons

$$\frac{\rho_i}{\rho_e}$$

Allongement géodésique

$$\frac{6}{\pi}\frac{V}{D_G^3}$$

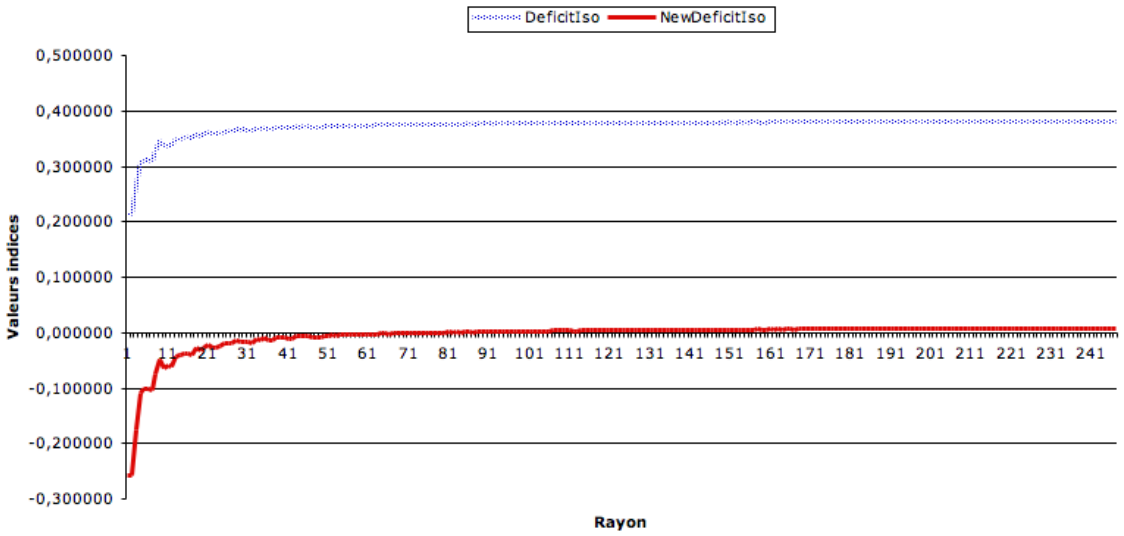


Figure B.3. Mesure des erreurs de valeur des différentes versions du déficit iso-périmétrique. On peut constater que dans sa nouvelle forme, l'erreur est quasiment nulle.

Convexité surfacique

$$\frac{S(C_H)}{S}$$

Convexité volumétrique

$$\frac{V}{V(C_H)}$$

Déficit

$$1-4\pi\frac{(\rho_e-\rho_i)^2}{S^2}$$

Déficit iso-surfacique

$$36\pi\frac{V^2}{S^3}$$

Ecart à la boule inscrite

$$\frac{3}{4}\frac{\pi\rho_i^3}{V}$$

Etalement de Morton Deux formes possibles

$$\frac{\pi}{6}\frac{V}{L_{AP}^3} \; ou \; \frac{\pi}{6}\frac{V}{L_{AP}L_{AP\perp 1}L_{AP\perp 2}}$$

Irrégularité

$$\frac{V+\sqrt{\pi}\max_{p\in F} d(p,B)}{\sqrt{V}}$$

Sphéricité

$$\frac{R_{min}}{R_{max}}$$

Symétrique de Bezicovitch

$$\max_{p \in F} V\left(F \bigcap Symetrique_p(F)\right)$$

Variance sphérique

$$\frac{1}{|Surface(F)|\mu_R^2} \sum_{p \in Surface(F)} (\|p-B\| - \mu_r)^2$$

B.5 Liste des caractéristiques Haralick

Dans le chapitre 7, il a été calculé les caractéristiques Haralick à partir de la matrice de cooccurrences. Cette section dresse la liste de l'ensemble des indices utilisés.
Pour les formules suivantes, on note :

- $p(x,y)$ l'élément de coordonnées (x,y) de la matrice.
- N_g le nombre de niveaux de gris.
- N la somme des occurrences de la matrice.
- μ_x et μ_y les moyennes en x et y des éléments de la matrice.
- σ_x et σ_y les variances en x et y des éléments de la matrice.
- $p_x(i)$ est la somme des éléments de la colonne i de la matrice.
- $p_y(i)$ est la somme des éléments de la ligne i de la matrice.
- $p_{x+y}(i)$ est la somme des éléments de la diagonale principale i de la matrice.
- $p_{x-y}(i)$ est la différence des éléments de la diagonale secondaire i de la matrice.

La moyenne

$$F_0 = \frac{1}{N} \sum_x \sum_y p(x,y)$$

L'énergie ou moment angulaire d'ordre 2. Il mesure l'uniformité de la texture. Plus la texture est uniforme, moins il y a de transitions de niveaux de gris et donc plus la somme des carrés des éléments de la matrice est faible. L'homogénéité est d'autant plus élevée que l'on retrouve souvent le même couple de pixels, ce qui est le cas lorsque le niveau de gris est uniforme ou quand il y a périodicité spatiale.

$$F_1 = \sum_x \sum_y p(x,y)^2$$

Le contraste Chaque terme de la matrice est pondéré par sa distance à la diagonale. Le contraste est élevé quand les termes éloignés de la diagonale de la matrice sont élevés, ce qui est le cas quand on a des différences importantes de niveaux de gris.

$$F_2 = \sum_{n=0}^{N_g-1} n^2 \left(\sum_{\substack{x=1 \\ |x-y|=n}}^{N_g} \sum_{y=1}^{N_g} p(x,y) \right)$$

La corrélation La corrélation est un indice qui mesure la dépendance linéaire des niveaux de gris dans l'image.

$$F_3 = \frac{1}{\sigma_x \sigma_y} \sum_x \sum_y (xy \cdot p(x,y) - \mu_x \mu_y)$$

L'écart type

$$F_4 = \sum_x \sum_y (p(x,y) - F_0)^2$$

Le moment des différences inverses

$$F_5 = \sum_x \sum_y \frac{1}{1+(x-y)^2} p(x,y)$$

La moyenne des sommes

$$F_6 = \sum_{i=2}^{2N_g-1} i \cdot p_{x+y}(i)$$

La variance des sommes

$$F_7 = \sum_{i=2}^{2N_g-1} (i - F_6)^2 p_{x+y}(i)$$

L'entropie de la somme

$$F_8 = -\sum_{i=2}^{2N_g-1} p_{x+y}(i) log(p_{x+y}(i))$$

L'entropie Elle rend compte de la complexité de l'image en fournissant un indicateur sur le désordre que peut présenter sa texture. L'entropie est faible si on rencontre souvent le même couple de pixels dans l'image et forte si chaque couple est peu représenté (texture aléatoire).

$$F_9 = \sum_x \sum_y p(x,y) log\,(p(x,y))$$

La variance des différences

$$F_{10} = \sum_{i=2}^{2N_g-1} (i - \mu_{x-y})^2 p_{x+y}(i) \text{ avec } \mu_{x-y} = \frac{1}{N} \sum_{i=0}^{N} p_{x-y}(i)$$

L'entropie des différences

$$F_{11} = -\sum_{i=2}^{2N_g-1} p_{x-y}(i) log(p_{x-y}(i))$$

L'inertie

$$\frac{1}{N} \sum_x \sum_y (x-y)^2 p(x,y)$$

L'homogénéité

$$\frac{1}{N} \sum_x \sum_y \frac{1}{1+|x-y|} p(x,y)$$

La dissimilarité

$$\frac{1}{N} \sum_x \sum_y |x-y| p(x,y)$$

B.6 Les indices de texture pour les run length et size zone matrix

Voici la liste des indices de texture [Xu et al. 2004] que nous avons utilisés pour construire les différents modèles basés sur les *run length* et *size zone matrix*. Toutes les descriptions le sont pour les longueurs de segments de la *run length matrix*, mais sont également applicables pour les surfaces des régions de la *size zone matrix*.
Pour les formules suivantes, on note :

- $p(i,j)$ l'élément de coordonnées (i,j) de la matrice.
- N le nombre de segments de la matrice, $N = \sum_{i,j} p(i,j)$.

Short Run Emphasis mesure la distribution des segments courts. L'inverse du carré de la longueur des segments rend la valeur très sensible aux segments courts (texture aléatoire) et beaucoup moins sensible aux segments longs (texture homogène, périodique).

$$SRE = \frac{1}{N}\sum_i\sum_j \frac{p(i,j)}{j^2}$$

Long Run Emphasis mesure la distribution des segments longs (inverse de SRE). On peut remarquer que sa construction possède une pondération inverse de la SRE. La valeur est donc d'autant plus grande que la texture possède des segments de longueur importante (texture homogène, périodique).

$$LRE = \frac{1}{N}\sum_i\sum_j p(i,j) * j^2$$

Low Gray Level Run Emphasis mesure la distribution des segments de faible intensité (valeur de niveaux de gris faible). La valeur sera d'autant plus grande que le nombre de segments de faible intensité sera important.

$$LGRE = \frac{1}{N}\sum_i\sum_j \frac{p(i,j)}{i^2}$$

High Gray Level Run Emphasis mesure la distribution des segments de haute intensité (valeur de niveaux de gris élevée, inverse de LGRE). La valeur est d'autant plus grande que le nombre de segments de haute intensité est important.

$$HGRE = \frac{1}{N}\sum_i\sum_j p(i,j) * i^2$$

Short Run Low Gray Level Emphasis mesure la distribution des segments courts avec une intensité faible, c'est une combinaison des deux indices SRE et LGRE. Les inverses des carrés pénalisent l'influence des segments longs et d'intensité élevée.

$$SRLGE = \frac{1}{N}\sum_i\sum_j \frac{p(i,j)}{i^2 * j^2}$$

Short Run High Gray Level Emphasis mesure la distribution des segments courts et d'intensité élevée (combinaison des indices SRE et HGRE). L'inverse du carré de la longueur pénalise les segments longs et le carré de l'intensité favorise les segments de haute intensité.

$$SRHGE = \frac{1}{N}\sum_i\sum_j \frac{p(i,j) * i^2}{j^2}$$

Long Run Low Gray Level Emphasis mesure la distribution des segments longs et de faible intensité (combinaison de LRE et LGRE). L'inverse du carré de l'intensité pénalise les segments d'intensité élevée et la pondération par le carré de la longueur pénalise les segments courts.

$$LRLGE = \frac{1}{N}\sum_i\sum_j \frac{p(i,j)*j^2}{i^2}$$

Long Run High Gray Level Emphasis mesure la distribution des segments longs et d'intensité élevée (combinaison des indices LRE et HGRE). La valeur est d'autant plus élevée que les segments sont longs (pondération par le carré de la longueur) et d'intensité élevée (pondération par le carré de l'intensité).

$$LRHGE = \frac{1}{N}\sum_i\sum_j p(i,j)*i^2*j^2$$

Gray Level Non Uniformity mesure la non uniformité des niveaux de gris. La pondération par le carré de la somme des valeurs d'une ligne permet de détecter des répartitions non uniformes entre les niveaux de gris. Si un niveau de gris possède un grand nombre de segments, le carré de la somme de la ligne influe de manière importante sur la valeur.

$$GLNU = \frac{1}{N}\sum_i\left(\sum_j p(i,j)\right)^2$$

Run Length Non Uniformity mesure la non uniformité des longueurs de segments. Même principe que pour l'indice GLNU, mais avec les longueurs de segments. Plus les longueurs de segments sont répartis entre les niveaux de gris, plus la valeur est faible.

$$RLNU = \frac{1}{N}\sum_j\left(\sum_i p(i,j)\right)^2$$

Run Percentage

$$RPC = N\sum_i\sum_j \frac{1}{p(i,j)*j}$$

Par la suite, nous avons conçu deux nouveaux indices afin de mesurer les variances des tailles et des niveaux de gris des régions. Dans ces formules, N désigne la hauteur de la matrice (le nombre de niveaux de gris) et S la largeur de la matrice (la surface de la plus grande région).

Ecart type des intensités mesure l'écart type pondéré (par la taille) des intensités des régions. Plus les régions sont grandes et l'écart entre leurs niveaux de gris élevé, plus la valeur de l'indice est élevée.

$$\psi_N = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}(n*M(n,s)-\mu_N)^2} \text{ avec } \mu_N = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} n*M(n,s)$$

Ecart type des tailles mesure l'écart type pondéré (par la taille) entre les tailles des régions. Plus l'écart entre la taille des plus grandes régions et des plus petites est grand, plus la valeur de l'indice est élevée.

$$\psi_S = \sqrt{\frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S}(s*M(n,s)-\mu_S)^2} \text{ avec } \mu_S = \frac{1}{N*S}\sum_{n=1}^{N}\sum_{s=1}^{S} s*M(n,s)$$

ANNEXE C

MARQUEURS FLUORESCENTS ET MICROSCOPE

Le chapitre 1 a présenté les enjeux et surtout le déroulement du travail des experts. Durant ce travail, on avons vu que les experts utilisaient des marqueurs et différents microscopes afin de visualiser certains éléments caractéristiques des noyaux.
Cette annexe présente succinctement les différents marqueurs utilisés par les experts ainsi que le principe du fonctionnement du microscope à fluorescence.

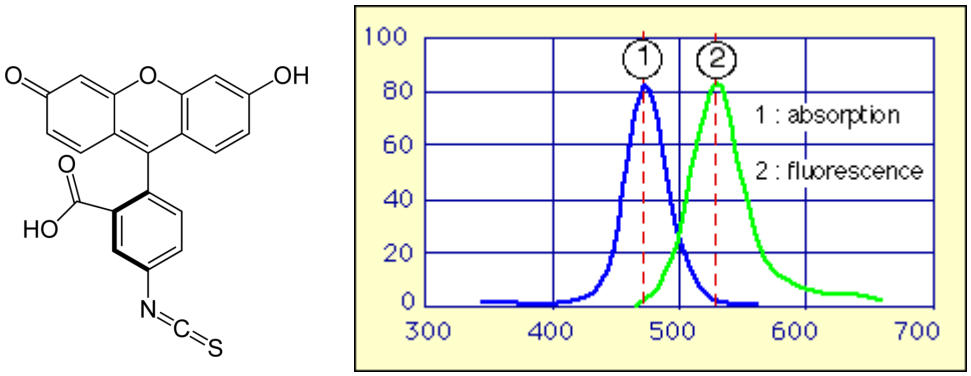
C.1 Les marqueurs

Avant de pouvoir présenter la définition exacte des marqueurs, il est nécessaire de donner deux définitions qui vont permettre de définir les marqueurs.

Définition C.1.1 (Fluorescence) – *La fluorescence est une émission lumineuse provoquée par diverses formes d'excitation autres que la chaleur. On parle parfois de "lumière froide".*

Définition C.1.2 (Fluorochrome) – *Un fluorochrome ou fluorophore est une substance chimique capable d'émettre de la lumière de fluorescence après excitation.*

Les marqueurs utilisés par les experts sont des fluorochromes. Ils réagissent à la présence de certaines substances, puis émettent une lumière fluorescente lorsqu'ils sont excités. Leur excitation est provoquée par une lumière dans une longueur d'onde qui leur est spécifique. Ils possèdent la propriété d'absorber une partie des longueurs d'ondes de la lumière excitatrice (spectre d'absorption), puis émettent une longueur d'onde qui leur est propre (spectre d'émission).
Le principal inconvénient de ces marqueurs est que la fluorescence n'est pas permanente : l'intensité de la fluorescence diminue avec le temps jusqu'à devenir indécelable.



(a) (b)

Figure C.1. Formule (a) et spectre (b) du FITC.

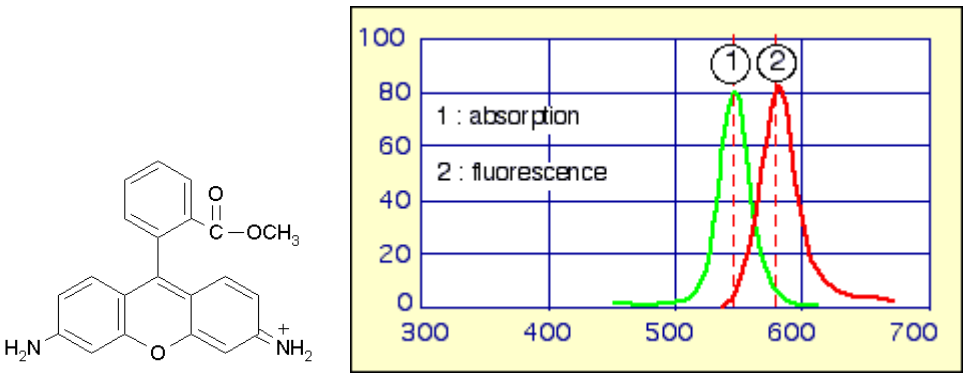
C.1.1 Le FITC

Les colorants fluorescents verts les plus connus sont des dérivés de la *fluorescéine* comme le FITC (Fluoresceine Iso Thio Cyanate, figure C.1).

Dans notre problème, le FITC marque la présence des lamines A et C, mais il est également utilisé pour marquer différentes biomolécules comme les immunoglobulines, les lectines, les différentes protéines (dont les lamines AC), les peptides, les acides nucléiniques, les polynucléotides, les oligo et polysaccharides. Son spectre d'absorption est dans le bleu (max 490*nm*) et il restitue une lumière verte (max 520*nm*).

C.1.2 Le TRITC

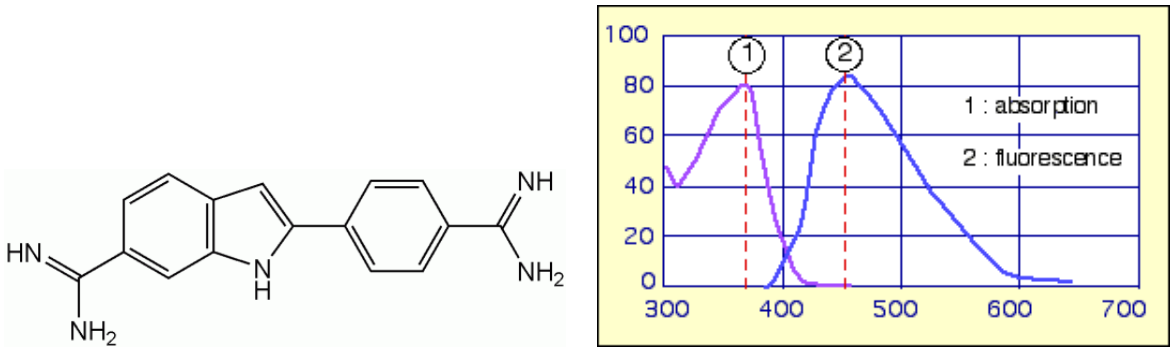
Les colorants fluorescents en rouge les plus connus sont des dérivés de la rhodamine comme le TRITC (Tetra methyl Rhodamine Iso Thio Cyanate, cf. figure C.2).



(a) (b)

Figure C.2. Formule (a) et spectre (b) de la rhodamine.

La rhodamine absorbe les radiations vertes (max 541*nm*) et restitue une fluorescence rouge (max 572*nm*). Elle marque essentiellement la présence des lamines B, mais n'est pas utilisée lors du diagnostic des noyaux.



(a) (b)

Figure C.3. Formule (a) et spectre (b) du DAPI.

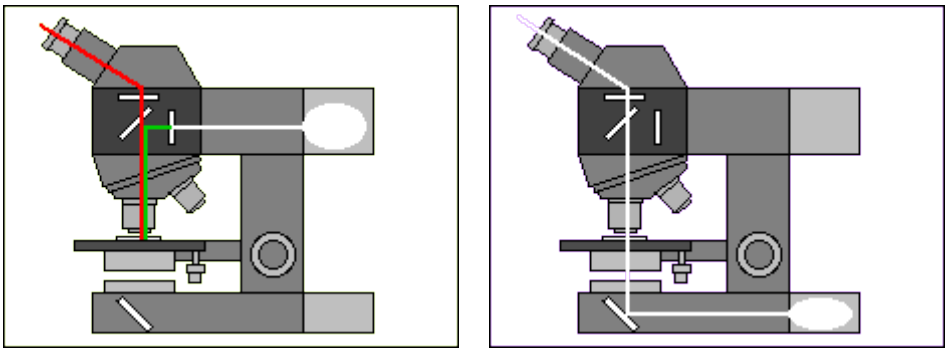
C.1.3 Le DAPI

Le DAPI (Di Aminido Phenyl lndol) fluorochrome est utilisé en cytochimie et se fixe spécifiquement sur l'ADN qui est essentiellement présent dans les noyaux (chromatine, cf. figure C.3).

Eclairé en lumière violette (max 372*nm*), il émet une fluorescence bleue (max 456*nm*). Bien que parfois utilisé dans le diagnostic, nous ne disposons pas d'expertise utilisant le DAPI, ce qui l'exclut de notre étude.

C.2 Le microscope à fluorescence

Les microscopes classiques éclairent la lame contenant la préparation avec une lumière blanche, mais dans le cas d'un microscope à fluorescence, il est nécessaire d'éclairer la préparation avec une lumière possédant une longueur d'onde particulière (figure C.4). En effet, les marqueurs ne réagissent qu'à une longueur d'onde qui leur est spécifique.



(a) (b)

Figure C.4. Schéma optique du microscope à fluorescence (a) et classique (b).

Afin de permettre cet éclairage différent, le microscope à fluorescence se compose schématiquement de sept parties différentes (figure C.5) :

1. Une lampe à arc : elle fournit la source lumineuse de l'éclairage.
2. Un filtre d'excitation : il permet la sélection des radiations absorbées par le fluorochrome.
3. Un miroir dichroïque : il réfléchit les radiations absorbables vers l'échantillon et ne laisse passer par transmission que les radiations émises par le fluorochrome.
4. Un objectif.
5. Une préparation : ce que l'on souhaite observer. Dans notre cas les noyaux de cellules.
6. Un filtre d'émission : il ne laisse passer par transmission que les radiations émises par le fluorochrome.
7. Un oculaire.

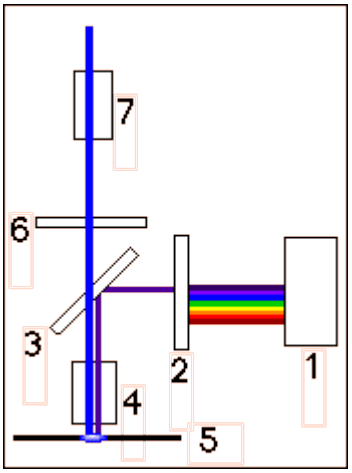


Figure C.5. Schéma optique du microscope à fluorescence. Dans cet exemple, le filtre d'excitation laisse passer la lumière violette et le marqueur fluoresce en bleu.

BIBLIOGRAPHIE

ALBREGTSEN, F. 1995. Statistical texture measures computed from gray level run length matrices. Tech. rep., University of Oslo, November.

AMIT, Y., AND GEMAN, D. 1997. Shape quantization and recognition with randomized trees. *Neural computation 9*, 7, 1545–1588.

AMMOR, O., LACHKAR, A., SLAOUI, K., AND RAIS, N. 2006. New efficient approach to determine the optimal number of clusters in overlapping cases. In *IEEE on Advances in Cybernetic Systems*, 26–31.

AMMOR, O., LACHKAR, A., SLAOUI, K., AND RAIS, N. 2008. Optimisation of pattern recognition in textile field. *Journal of the Textile Institute 99*, 3 (June), 227–233.

ANDRIAMAMPIANINAO, L., STAMON, G., SIMON, J., AND POULENAR, M. 1994. Transformations géométriques et extraction de caractéristiques d'une forme 2d en représentation par primitives angulaires. In *African Conference on research in computer science*, 317–331.

ARAGON, C. R., ARAGON, D. B., AND BERKELEY, L. 2007. A fast contour descriptor algorithm for supernova image classification. In *Real-time image processing*, Society of Photo-Optical Instrumentation Engineers, San Jose, CA, USA, vol. 6496, 649607.1–649607.12.

BENJELLOUN, M., TÉLLEZ, H., AND MAHMOUDI, S. 2006. A new template matching method for vertebrae contours detection in x-ray images. In *Visualization, Imaging and Image Processing*, vol. 1, 1119–1124.

BERKSON, J. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association 39*, 357–365.

BLUM, H. 1964. A transformation for extracting new descriptors of shape. In *Symposium on Models for the Perception of Speech and Visual Form*, M.I.T. press, W. Wathendunn, Ed., 139–146.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. 1984. *Classification And Regression Trees*. CRC Press.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning 24*, 2, 123–140.

BREIMAN, L. 2001. Random forests. *Machine Learning 45*, 1, 5–32.

BRIBIESCA, E., AND GUZMAN, A. 1980. How to describe pure form and how to measure differences in shapes using shape numbers. *Pattern recognition 12*, 2, 101–112.

BRODATZ, P. 1966. *Textures : A Photographic Album for Artists and Designers*. Dover.

BRUNO, O. M., JR, R. M. C., CONSULARO, L. A., AND COSTA, L. F. 1998. Automatic feature selection for biological shape classification in synergos. In *International Symposium on Computer Graphics, Image Processing and Vision*, IEEE Computer Society Press, 363–370.

BUHLMANN, P., AND HOTHORN, T. 2007. Boosting algorithms : Regularization prediction and model fitting. *Statistical Science 22*, 4, 477–405.

CAKMAKOV, D., RADEVSKI, V., BENNANI, Y., AND DEJANGORGEVIK. 2002. Decision fusion and reliability control in handwritten digit recognition system. In *Journal of Computing and Information Technology*, 283–293.

CASTANON, C. A., FRAGA, J. S., FERNANDEZ, S., GRUBER, A., AND DA F. COSTA, L. 2007. Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus eimeria. *Pattern Recognition 40*, 7, 1899–1910.

CHANG, C., DU, Y., WANG, J., GUO, S., AND THOUIN, P. 2006. Survey and comparative analysis of entropy and relative entropy thresholding techniques. *Vision, Image and Signal Processing, IEE Proceeding 153*, 6 (December), 837–850.

CHASSERY, J.-M., AND MONTANVERT, A. 1991. *Géométrie discrète*. Hermes, Paris.

CHEN, G. Y., AND KEGL, B. 2009. Invariant pattern recognition using dual tree complex wavelets and fourier features. *Pattern Recognition 42*, 9 (September), 2013–2019.

CHEN, Y. Q., DIXON, M. S., AND THOMAS, D. W. 1995. Statistical geometrical features for texture classification. *Pattern Recognition 28*, 4, 537–552.

CHINGA, G., JOHNSEN, P. O., DOUGHERTY, R., BERLI, E. L., AND WALTER, J. 2007. Quantification of the 3d microstructure of sc surfaces. *Journal of Microscopy 227* (September), 254–265.

CHU, A., SEHGAL, C. M., AND GREENLEAF, J. F. 1990. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters 11*, 6, 415–419.

COLLEWET, C. 1999. *Contribution à l'élargissement du champ applicatif des asservissements visuels 2D*. PhD thesis, Université de Rennes I.

COSTA, J.-P. D. 2001. *Analyse statistiques de textures directionnelles*. PhD thesis, Université Bordeaux I.

COSTER, M., AND CHERMANT, J.-L. 1985. *Précis d'analyse d'images*. Editions du CNRS.

COX, D. R., AND HINKLEY, D. V. 1978. *Theoretical Statistics*, 3rd ed. Chapman & Hall/Crc.

DAVIS, L., JOHNS, S., AND AGGARWAL, J. 1979. Texture analysis using generalized cooccurrence matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1*, 3 (July), 251–259.

DIDAY, E., LEMAIRE, J., POUGET, J., AND TESTU, F. 1982. *Eléments d'analyse des données*. Dunod-Bordas.

DIDAY, E. 1972. Optimisation en classification automatique et reconnaissance des formes. *Revue française d'Automatique, Informatique et Recherche Opérationnelle 3*, 61–95.

DIETTERICH, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation 10*, 1895–1923.

DO, T., PHAM, N., LENCA, P., AND LALLICH, S. 2008. Expérimentation de l'entropie décentrée pour le traitement des classes déséquilibrées en induction par arbres. In *4ème Atelier Qualité des Données et des Connaissances*, 39–50.

DOMINGO, P. 1999. Metacost : A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, ACM Press, New York, NY, USA, 155–164.

DRUMMOND, C., AND HOLTE, R. C. 2003. C4.5, class imbalance, and cost sensitivity : Why undersampling beats over-sampling. In *International Conference on Machine Learning, workshop on learning from imbalanced datasets*.

DUDA, R. O., AND HART, P. E. 1972. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM 15*, 1 (January), 11–15.

DUDA, R. O., HART, P. E., AND STARK, D. G. 2000. *Pattern Classification*. Wiley Interscience.

EFRON, B. 1979. Bootstrap methods : Another look at the jackknife. *The Annals of Statistics 7*, 1 (January), 1–26.

ERIKSSON, M., BROWN, T. W., GORDON, L., GLYNN, M. W., SINGER, J., SCOTT, L., ERDOS, M. R., ROBBINS, C. M., MOSES, T. Y., BERGLUND, P., DUTRA, A., PAK, E., DURKIN, S., CSOKA, A. B., BOEHNKE, M., GLOVER, T. W., AND COLLINS, F. S. 2003. Recurrent de novo point mutations in lamin a cause hutchinson-gilford progeria syndrome. *Nature 423* (April), 6937.

FEARS, T. R., BENICHOU, J., AND GAIL, M. H. 1996. A reminder of the fallibility of the wald statistic. *The American Statistician 50*, 3 (August), 226–227.

FILLÈRE, I. 1995. *Outils mathématiques pour la reconnaissance de formes*. PhD thesis, Université de St Etienne.

FIX, E., AND HODGES, J. 1951. Discriminatory analysis : Nonparametric discrimination : Consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine, Randolph Field, Texas.

FIX, E., AND HODGES, J. 1952. Discriminatory analysis : Nonparametric discrimination : Consistency properties. Tech. Rep. 11, USAF School of Aviation Medicine, Randolph Field, Texas.

FORGY, E. W. 1965. Cluster analysis of multivariate data : efficiency vs interpretability of classifications. *Biometrics 21*, 768–769.

FREEMAN, H. 1961. On the encoding of arbitrary geometric configurations. *TC 10*, 2 (June), 260–268.

FREEMAN, H. 1974. Computer processing of line-drawing images. In *ACM Computer survey*, vol. 6, 57–97.

FREUND, Y., AND SCHAPIRE, R. E. 1999. A short introduction to boosting. *Japanese Society of Artificial Intelligence 14*, 5 (September), 771–780.

GAGALAWICZ, A. 1983. *Vers un modèle de texture*. PhD thesis, Université Paris VI.

GALLER, B. A., AND FISCHER, M. J. 1964. An improved equivalence algorithm. *Communications of the ACM 7*, 5 (May), 301–303.

GALLOWAY, M. M. 1975. Texture analysis using grey level run lengths. *Computer Graphics Image Processing 4* (July), 172–179.

GÄRTNER, B. 1999. Fast and robust smallest enclosing ball. In *European Symposium on Algorithms (ESA)*, Springer-Verla, London, U, vol. 1643, 325–338.

GINI, C. 1921. Measurement of inequality of income. *Economic Journal 31*, 22–43.

GLOVER, F. 1986. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research 13*, 5 (May), 533–549.

GLOVER, F. 1990. Tabu search : A tutorial. *Interfaces 20*, 1, 74–94.

GOLUB, G., AND LOAN, C. V. 1989. *Matrix computations*. Johns Hopkins University Press, Baltimore.

GOSH, B. K. 1979. A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association 74*, 894–900.

HALKIDI, M., AND VAZIRGIANNIS, M. 2001. Clustering validity assessment : Finding the optimal partitioning of a data set. In *IEEE Internationnal Conference On Data Mining*, IEEE Computer Society, Washington, DC, USA, 187–194.

HARALICK, R. M., SHANMUGAM, K., AND DINSTEIN, I. 1973. Textural features for image classification. In *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, 610–621.

HARALICK, R. M. 1979. Statistical and structural approaches to texture. In *Proceedings of the IEEE*, vol. 67, 786–804.

HARTIGAN, J. A., AND WONG, M. A. 1979. A K-means clustering algorithm. *Applied Statistics 28*, 1, 100–108.

HEUTTE, L., PAQUET, T., MOREAU, J., LECOURTIER, Y., AND OLIVIER, C. 1998. A structural/statistical feature based vector for handwritten character recognition. In *IEEE Transaction on Pattern Recognition Letters*, Elsevier Science, vol. 19, 629–641.

HOSMER, D., AND LEMESHOW, S. 1989. *Applied Logistic Regression*. John Wiley & Sons, Toronto.

HOTELLING, H. 1933. Analysis of a complex of statistical variables with principal components. In *Journal of Educational Psychology*.

HOUGH, P. 1962. Method and means for recognizing complex patterns. In *U.S. Patent 3,069,654*.

IIVARINEN, J., AND PEURA, M. 1997. Efficiency of simple shape descriptors. In *International Workshop on Visual Form*, 28–30.

IIVARINEN, J., AND VISA, A. 1996. Shape recognition of irregular objects. In *Intelligent Robots and Computer Vision*, D. Casasent, Ed., 25–32.

IIVARINEN, J., PEURA, M., SÄRELÄ, J., AND VISA, A. 1997. Comparison of combined shape descriptors for irregular objects. In *British Machine Vision Conference (BMVC)*, vol. 2, 430–439.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. 1999. Data clustering : A review. *ACM Computing Surveys 31*, 3 (September), 264–323.

JAIN, A. K., DUIN, R. P. W., AND MAO, J. 2000. Statistical pattern recognition : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1, 4–37.

JALBA, A. C., WILKINSON, M. H., AND ROERDINK, J. B. 2004. Morphological hat-transform scale spaces and their use in pattern classification. In *Pattern recognition*, E. Science, Ed., vol. 37, 901–915.

JALBA, A. C., WILKINSON, M. H. F., AND ROERDINK, J. B. T. 2006. Shape representation and recognition through morphological curvature scale spaces. In *IEEE Transactions on Image Processing*, Institute of Electrical and Electronics Engineers, New York, NY, ETATS-UNIS, I. of Electrical and E. Engineers, Eds., vol. 15, 31–341.

JAPKOWICZ, N., AND STEPHEN, S. 2002. The class imbalanced problem : A systematic study. *Intelligent Data Analysis Journal 6*, 5 (November), 429–449.

JAPKOWICZ, N. 2000. The class imbalanced problem : Significiance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, 111–117.

JAPKOWICZ, N. 2000. Learning from imbalanced data sets : A comparison of various strategies. In *AAAI Workshop on Learning from Imbalanced Data Sets*, AAAI Press, 10–15.

KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R., AND WU, A. Y. 2002. An efficient k-means clustering algorithm : Analysis and implementation. In

IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Computer Society, Washington, DC, USA, vol. 24, 881–892.

KAPUR, J. N., SAHOO, P. K., AND WONG, A. K. C. 1985. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing 29*, 3, 273–285.

KÉGL, B., AND KRZYZAK, A. 2002. Piecewise linear skeletonization using principal curves. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 59–74.

KIM, D.-J., PARK, Y.-W., AND PARK, D.-J. 2001. A novel validity index for determination of the optimal number of clusters. In *IEICE Transaction on Information and System*, vol. E84-D, 281–285.

KIMA, D.-W., LEEA, K. H., AND LEE, D. 2004. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition 37*, 10, 2009–2025.

KOHAVI, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137–1145.

KPALMA, K., AND RONSIN, J. 2003. A multi-scale curve smoothing for generalised pattern recognition (msgpr). In *Seventh International Symposium on Signal Processing and its Applications*, ISSPA, 427–430.

KPALMA, K., AND RONSIN, J. 2006. Multiscale contour description for pattern recognition. In *IEEE Transactions on Pattern Recognition Letters*, Elsevier, vol. 27, 1545–1559.

KUBAT, M., AND MATWIN, S. 1997. Addressing the curse of imbalanced training sets : One-sided selection. In *In Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann, 179–186.

LALLICH, S., LENCA, P., AND VAILLANT, B. 2007. Construction d'une entropie décentrée pour l'apprentissage supervisé. In *Qualité des Données et des Connaissances (QDC)*, 45–54.

LAZEBNIK, S., SCHMID, C., AND PONCE, J. 2005. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence 27*, 8, 1265–1278.

LENCA, P., LALLICH, S., DO, T.-N., AND PHAM, N.-K. 2008. A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi, Eds., vol. 5012, 634–643.

LI, X., MAK, M. W., AND LI, C. K. 1999. Determining the optimal number of clusters by an extended rpcl algorithm. *Journal of Advanced Computational Intelligence and Intelligent Informatics 3*, 6 (April), 467–473.

LIU, X.-Y., AND ZHOU, Z.-H. 2006. The influence of class imbalance on cost-sensitive learning : An empirical study. In *International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, 970–974.

LIU, J., VAN DER PUTTEN, P., HAGEN, F., CHEN, X., AND BOEKHOUT, T. 2006. Detecting virulent cells of cryptococcus neoformans yeast : Clustering experiments. In *International Conference on Pattern Recognition (ICPR)*, IEEE Computer Society, Washington, DC, USA, vol. 1, 1112–1115.

LIU, X.-Y., WU, J., AND ZHOU, Z.-H. 2006. Exploratory under-sampling for class-imbalance learning. *IEEE International Conference on Data Mining 0*, 965–969.

LIU, A., GHOSH, J., AND MARTIN, C. 2007. Generative oversampling for mining imbalanced datasets. In *International Conference on Data Mining*, CSREA Press, R. Stahlbock, S. F. Crone, and S. Lessmann, Eds., 66–72.

LIU, X.-Y., WU, J., AND ZHOU, Z.-H. 2009. Exploratory under-sampling for class-imbalance learning. *IEEE Transactions on Systems, Man and Cybernetics*, 1–14.

LOHMANN, G. 1995. Analysis ans synthesis of textures : a cooccurrence-based approach. In *CG*, vol. 19, 29–36.

LORIGO, L. M., AND GOVINDARAJU, V. 2006. Off-line arabic handwriting recognition : A survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Washington, DC, USA, vol. 28, 712–724.

MACQUEEN, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, L. M. L. C. . J. Neyman, Ed., vol. 1, 281–297.

MARCELLIN, S., ZIGHED, D.-A., AND RITSCHARD, G. 2008. Evaluating decision trees grown with asymmetric entropies. In *Foundations of Intelligent Systems, ISMIS*, 58–67.

MARCELLIN, S., ZIGHED, D.-A., AND RITSCHARD, G. 2008. Évaluation des critères asymétriques pour les arbres de décision. In *Extraction et Gestion des Connaissances (EGC)*, vol. 11, 31–38.

MARI, J.-L. 2002. *Modélisation de formes complexes intégrant leurs caractéristiques globales et leurs spécificités locales*. PhD thesis, Université de la Méditerranée.

MARTENS, H., AND DARDENNE, P. 1998. Validation and verification of regression in small data sets. *Chemometrics and intelligent laboratory systems 44*, 1-2, 99–121.

MATHERON, G., AND SERRA, J. 2002. The birth of mathematical morphology. In *Proceedings of VIth International Symposium on Mathematical Morphology*, Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia, H. Talbot and R. Beare, Eds., 1–16.

MAVROMATIS, S. 2001. *Analyse de texture et Visualisation scientifique*. PhD thesis, Université de la Méditerranée.

MCCULLOCH, W. S., AND PITTS, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

MCFADDEN. 1973. Conditional logit analysis of qualitative choice behaviour. *Frontiers in econometrics*, 105–142.

MICHEL, L., AND HENTENRYCK, P. V. 2004. A simple tabu search for warehouse location. *European Journal of Operational Research 157*, 576–591.

MILLIGAN, G. W., AND COOPER, M. C. 1988. A study of standardization of variables in cluster analysis. *Journal of Classification 5*, 2 (September), 181–204.

MOORE, D. S., AND MCCABE, G. P. 1998. *Introduction to the Practice of Statistics*, 3rd ed. W. H. Freeman, New York, NY, USA.

MORGAN, J. N., AND SONQUIST, J. A. 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*, 415–435.

NETTEN, H., VAN VLIET, L. J., VROLIJK, H., SLOOS, W. C., TANKE, H. J., AND YOUNG, I. T. 1996. Fluorescent dot counting in interphase cell nuclei. In *Bioimaging*, IOP, vol. 4, 93–106.

NEWCOMBE, R. G. 1998. Two-sided confidence intervals for the single proportion : comparison of seven methods. *Statistics in Medicine 17*, 8 (December), 857–872.

ORLOV, N., SHAMIR, L., MACURA, T., JOHNSTON, J., ECKLEY, D., AND GOLDBERG, I. 2008. Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters 29*, 11 (October), 1684–1693.

PARKKINEN, J., SELKAINAHO, K., AND OJA, E. 1990. Detecting texture periodicity from the cooccurrence matrix. In *IEEE Transaction on Pattern Recognition Letters*, Elsevier, Amsterdam, PAYS-BAS (1982), vol. 11, 43–50.

PERNER, P., PERNER, H., AND MÜLLER, B. 2002. Mining knowledge for hep-2 cell image classification. *Journal of Artificial Intelligence in Medicine 26*, 161–173.

PUN, T. 1980. A new method for grey-level picture thresholding using the entropy of the histogram. *Signal Processing 2*, 223–237.

PUN, T. 1981. Entropic thresholding : a new approach. *Computer Graphics Image Processing 16*, 210–239.

REMY, E., AND THIEL, E. 2002. Medial axis for chamfer distances : computing look-up tables and neighbourhoods in 2d or 3d. In *IEEE Transaction on Pattern Recognition Letters*, Elsevier, vol. 23, 649–661.

REMY, E. 2001. *Normes de Chanfrein et axe médian dans le volume discret*. PhD thesis, Faculté des sciences de Luminy.

RITSCHARD, G., ZIGHED, D. A., AND MARCELLIN, S. 2007. Données déséquilibrées, entropie décentrée et indice d'implication. In *Nouveaux apports théoriques à l'analyse statistique implicative et applications*, Département de Mathématiques, Université Jaume I, ASI4, 315–327.

ROSENBLATT, F. 1958. The perceptron : probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 386–408.

ROSIN, P. L. 2004. Measuring sigmoidality. In *IEEE Transaction on Pattern Recognition*, vol. 37, 1735–1744.

SAHOO, P. K., SOLTANI, S., WONG, A. K., AND CHEN, Y. C. 1988. A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing 41*, 2, 233–260.

SANDRE-GIOVANNOLI, A. D., BERNARD, R., CAU, P., NAVARRO, C., AMIEL, J., BOCCACCIO, I., LYONNET, S., STEWART, C. L., MUNNICH, A., MERRER, M. L., AND LEVY, N. 2003. Lamin a truncation in progeria. *Science 300*, 5628 (April), 2055.

SANTALO, L. 1976. *Integral Geometry and Geometric Probability*. Addison Wesley.

SAPORTA, G. 2006. *Probabilité, analyse des données et statistiques*, 2nd ed. Edition Technip.

SERRA, J. 1982. *Image Analysis and Mathematical Morphology*. Academic Press.

SHAKHNAROVICH, G., DARRELL, T., AND INDYK, P. 2006. *Nearest-Neighbor Methods in Learning and Vision : Theory and Practice*. The MIT Press, March.

SIDDIQI, K., SHOKOUFANDEH, A., DICKINSON, S. J., AND ZUCKER, S. 1999. Shock graphs and shape matching. In *International Journal of Computer Vision*, Kluwer Academic, vol. 35, 13–32.

SMITH, B., BOYLE, J., DONGARRA, J., GARBOW, B., IKEBE, Y., KLEMA, V., AND MOLER, C. 1976. *Matrix Eigensystem Routines*, 2nd ed ed., vol. 6. Springer-Verlag, Berlin : Heidelberg.

SOLTANIAN ZADEH, H., RAFIEE RAD, F., AND POURABDOLLAH NEJAD, S. 2004. Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms. *Pattern Recognition 37*, 10 (October), 1973–1986.

SOLTANZADEH, H., AND RAHMATI, M. 2004. Recognition of persian handwritten digits using image profiles of multiple orientations. In *IEEE Transactions on Pattern Recognition Letters*, Elsevier, vol. 25, 1569–1576.

STOJMENOVIĆ, M., NAYAK, A., AND ZUNIC, J. 2006. Measuring linearity of a finite set of points. In *IEEE Cybernetics and Intelligent Systems*, 1–6.

STONE, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society 36*, 2, 111–147.

STREET, W., WOLBERG, W., AND MANGASARIAN, O. 1993. Nuclear feature extraction for breast tumor diagnosis. In *International Symposium on Electronic Imaging*, vol. 1905, 861–870.

SUN, C., AND WEE, W. 1983. Neighboring gray level dependence matrix for texture classification. *CVGIP 23*, 3 (September), 341–352.

TAO, Y., LAM, E. C., AND TANG, Y. Y. 2001. Features extraction using wavelet and fractal. In *IEEE Transactions on Pattern Recognition Letters*, 271–287.

THIBAULT, G., DEVIC, C., FERTIL, B., SEQUEIRA, J., AND MARI, J.-L. 2007. Indices de formes : de la 2d vers la 3d. application au classement de noyaux de cellules. In *Association Française de l'Informatique Graphique (AFIG)*, 17–24.

THIBAULT, G., DEVIC, C., HORN, J.-F., FERTIL, B., SEQUEIRA, J., AND MARI, J.-L. 2008. Classification of cell nuclei using shape and texture indexes. In *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 25–28.

THIBAULT, G., FERTIL, B., SEQUEIRA, J., AND MARI, J.-L. 2008. Indices de textures : application au classement de noyaux de cellules. In *MajecSTIC*, 65–73.

THIBAULT, G., FERTIL, B., NAVARRO, C., PEREIRA, S., CAU, P., LEVY, N., SEQUEIRA, J., AND MARI, J.-L. 2009. Texture indexes and gray level size zone matrix application to cell nuclei classification. In *Pattern Recognition and Information Processing (PRIP)*, 140–145.

THIEL, E. 1994. *Les distances de Chanfrein en analyse d'images : fondements et applications*. PhD thesis, Université Joseph Fourier, Grenoble 1.

THIRAN, J.-P., AND MACQ, B. 1996. Morphological feature extraction for the classification of digital image of cancerous tissues. *IEEE Transaction on Biological Engineering 43* (October), 1011–1020.

TOUMI, A., HOELTZENER, B., AND KHENCHAF, A. 2006. Classification des images ISAR pour la reconnaissance des cibles. In *XIIIème Rencontres de la Société Francophone de Classification (SFC)*.

TRIER, Ø. D., JAIN, A. K., AND TAXT, T. 1996. Feature extraction methods for character recognition-a survey. In *IEEE Transactions on Pattern Recognition Letters*, IEEE, vol. 29, 641–662.

TUCERYAN, M., AND JAIN, A. K. 1998. Texture analysis. 207–248.

TUFFÉRY, S. 2007. *Data Mining et statistiques décisionnelles*, 2nd ed. Edition Technip, June.

TUSET, V. M., LOZANO, I. J., GONZÁLEZ, J. A., PERTUSA, J. F., AND GARCÍA-DÍA, M. M. 2003. Shape indixes to identify regional differences in otolith morphology of comber. In *Journal of Applied Ichthyology*, vol. 19, 88–93.

UNG, A., RANCHIN, T., WALD, L., WEBER, C., HIRSCH, J., PERRON, G., AND KLEINPETER, J. 2002. Cartographie de la pollution de l'air : une nouvelle approche basée sur la télédétection et les bases de données géographiques. application à la ville de Strasbourg. In *Journées CASSINI 2002 du GDR CASSINI-SIGMA*.

VISA, S., AND RALESCU, A. 2005. Issues in mining imbalanced data sets - a review paper. In *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference*, 67–73.

WEISS, G. M., AND PROVOST, F. 2003. Learning when training data are costly : The effect of class distribution on tree induction. *Artificial Intelligence Research 19*, 315–354.

WIRJADI, O., BREUEL, T. M., FEIDEN, W., AND KIM, Y.-J. 2006. Automated feature selection for the classification of meningioma cell nuclei. *Bildverarbeitung für die Medizin 19*, 76–80.

WONNACOTT, T. H., AND WONNACOTT, R. J. 1998. *Statistique : Economie - Gestion - Sciences - Médecine (avec exercices d'application)*, 4th ed. Economia, March.

WOUWER, G. V. D., SCHEUNDERS, P., AND DYCK, D. V. 1999. Statistical texture characterization from discrete wavelet representations. In *IEEE Transactions on Image Processing*, vol. 8, 592–598.

XU, D., KURANI, A., FURST, J., AND RAICU, D. 2004. Run-length encoding for volumetric texture. In *International Conference on Visualization, Imaging and Image Processing (VIIP)*, 452–458.

YANG, Q., AND WU, X. 2006. 10 challenging problems in data mining research. *Information Technology & Decision Making 5*, 4, 597–604.

YOGESAN, K., JØRGENSEN, T., ALBREGTSEN, F., TVETER, K. J., AND DANIELSEN, H. E. 1996. Entropy based texture analysis of chromatin structure in advanced prostate cancer. *Cytometry 24* (July), 268–276.

ZAHN, C., AND ROSKIES, R. 1971. Fourier descriptors for plane closed curves. In *IEEE Transaction on Computer*, vol. 21, 269–281.

ZHANG, D., AND LU, G. 2001. A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *International Conference on Multimedia and Distance Education*, 1–9.

ZHANG, D., AND LU, G. 2002. Generic fourier descriptor for shape-based image retrieval. In *Signal processing. Image communication*, Elsevier, Amsterdam, PAYS-BAS, vol. 17, 825–848.

ZHANG, D., AND LU, G. 2004. Review of shape representation and description techniques. *Pattern Recognition 37*, 1–19.

ZHANG, J., AND MANI, I. 2003. knn approach to unbalanced data distributions : A case study to involving information extraction. In *The Twentieth International Conference on Machine Learning (ICML)*.

ZIGHED, D.-A., MARCELLIN, S., AND RITSCHARD, G. 2007. Mesure d'entropie asymétrique et consistante. In *Extraction et Gestion des Connaissances (EGC)*, Cépaduès, Toulouse, G. Venturini and M. Noirhomme, Eds., RNTI, EGC, 81–86.

ZUCKER, S., AND TERZOPOULOS, D. 1980. Finding structure in co-occurrence matrices for texture analysis. *CGIP 12*, 3 (March), 286–308.

ZUNIC, J., AND ROSIN, P. L. 2004. A new convexity measure for polygons. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, 923–934.

INDEX

RÉSUMÉ

Dans cette thèse, nous présentons une approche de reconnaissance de forme pour la caractérisation et le classement de noyaux de cellules prélevés chez des patients atteints par la maladie de la progéria. Les noyaux sont marqués par immunofluorescence puis observés au microscope à fluorescence. L'analyse des noyaux doit permettre de diagnostiquer s'ils sont normaux ou pathologiques. L'approche est basée sur une modélisation systématique des éléments de diagnostic (forme, texture, trous et focis) par différentes caractéristiques et algorithmes de classement par apprentissage supervisé.
La première partie s'intéresse au classement des noyaux en fonction de leur forme. Nous effectuons une caractérisation efficace par indices de forme, parmi lesquels quatre nouveaux indices de notre conception. Ces indices permettent de discriminer la forme et ainsi construire un sous-modèle de classement efficace.
En début de deuxième partie, nous proposons une nouvelle méthode de caractérisation statistique de l'homogénéité de la texture. Cette technique est basée sur le dénombrement des régions de même niveau de gris. Les informations extraites sont stockées sous forme matricielle puis caractérisées à l'aide d'indices de texture dont deux nouveaux qui détectent les textures ayant de grandes zones homogènes mais d'intensités différentes. Ces indices et cette méthode se sont révélés pertinents pour améliorer le classement des noyaux.
L'expertise des noyaux est également basée sur la présence d'éléments locaux de textures spécifiques (les trous et les focis), parfois constitués de quelques pixels. Pour les modéliser, nous proposons une succession d'étapes qui forment un procédé générique d'analyse statistique. Celui-ci est basé sur la représentation de la texture par sa carte d'élévation 3D. Les éléments sont extraits, caractérisés par indices de forme 3D, filtrés par classement pour affiner leur pouvoir de caractérisation et enfin dénombrés.
Après avoir modélisé tous les éléments de diagnostic, nous fusionnons les sous-modèles dans un modèle final qui classe de manière fiable et efficace les noyaux de cellules.

Mots-clés : Caractérisation de forme et de texture, indices de forme 2D et 3D, indices de texture, matrice de taille de zone, volume sous la nappe, classement de noyaux de cellules.

ABSTRACT

SHAPE AND TEXTURE INDEXES : FROM 2D TO 3D. APPLICATION TO CELL NUCLEI CLASSIFICATION

This PhD is dedicated to pattern recognition methods applied to characterization and classification of blood cell nuclei from patients suffering from Progeria. Immunodetection through an epifluorescent microscope is used to sort nuclei into pathological and normal groups. The method relies on a systematic modeling of the characteristics (shape, texture, holes, focis) and learning classification algorithms.
First, the shape classification is discussed : the use of shape indexes, four of them being designed especially for this study, allows to discriminate the nuclei and to build an efficient classification submodel.
Then, a new statistical method to evaluate the texture homogeneity is suggested, based on the enumeration of areas displaying a similar gray level. Matrices are used to store the extracted pieces of information on which texture indexes are applied. Two new textures indexes, focusing on large homogeneous areas of far intensity levels, induce a noticeable improvement of the classification results.
Furthermore, the detection of small patterns of only few pixels, representing local and specific elements (holes and focis) for the texture, that are modeled through a several steps process defining a generic statistics analysis method. Using a three dimensional map of the texture, the data are first extracted and sorted thanks to 3D shape indexes, then filtered by classification process in order to improve their characterization power, and finally enumerated.
Eventually, the above elements are merged in a complete model which evaluation proves to be both efficient and reliable in cell nuclei analysis in a diagnosis context.
After the modelling of each diagnoses elements, we merge submodels in a final model which classifiy with reliability and efficiency cell nuclei.

Keywords : Shape and texture characterization, shape indexes 2D and 3D, texture indexes, gray level size zone matrix, volume under elevation map, cell nuclei classification.

Thèse préparée au sein du Laboratoire LSIS – Equipe I&M