



**HAL**  
open science

# Qualité géométrique des entités géographiques surfaciques. Application à l'appariement et définition d'une typologie des écarts géométriques

Atef Bel Hadj Ali

## ► To cite this version:

Atef Bel Hadj Ali. Qualité géométrique des entités géographiques surfaciques. Application à l'appariement et définition d'une typologie des écarts géométriques. Sciences de l'ingénieur [physics]. université Gustave Eiffel; Anciennement Université de Marne La vallée, 2001. Français. NNT: . tel-03244834

**HAL Id: tel-03244834**

**<https://hal.science/tel-03244834>**

Submitted on 1 Jun 2021

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

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



Numéro d'identification:



UNIVERSITE DE MARNE-LA-VALLEE

THESE DE DOCTORAT  
Spécialité: Sciences de l'Information Géographique

**Qualité géométrique des entités géographiques surfaciques  
Application à l'appariement et définition d'une typologie des  
écarts géométriques**

Par :

**Atef BEL HADJ ALI**

Soutenue publiquement le **22 octobre 2001**, devant le jury composé de :

<b>Bernard CERVELLE</b> , Professeur, Université de Marne-La-Vallée	Président
<b>Maurice MILGRAM</b> , Professeur, Université Pierre & Marie Curie, Paris VI	Rapporteur
<b>Mohamed Rached BOUSSEMMA</b> , Professeur, Ecole Nationale des Ingénieurs de Tunis	Rapporteur
<b>Geoffrey EDWARDS</b> , Professeur, Université LAVAL , Québec, Canada	Rapporteur
<b>Robert JEANSOULIN</b> , Chercheur CNRS, Université de Provence, Marseille	Directeur de thèse
<b>Patrice BOURSIER</b> , Professeur, Université de la Rochelle	Examineur
<b>François VAUGLIN</b> , Docteur, Institut Géographique National	Examineur

*à mon père,  
à ma mère,  
à mon épouse  
et à mes deux enfants ....*

*Atef.*

## AVANT-PROPOS

*Ce travail de thèse a été effectué au sein du laboratoire COGIT (Conception Objet et Généralisation d'Information Topographique) de l'Institut Géographique National à Saint-Mandé. Que soient ici remerciés **Jacques Poulain**, directeur technique de l'IGN, **Hervé Le Men**, directeur technique adjoint, **Serge Motet**, chef du service de la recherche, **Sylvie Lamy**, ex-directrice du laboratoire COGIT et **Anne Ruas**, l'actuelle directrice du laboratoire.*

*Mes remerciements s'adressent aussi au gouvernement français et au ministère tunisien de la défense pour le financement de cette thèse. Merci à **Jean-Michel Namur** (CROUS Creteil) qui a géré mon dossier avec efficacité et professionnalisme. je remercie également le Colonel-Major **Mahmoud Mezougui** pour son soutien moral durant ces dernières années.*

*La direction universitaire de cette thèse était assurée par **Robert Jeansoulin**, chercheur CNRS, du Laboratoire d'Informatique de Marseille, dont le rôle a été délicat pendant ces années de thèse du fait de la distance spatiale qui nous séparait, mais qui a su être patient et m'a fait confiance. La direction technique au sein du laboratoire était menée par **François Vauglin**, à qui je suis très reconnaissant pour le temps qu'il a consacré pour le suivi de mes travaux ainsi que pour sa disponibilité et sa participation active à l'élaboration de cette thèse.*

*Je tiens à exprimer ma profonde reconnaissance à **Mohamed Rached Boussema**, **Geoffrey Edwards** et **Maurice Milgram** qui ont accepté la lourde tâche de rapporteurs. Je les remercie pour les remarques et les critiques qu'ils m'ont prodiguées et qui ont permis de perfectionner le présent manuscrit.*

*Mes remerciements vont également à **Bernard Cervelle**, pour m'avoir initialement accueilli au sein de la formation doctorale (pour un DEA en sciences de l'Information Géographique), pour m'avoir aidé à monter mon dossier de thèse, et finalement pour l'honneur qu'il m'a accordé en présidant mon jury de thèse.*

*Un grand merci à **Nicolas Chrisman**, professeur à l'université de Washington à Seattle, pour le temps qu'il m'a consacré pour analyser mes travaux et pour les conseils qu'il m'a prodigués ... qu'il reçoive ici en un peu de mots toute ma gratitude.*

*J'adresse mes remerciements à toutes celles et à tous ceux que j'ai côtoyés durant ces dernières années au COGIT ..... **Sébastien Delattre** (le stagiaire qui a travaillé avec moi et a contribué à l'avancement de ce travail), **JeF** (pour son anglais "shakespearien" qui m'a été d'un grand recours lors de la rédaction de mes articles), **Fred** et **Sylvain** (pour leur lecture minutieuse de ce manuscrit), **Thierry** (mon collègue de bureau, qui m'a supporté depuis que j'étais stagiaire en DEA), **Olivier**, **Jean-Georges**, **Hakima**, **Cécile** (-s: L et D). A **Alain**, pour m'avoir facilité la vie pour toutes sortes de choses, des démarches administratives aux préparations des missions et voyages. De peur d'en oublier, merci à tout le personnel du COGIT.*

*Enfin, un travail de thèse n'est guère un travail individuel. En effet je remercie tous ceux qui ne sont pas nommément cités et qui m'ont aidé de près ou de loin à l'élaboration de cette thèse. Merci aussi pour les âmes dévouées qui ont supporté mon éloignement durant ces dernières années.*

*Et plus particulièrement, par ce document, je tiens à montrer toute ma reconnaissance à ma famille qui a permis l'aboutissement de mes "très" longues années d'étude. A mes parents, à mon épouse et à mes deux enfants.*

## **RESUME**

Ce travail de thèse est consacré à l'étude de la géométrie des entités surfaciques dans les bases de données géographiques, dans un contexte d'évaluation de la qualité de la forme et de la position.

Le rapport est constitué de trois chapitres. Il débute par l'étude des concepts sur la qualité des données géographiques. L'étude s'est orientée vers l'analyse des outils et méthodes qui s'intéressent à la qualité des données géographiques ponctuelles et linéaires. Nous avons passé en revue les tentatives qui les ont utilisés pour les adapter aux entités surfaciques, en montrant leurs limites afin de présenter l'intérêt du sujet étudié.

La représentation vecteur manque de pertinence pour représenter les entités surfaciques. En effet, il est difficile d'étendre l'emploi des outils et méthodes d'évaluation de la qualité géométrique des primitives linéaires utilisant cette représentation pour les primitives surfaciques. Nous avons donc proposé dans le deuxième chapitre un ensemble de représentations dont l'utilisation permet de mieux cerner les caractéristiques géométriques des entités. De surcroît, l'association des métriques appropriées à ces espaces de représentation permet de mieux caractériser les écarts de formes et de position dans le contexte du contrôle de qualité.

Une évaluation de la qualité passe éventuellement par une étape d'appariement dans laquelle, les différentes entités qui correspondent à différentes représentations du même phénomène physique sont identifiées. En effet, une méthode de mise en correspondance des entités surfaciques utilisant l'appariement géométrique est proposée dans le troisième chapitre. Les écarts de forme ont été répartis en trois classes (fort, moyen, faible et Nul - pas d'écart -), et la même classification est adoptée pour les écarts de position. L'utilisation combinée de ces écarts permet d'aboutir à une typologie de 13 configurations pertinentes d'écarts. Chacune des 13 classes d'écarts est caractérisée par un ensemble de règles utilisant les mesures développées dans le cadre de ce travail. La méthode permettant de générer ces règles utilise les techniques de classification non supervisée et les techniques d'apprentissage supervisé.

L'application de ces techniques sur des données réelles est présentée dans ce rapport, ce qui montre leur intérêt pratique pour l'évaluation de la qualité géométrique des entités surfaciques.

# SOMMAIRE

<b>Avant-propos</b>	<b>3</b>
<b>Résumé</b>	<b>4</b>
<b>Sommaire</b>	<b>5</b>
<b>Liste des figures</b>	<b>8</b>
<b>Introduction</b>	<b>10</b>
<b>Chapitre I : Qualité des données géographiques</b>	<b>14</b>
<i>I.1. L'information Géographique Numérique</i>	15
I.1.1. Modélisation de l'univers physique (Spécifications)	16
<i>I.2. Qualité de l'information géographique.</i>	17
I.2.1. Univers nominal	18
I.2.2. Composantes de la qualité	21
I.2.2.1. Actualité	21
I.2.2.2. Généalogie	22
I.2.2.3. Cohérence logique	23
I.2.2.4. Qualité géométrique	23
I.2.2.5. Qualité sémantique et exhaustivité	24
I.2.3. Assurance qualité	25
I.2.4. Contrôle qualité	25
<i>I.3. Qualité des primitives géométriques dans les BDG</i>	28
I.3.1. Indicateurs géométriques	28
I.3.2. Qualité géométrique du ponctuel et du linéaire	30
I.3.2.1. Description des erreurs des primitives ponctuelles	30
I.3.2.2. Description des erreurs des primitives linéaires	31
I.3.3. Modèles statistiques : bruitage et simulation	33
I.3.4. Mesure des écarts entre les primitives géométriques	34
I.3.5. Approches utilisées pour qualifier la géométrie des objets surfaciques.	41
<i>I.4. Appariement des données géographiques</i>	44
I.4.1. Définition de l'appariement	44
I.4.2. A quoi ça sert?	44
I.4.2.1. Mise à jour de base de données géographiques	44
I.4.2.2. Unification des bases de données et création des serveurs multi-échelles	45
I.4.2.3. Contrôle de la qualité	45
I.4.3. Etat de l'art	45
<i>I.5. Synthèse</i>	47
<b>Chapitre II : Représentations de la géométrie des entités surfaciques et Mesures</b>	<b>51</b>
<i>II.1. Introduction</i>	52
II.1.1. Pourquoi de nouvelles représentations?	52
<i>II.2. Critères de choix d'une représentation</i>	54

II.2.1. L'unicité	54
II.2.2. L'inversibilité	55
II.2.3. La stabilité	55
II.2.4. L'invariance	55
II.2.5. L'accès aux propriétés géométriques	55
II.2.6. Efficience et complexité algorithmique	56
<i>II.3. Représentations des objets surfaciques</i>	<i>56</i>
II.3.1. Pour les objets simples -représentation du contour-	57
II.3.1.1. Représentation cartésienne	57
II.3.1.2. Représentation angulaire	58
II.3.1.3. Représentation par les signatures polygonales	59
II.3.1.4. Représentation dans l'espace des fréquences -Descripteurs de Fourier-	61
II.3.2. Pour les objets complexes (représentation par l'intérieur de l'entité surfacique)	65
II.3.2.1. Représentation cartésienne (mode maillé)	65
II.3.2.2. Représentation par les moments mathématiques	66
II.3.2.2.1. Définition générale des moments	66
II.3.2.2.2. Représentation par les moments géométriques	67
II.3.2.2.3. Relation entre moments géométriques et la transformée de Fourier	68
II.3.2.2.4. Implémentation et mode de calcul	69
II.3.2.2.5. Signification physique des moments géométriques	69
II.3.2.2.6. Problèmes liés au pas d'échantillonnage	72
II.3.2.2.7. Moments de Legendre	75
II.3.2.2.8. Recherche de l'ordre optimal des moments de Legendre	77
II.3.2.2.9. Moments de Zernike	79
II.3.2.2.10. Recherche de l'ordre optimal des moments de Zernike	81
II.3.2.2.11. Relations entre moments	82
II.3.3. Synthèse sur la représentation par les moments	83
<i>II.4. Métriques et Distances</i>	<i>84</i>
II.4.1. Rappels mathématiques	85
II.4.1.1. Distance ou métrique	85
II.4.1.2. Indicateur de similarité normalisé	85
II.4.2. Distances associées à l'espace cartésien	86
II.4.2.1. Distance de Hausdorff	86
II.4.2.2. Distance de Fréchet	89
II.4.2.3. Probabilité d'association	91
II.4.2.4. Distance surfacique	92
II.4.3. Distance entre fonctions angulaires	92
II.4.4. Distance entre signatures de polygones	94
II.4.5. Distances associées à l'espace des fréquences	96
II.4.6. Distances associées à l'espace des moments	98
<i>II.5. Tests – Calibrage des mesures</i>	<i>100</i>
II.5.1. Stratégie utilisée (bruitage et simulations)	100
II.5.2. Robustesses des indicateurs face aux perturbations (artefacts)	101
II.5.2.1. Comportement de la fonction angulaire face au bruit	101
II.5.2.2. Comportement de la signature polygonale face aux bruits	103
II.5.2.3. Comportement des descripteurs de Fourier face aux bruits	105
II.5.2.4. Comportement des moments face aux bruits	106
II.5.3. Comportement des indicateurs face aux déformations	108
<i>II.6. Synthèse sur les représentations et les indicateurs</i>	<i>111</i>
II.6.1. Pour les entités simples	111
II.6.2. Pour les entités complexes	113
<i>II.7. Bilan et critiques</i>	<i>114</i>

<b>Chapitre III : Appariement, analyse des mesures et contrôle qualité</b> -----	<b>116</b>
<i>III.1. Introduction et approche</i> -----	117
<i>III.2. appariement de données géographiques</i> -----	118
III.2.1. Appariement géométrique des données surfaciques -----	118
III.2.1.1. Liens d'association -----	119
III.2.1.2. Appariement complet -----	122
III.2.1.2.1. Description statistique de la méthode de croisement de données -----	122
III.2.1.2.2. Etablissement des liens multiples -----	123
III.2.1.2.3. Liens particuliers -----	127
III.2.2. Tests de fiabilité de l'algorithme -----	128
III.2.2.1. Test #1 : appariement entre deux actualités différentes d'une même base de données	129
III.2.2.2. test #2 : appariement des données n'ayant pas les mêmes spécifications -----	131
III.2.2.3. test #3 : appariement des données à pavage "presque" complet et ayant les mêmes spécifications -----	136
III.2.2.4. test #4 : appariement des données à pavage complet avec des spécifications différentes, sources de saisie différentes et actualité différente -----	138
III.2.2.5. Critique des tests -----	142
III.2.3. Présentation du prototype réalisé -----	143
<i>III.3. Analyse des mesures</i> -----	144
III.3.1. Mesures et analyse statistique -----	145
III.3.2. Classification des liens sur la base des mesures -----	149
III.3.2.1. Classification par partitions -----	152
III.3.2.2. Classifications hiérarchiques -----	153
III.3.2.2.1. Classifications hiérarchiques par division (Descendante) -----	153
III.3.2.2.2. Classifications hiérarchiques par agglomérations (Ascendante) -----	154
III.3.3. Analyse des classifications et contrôle qualité -----	157
III.3.3.1. Interprétation des classes -----	159
III.3.4. Règles de classification -----	162
III.3.4.1. Interprétation des règles et contrôle qualité -----	165
III.3.5. Application de la méthode des mesures sur des jeux de données "occupation du sol" -----	167
<i>III.4. Synthèse</i> -----	176
<b>CONCLUSION GENERALE</b> -----	<b>179</b>
<b>ANNEXES</b> -----	<b>186</b>
<b>Annexe A : Calcul des moments géométriques par l'utilisation des contours</b> -----	<b>187</b>
<b>Annexe B : Influence du choix du pas d'échantillonnage –Exemples</b> -----	<b>191</b>
<b>Annexe C : prototype d'appariement</b> -----	<b>192</b>
<b>Annexe D : Module de contrôle de la qualité par l'utilisation des règles de classification</b> -----	<b>193</b>
<b>REFERENCES BIBLIOGRAPHIQUES</b> -----	<b>194</b>



# LISTE DES FIGURES

## Chapitre I

Figure I-1: terrain nominal et qualité interne	18
Figure I-2 : modèles de représentation et de stockage de l'information géographique (exemple de données surfaciques)	19
Figure I-3 : exemple d'une fonction d'adhésion pour la limite "floue" entre deux parcelles "forêt" et "broussailles"	20
Figure I-4 : construction d'un super terrain nominal [De Groeve & Lowell 1998]	21
Figure I-5 : Exemple de modélisation des données de la généalogie [BI 1997]	22
Figure I-6 : Processus d'assurance et de contrôle qualité	26
Figure I-7 : Erreur géométrique des primitives linéaires	32
Figure I-8 : exemples d'utilisation de la bande $\epsilon$	35
Figure I-9 : bande $\epsilon$ et mesure de la précision géométrique des primitives linéaires	35
Figure I-10: Indicateur de [Goodchild & Hunter 1997]	36
Figure I-11 : évaluation de l'écart géométrique par l'intersection des bandes $\epsilon$	37
Figure I-12 : indice de concavité	39
Figure I-13 : test de [Chrisman & Lester 1991]	42
Figure I-14 : insuffisance du contrôle ponctuel pour décrire les écarts entre les polygones	49
Figure I-15 : insuffisance de la distance de Hausdorff pour mesurer les écarts de forme	49

## Chapitre II

Figure II-1 : exemple d'une entité complexe	53
Figure II-2 : fonction angulaire d'un polygone simple	58
Figure II-3 : Signature polygonale d'un polygone représentant une parcelle de bois	59
Figure II-4 : Influence du pas d'échantillonnage sur la construction de la signature du polygone	60
Figure II-5 : Descripteurs de Fourier	63
Figure II-6 : Reconstruction d'un polygone à partir de ses descripteurs de Fourier	64
Figure II-7 : passage du mode vecteur au mode raster	65
Figure II-8 : Moments géométriques et caractérisation d'un polygone	71
Figure II-9 : Invariance de forme sous l'effet des transformations affines	72
Figure II-10 : influence du pas d'échantillonnage sur le calcul des moments	73
Figure II-11 : Différence entre moments exacts et moments calculés	74
Figure II-12 : moyenne et variance mobile d'un échantillon mobile de 10 individus de l'erreur relative au pas d'échantillonnage	75
Figure II-13 : Polynômes de Legendre	76
Figure II-14 : Erreur de reconstruction d'une entité surfacique à partir de ses moments de Legendre	78
Figure II-15 : Reconstruction d'un agrégat de polygones par l'utilisation de ses moments de Legendre	79
Figure II-16 : Estimation de l'angle de rotation relatif entre deux entités surfaciques par l'utilisation des moments de Zernike	81
Figure II-17 : Reconstruction d'un polygone par ses moments de Zernike jusqu'à l'ordre 12	82
Figure II-18 : distance de Hausdorff	86
Figure II-19 : Variation des deux composantes de la distance de Hausdorff en fonction de l'abscisse curviligne	87
Figure II-20 : Calcul de la distance de Hausdorff par l'utilisation des diagrammes de Voronoï	88
Figure II-21 : Distance de Hausdorff entre entités surfaciques	89
Figure II-22 : Distance de Hausdorff et mesure de forme	89
Figure II-23 : Les cinq configurations possibles décrites par Venn [Venn, 1881]	91
Figure II-24 : Détermination du coût de mise en correspondance .... Programmation dynamique	96
Figure II-25 : Différents types de normalisation	99
Figure II-26 : Entités tests	101
Figure II-27 : Fonctions angulaires relatives aux entités (figure II-25(a))	102
Figure II-28 : Fonctions angulaires relatives aux entités (figure II-25(b))	103
Figure II-29 : Signatures polygonales relatives aux polygones de la figure II-25(a)	104
Figure II-30 : Signatures polygonales relatives aux polygones de la figure II-25(b)	104
Figure II-31 : Projection de l'objet O sur les polynômes $R_{pq}$ ; $p = 8$	107
Figure II-32 : Evolution des moyennes des distances vs. l'erreur moyenne quadratique	110

## Chapitre III

Figure III-1 : Approche générale	117
Figure III-2 : Etablissement du graphe d'association	120
Figure III-3 : influence de l'effet de bord sur le choix du seuil de filtrage	121
Figure III-4 : Evaluation des liens d'appariement par l'analyse de leur "Détermination"	123
Figure III-5 : matrice d'association	124
Figure III-6 : Exactitude et complétude du lien multiple 5-à-3 de l'exemple de la Figure III-2	125
Figure III-7 : Distance surfacique en fonction de l'exactitude et de la complétude	126
Figure III-8 : Exemple d'un lien circulaire	127
Figure III-9 : Exactitude vs. complétude des liens d'appariement relatifs au jeu de données #1	130
Figure III-10 : exemples de liens d'appariement douteux	131
Figure III-11 : Spécification de saisie d'un bâtiment dans le Cadastre et dans la BDTopo	131

## Liste des figures

---

<b>Figure III-12</b> : Histogrammes de l'exactitude et de la complétude -jeux de données Cadastre-BDTopo-	132
<b>Figure III-13</b> : Histogrammes de l'exactitude et de la complétude relatifs au jeu de données #3	133
<b>Figure III-14</b> : Nombre de liens invalidés vs. seuil de coupure	133
<b>Figure III-15</b> : Exemples de liens invalidés	134
<b>Figure III-16</b> : Evolution de l'exactitude et de la complétude en la présence d'un biais généralisé	135
<b>Figure III-17</b> : Deux saisies de l'occupation de sol - BDTopo - Bédarioux	136
<b>Figure III-18</b> : Exemple d'un appariement multiple de type 17:5	136
<b>Figure III-19</b> : Histogramme de la distance surfacique et pourcentage cumulé	137
<b>Figure III-20</b> : Exemples de mauvaise interprétation	137
<b>Figure III-21</b> : Extraits des jeux de données utilisés pour le test #4	139
<b>Figure III-22</b> : BD forestière utilisées pour le test #4	140
<b>Figure III-23</b> : Répartition des liens selon le seuil de coupure utilisé	141
<b>Figure III-24</b> : Exemples de liens invalidés	142
<b>Figure III-25</b> : Modèle de stockage des résultats de l'appariement	143
<b>Figure III-26</b> : Méthode proposée pour l'analyse des mesures	145
<b>Figure III-27</b> : Exemple de l'insuffisance de l'utilisation d'une seule mesure pour quantifier les écarts de positions	146
<b>Figure III-28</b> : Représentation des mesures initiales sur les 3 plans principaux	148
<b>Figure III-29</b> : Importance des composantes principales	148
<b>Figure III-30</b> : Echantillon utilisé pour les classifications	150
<b>Figure III-31</b> : Exemple de classification hiérarchique par divisions	155
<b>Figure III-32</b> : Classification par les méthodes hiérarchiques par agglomérations	156
<b>Figure III-33</b> : Arbre de classification des mesures entre les entités simples	158
<b>Figure III-34</b> : Répartition des individus analysés en 6 classes	158
<b>Figure III-35</b> : Classification par partitions	160
<b>Figure III-36</b> : Règles de classification générées par les méthodes d'apprentissage supervisé	164
<b>Figure III-37</b> : Analyse de l'information sémantique	169
<b>Figure III-38</b> : Représentation des mesures initiales sur le premier plan principal	170
<b>Figure III-39</b> : Arbre de classification	171
<b>Figure III-40</b> : Exemple d'une carte de qualité	175
<b>Figure III-42</b> : Fonctionnement global de la méthode (de l'appariement au contrôle qualité)	178

## INTRODUCTION

L'essor de la technologie informatique en terme de moyens et d'algorithmes de traitement a donné un élan inestimable aux systèmes d'information géographique (SIG). Cet élan se manifeste sur plusieurs domaines notamment l'acquisition des données géographiques, leur stockage, leur manipulation et de surcroît leur analyse à des fins de planification de l'espace et d'aide à la décision. De ce fait, un nombre croissant d'outils SIG est mis en place et est devenu inévitable pour l'accomplissement de tâches qui autrefois étaient accomplies manuellement. L'exemple classique est la fabrication des cartes topographiques dont la production est très coûteuse en moyens et en temps. En outre, l'information conservée pour une exploitation ultérieure se réduit aux planches mères, qui constituent des supports analogiques avec les distorsions et dilatations que peut subir leur support physique.

Les SIG ont permis, entre autres, l'acquisition des données et leur conservation dans des bases de données numériques en automatisant les chaînes de production manuelle. En effet, à titre d'exemple, on ne parle plus de la distorsion géométrique "des planches mères" mais plutôt d'autres problèmes liés à l'outil informatique. Les bases de données géographiques (BDG) traitent d'autres notions [Goodchild 1991], telles la précision ou le degré de détails dans les mesures, l'erreur qui représente l'écart entre une mesure stockée et une mesure idéale, la résolution géométrique définissant la plus petite entité représentable ou encore la résolution sémantique qui décrit le niveau de détail dans la précision sémantique des entités. De ce fait, les producteurs ainsi que les utilisateurs d'informations géographiques se trouvent confrontés à de nouvelles problématiques qui n'étaient pas d'actualité du temps où l'information géographique se limitait à la production des cartes papiers. Ces problématiques ont été identifiées et classées selon la nature des difficultés qu'elles soulèvent.

Le problème sur lequel nous portons spécialement notre intérêt concerne la précision géométrique des entités géographiques dans les bases de données numériques. En effet, nous ne pouvons pas affirmer avec certitude que les données saisies ou acquises sont exactes. Cela peut être aisément constaté même pour plusieurs saisies dans des conditions rigoureusement identiques d'un même phénomène physique, du fait qu'aucune saisie ne correspond exactement à une autre.

En plus des problèmes cités précédemment, et bien que les efforts de développement des outils SIG aient suivi un rythme soutenu, il reste beaucoup à faire dans le domaine de l'analyse statistique des données spatiales. Ce constat a été fait en 1992 par [Anselin & Getis 1992, p. 20]:

*[...] Most commercial GIS implementations are rather limited in what they offer in term of statistical tools for the analysis of spatial data [...] There are two aspects of this. First, there is the incorporation of spatial statistics techniques as part of the toolbox provided with a GIS, by adding statistical functions of the menu of GIS capabilities, or by providing an easy link between a GIS and a statistical package. A second, and potentially more interesting aspect is*

*the extend to which statistical and even spatial statistical techniques are appropriate for use with GIS, and the resulting need to develop new "spatial" analysis tools.*

Bien que ce constat ait permis de mettre le doigt sur une des lacunes des SIG actuels (du moins commerciaux), les efforts de développement n'ont pas été à la hauteur des attentes des utilisateurs et notamment les géographes, dont ils ont reformulé les mêmes attentes huit ans plus tard [Rushton 2000].

A cela d'autres problèmes s'ajoutent, liés à la modélisation et à la représentation des primitives géométriques dans les BDG. En effet, depuis la création du premier SIG vecteur, les primitives géométriques sont souvent gérées et analysées en s'articulant exclusivement sur la notion du point, forçant l'utilisateur à traiter les données dans un espace discret. Un cercle, par exemple, est stocké dans une BDG par une suite de segments, bien qu'il puisse être stocké par un couple d'information indiquant son centre et son rayon. Des travaux ont été engagés dans ce sens en tentant de représenter les tronçons de routes par des segments de droites, des arcs de cercles et des cubiques en approchant des arcs de clothoïdes [Affholder 1994]. Cependant ce type d'approche n'est possible que sur des objets dont le modèle géométrique est connu.

Malgré la reconnaissance accrue de la pertinence et de l'utilité de l'information sur la qualité des données géographiques, et bien qu'on puisse la considérer comme une partie intégrante des données elles-mêmes, un autre constat a été récemment fait par [Tveite 1999, p. 28] en pointant notamment le manque de mesures nécessaires pour l'évaluation de la qualité.

*Despite the increased recognition of spatial data quality information, both qualitative and quantitative, as an integral part of spatial data sets for use in GIS, there has not been a matching increase in the elaboration of such information according to the meta data standards recently developed or under development. This can be explained by various causes, such as:*

*\* lack of conclusive, relevant and convenient measures and metrics*

*\* lack of "toolboxes" for data quality assessments and descriptions*

*\* lack of institutional and personal attitudes, competence, priorities and funding within the spatial information society.*

Ces constats ont incité la communauté scientifique à pousser les recherches dans ce sens. Les travaux sur la modélisation des erreurs géométriques dans les BDG ont bénéficié d'une large attention essentiellement pour les primitives linéaires et ponctuelles [Perkel 1956; Chrisman 1982; Abbas 1994; Vauglin 1997; Leung & Yan 1998].

Le contrôle de la qualité géométrique des entités géographiques a souvent été synonyme de contrôle de leurs écarts par rapport à la position nominale. Malgré cela, le contrôle de la forme des entités reste peu étudié lors du contrôle de la qualité géométrique du linéaire. De même, pour les entités surfaciques, nous considérons que le contrôle de leur forme à la même importance que le contrôle de leur position.

Cependant, on rencontre peu de travaux de recherche traitant les entités surfaciques dans les bases de données géographiques, tant au niveau de leur représentation qu'au niveau de la qualification de leur géométrie. En effet, l'étude de la qualité des données surfaciques fait surgir des besoins que nous allons aborder dans les travaux de cette thèse à savoir : un besoin de mesures, un besoin de représentations, un besoin de calibrage et d'analyse et un besoin de méthodes d'appariement des données surfaciques.

Les entités surfaciques ont été souvent traitées par leurs contours en leur appliquant les méthodes déjà testées pour les primitives linéaires sans pour autant tenir compte de leur intérieur. Le travail de recherche mené dans le cadre de cette thèse s'intéresse à la recherche de nouvelles représentations et éventuellement de nouveaux espaces métriques pouvant refléter aux mieux les caractéristiques géométriques des objets surfaciques. Le but est d'identifier le plus facilement possible les différences de forme et de position entre les objets issus de deux bases différentes et qui représentent le même phénomène du monde réel. L'identification de ces différences entre les objets peut servir à une multitude d'applications, notamment pour la mise en œuvre des opérations de mise à jour, des serveurs multi-échelles de données géographiques, des études qualitatives des résultats d'un processus de généralisation et de dérivation ou des opérations de contrôle de qualité.

Le travail de cette thèse se situe dans le cadre de l'étude de la qualité géométrique des primitives surfaciques dans les bases de données géographiques. Elle s'inscrit dans un axe de recherche engagé par le laboratoire COGIT de l'Institut Géographique National depuis 1993 pour l'étude de la qualité des données numériques. En effet, cette thèse étudie les questions liées à la qualité de l'information de localisation et de la forme des données géographiques numériques, notamment dans les opérations de mise en correspondance entre deux bases de données à une échelle égale (contrôle de la qualité) ou à des échelles différentes (mise en œuvre des serveurs multi-échelles par exemple).

L'objectif de cette thèse est donc de mettre au point une méthode de mise en correspondance (ou d'appariement) des bases de données surfaciques qui soit la plus automatique possible, et d'évaluer la qualité de la forme et de la position des entités appariées en utilisant des nouvelles représentations appropriées au traitement de ce type de problème.

Ce rapport est organisé comme suit :

Dans le premier chapitre, nous rappelons les concepts généraux de la qualité des données géographiques numériques, et nous retraçons un état de l'art en synthétisant les travaux de recherche réalisés et les outils développés pour évaluer la qualité de l'information de localisation des données géographiques ponctuelles et linéaires.

Le deuxième chapitre est consacré à la recherche de nouvelles représentations, ainsi qu'aux espaces métriques qui leurs sont associés, dont l'objet est de refléter au

mieux les caractéristiques géométriques des entités surfaciques. Un ensemble de métriques et de mesures sera présenté dans ce chapitre, ainsi que les tests de calibrage permettant de les valider face à différents types de perturbations.

Enfin, dans un troisième et dernier chapitre, nous traitons le problème de l'appariement des données géographiques et nous présentons un algorithme de mise en correspondance entre des données surfaciques. Une démarche globale pour analyser les mesures entre les entités appariées est également présentée avec la mise place d'une typologie d'écart de forme et de position.

# **CHAPITRE I : QUALITE DES DONNEES GEOGRAPHIQUES**

## **I.1. L'INFORMATION GEOGRAPHIQUE NUMERIQUE**

Depuis le développement des outils et techniques informatiques, les domaines de la géographie et de la cartographie ont subi un changement majeur en adoptant de nouvelles procédures d'analyse, de gestion, d'acquisition et d'archivage des données. En cartographie traditionnelle, la carte "papier" était le support de stockage et de visualisation de l'information géographique. Par conséquent, elle est utilisée comme un vecteur pour véhiculer une quantité condensée de l'information, sans pouvoir pour autant la dissocier, ni facilement l'analyser. En effet, l'apparition des systèmes d'information géographique (SIG) a permis de changer plusieurs habitudes, et de redéfinir un nouveau schéma d'utilisation de l'information géographique, mais en contrepartie, de nouveaux problèmes ont pu apparaître, et sont relatifs à l'outil SIG et aux techniques utilisées.

Les données géographiques sont donc stockées dans des bases de données, ce qui facilite l'accessibilité, la manipulation, l'analyse et la mise à jour. Le stockage de l'information géographique dans les bases de données est généralement structuré en plusieurs volets : le stockage de l'information de localisation, le stockage de l'information descriptive (ou factuelle), le stockage de la topologie et le stockage de la description des données, ainsi que les schémas et les modèles de données. Toutes ces informations nécessitent une modélisation de l'objet à saisir dans la base de données afin de standardiser sa représentation en un objet géographique. La modélisation des objets se fait par abstraction de la réalité physique en un modèle facilement interprétable par l'être humain. Représenter un arbre dans une base de données peut être fait en le modélisant sous la forme d'un point dont on stocke les informations relatives à sa localisation (coordonnées) et les informations sémantiques le décrivant (hauteur, âge, espèce, etc.).

Les difficultés liées à l'information descriptive ont souvent trouvé une réponse dans les techniques classiques des bases de données. Des travaux de recherche s'efforcent de résoudre les problèmes persistants. D'autre part, la composante géométrique n'est plus totalement étrangère aux techniques des bases de données classiques depuis l'apparition sur le marché des systèmes de gestion de base de données (SGBD) majeurs permettant la création et la gestion des objets géographiques. Néanmoins, elle suscite toujours l'intérêt de la recherche en essayant d'apporter des solutions relatives à son stockage, à sa gestion dans la base de données et à son interrogation par des systèmes de requêtes dédiés.

Usuellement, les entités du monde physique sont modélisées par trois types de primitives géométriques dans les bases de données : le point, la ligne et le polygone. Le point est défini par les coordonnées de l'entité, la ligne est une suite de points orientée (ou non) et le polygone est une ligne fermée délimitant une partie finie de l'espace. Ces trois primitives sont liées entre-elles par des relations topologiques, ainsi, une ligne possède un point de départ et un point d'arrivée, et elle se trouve à la droite ou à la gauche d'un polygone selon son orientation. Depuis la création des SIG vecteurs, cette modélisation s'est très vite répandue. Cependant, les travaux de recherche sur la qualité



des données ont pu démontrer les limites de cette modélisation lorsqu'elle est utilisée, par exemple, pour représenter les objets naturels dont on ne peut pas définir avec rigueur les limites géographiques. En effet, de nouvelles modélisations ont été développées pour pallier ces lacunes en se basant sur la théorie des "ensembles flous" afin de mieux cerner les incertitudes des trois primitives [Molenaar 1998; Cheng & Molenaar 1999], et ce, en proposant une nouvelle approche pour définir la topologie entre les primitives géométriques en tenant compte de l'incertitude qui entache leur localisation dans l'espace géographique [Shi & Guo 1999]. Bien que la mise en œuvre de cette modélisation floue pour la représentation de la géométrie commence à faire ses premiers pas pour la représentation des pixels en mode *raster* [Fonte & Lodwick 2000], elle reste encore difficile à mettre en œuvre pour les primitives géométriques en mode vecteur. Les techniques utilisant la notion des ensembles flous ne sont pas traitées dans le cadre de cette thèse car le "flou" nécessite le contrôle de qualité pour être instancié, et par conséquent ces techniques interviennent en aval de nos travaux qui portent sur la qualification et la caractérisation de la qualité géométrique.

### **I.1.1. Modélisation de l'univers physique (Spécifications)**

Pour être capable de définir les paramètres de qualité d'un jeu de données géographiques, il est nécessaire de modéliser d'une manière précise le monde réel (appelé aussi univers physique) pour aboutir au jeu de données. Ce processus est décomposé en trois étapes : une première étape, dite d'abstraction, permettant de définir un modèle géographique, une seconde étape d'instanciation du modèle géographique en schémas et une troisième étape de production qui permet d'instancier les objets géographiques en tenant compte des schémas.

La réalité est trop complexe pour être décrite dans le détail. En conséquence, il est nécessaire d'établir des spécifications aidant à l'abstraction de l'univers physique. Par conséquent, les spécifications des données définissent une vue particulière du monde réel. La définition des spécifications d'un produit a été donnée par l'Organisation Internationale de Standardisation (ISO):

*Document qui prescrit les exigences auxquelles le produit ou le service doit se conformer.*

On fait remarquer, à ce niveau, qu'il faut faire la distinction entre les spécifications de contenu et les spécifications de saisie. Les premières décrivent quelle information doit être présente dans la base de données et sous quelle forme. Elles sont formées par une collection exhaustive des règles en vigueur dans la base de données concernée. Les secondes sont utilisées par les opérateurs des chaînes de production pour savoir comment représenter l'information désirée, en termes de techniques et de méthodes.

## I.2. QUALITE DE L'INFORMATION GEOGRAPHIQUE.

L'utilisation des données géographiques numériques s'est développée d'une manière exponentielle, notamment pour la prise de décision dans les domaines de l'aménagement de l'espace géographique ou dans les autres domaines dont les données géographiques constituent la pierre angulaire des systèmes de prise de décision. Ces décisions sont généralement importantes et coûteuses en terme de conséquences sur l'environnement, mais leur importance est aussi accrue du fait que les décideurs ne sont pas généralement des experts en matière de SIG. Ils ne sont pas impliqués dans les processus d'acquisition et d'analyse des données géographiques qu'ils utilisent. En conséquence, la qualité des données géographiques a un effet considérable sur les résultats de l'analyse et de ce fait sur la prise de décision correspondante [Ehlschlaeger 1996; Couget 1997].

D'une manière générale, la qualité est définie par l'ISO comme un:

*Ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites.*

Cette définition évoque la notion de l'adéquation des données (le produit) à l'utilisation *-fitness for use-*. Cette notion a été pour la première fois suggérée par [Juran & al. 1974] et introduite dans la communauté de l'information géographique par [Chrisman 1983]. En effet, la qualité d'un même jeu de données peut varier d'un utilisateur à un autre, et pour un même utilisateur, d'une application à une autre. Ceci oblige à introduire à nouveau paramètre qui consiste en la définition du contexte d'utilisation des données.

Pour les données géographiques, on peut voir la qualité sous deux angles [BI 1997] : une qualité interne et une qualité externe. La qualité interne consiste à estimer les écarts entre un jeu de données et l'univers nominal, et est définie comme suit :

*Ensemble des propriétés et caractéristiques d'un produit ou service qui lui confère l'aptitude à satisfaire aux spécifications de contenu de ce produit ou de ce service*

La qualité externe décrit l'adéquation des spécifications des bases de données géographiques aux besoins des utilisateurs. Elle est souvent abordée en amont de la construction des données.

La complexité des données géographiques et la diversité de l'information qu'elles contiennent rendent trop complexe la définition de leur qualité d'une manière globale. En effet, les travaux de la recherche internationale en la matière ainsi que les différents comités de normalisation ont tenté de décomposer le concept de qualité en sous classes. Ces sous classes, qu'on convient d'appeler composantes de la qualité, varient d'un auteur à un autre. Cependant, on trouve une certaine convergence au niveau de la définition de quelques-unes parmi elles, lesquelles seront présentées en § I.2.2.

### I.2.1. Univers nominal

L'univers nominal est l'image de l'univers physique vu à travers les spécifications de la base de données [BI 1997], en explicitant les modélisations qu'il faut appliquer pour aboutir à une base de données "idéale". Cette base de données idéale est appelée univers nominal ou encore terrain nominal. Le terrain nominal est donc une image virtuelle de la base de données, qui reste souvent une notion abstraite généralement non accessible aux instruments de mesures<sup>1</sup>. Ceci sous-entend la construction d'une base de données qui soit le plus parfaitement possible mise à jour et exhaustive sans aucune incohérence logique et topologique. On se contente souvent de données à plus grande échelle.

Par ailleurs, le terrain nominal trouve tout son intérêt lorsqu'on traite la qualité interne qui peut être reformulée comme l'adéquation entre le terrain nominal et le jeu de données réellement produit (cf. figure I-1).

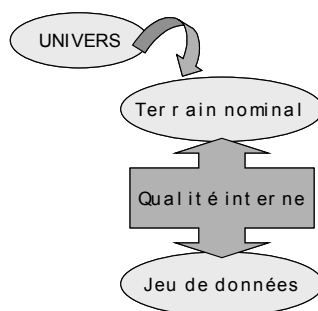


Figure I-1: terrain nominal et qualité interne

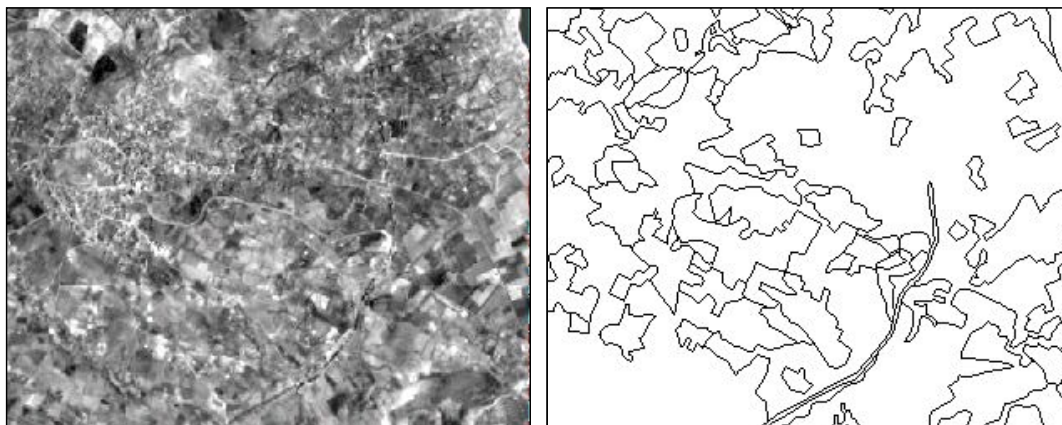
Pour pouvoir évaluer la qualité interne, il est nécessaire de construire une estimation du terrain nominal. Le principe retenu pour la construction d'une estimation du terrain nominal consiste à reproduire le jeu de données avec un soin tel qu'on aura éliminé, le plus possible, les chances d'avoir des écarts entre les données du nouveau jeu et les données nominales. Or, il n'est pas toujours possible d'avoir des données beaucoup plus proches du terrain nominal. En pratique, on restitue le terrain nominal à des données simplement plus précises, voire au pire, de précision comparable.

Par ailleurs, pour des spécifications de contenu données, aussi précises soient-elles, il n'y a pas unicité de terrain nominal en raison de l'ambiguïté et de l'imprécision souvent inévitables des spécifications de contenu (causées essentiellement par des défauts des spécifications ou par l'impossibilité d'y prévoir tous les cas particuliers de l'univers réel). Ce type d'ambiguïté est souvent rencontré lors de la construction des données représentant des phénomènes naturels tels que la délimitation d'une forêt, la représentation des berges d'un lac, l'identification des limites entre deux thèmes d'occupation du sol, etc. Par conséquent, le terrain nominal reste souvent entaché d'incertitudes.

<sup>1</sup> Cette règle n'est toutefois pas absolue. Par exemple, l'univers nominal de la BDCarto® est partiellement inclus dans la carte au 1:50.000, qui a une réalité physique : la carte elle-même [Vauglin 1997]

La majorité des systèmes d'information géographique actuels stocke l'information de localisation d'un objet géographique au moyen de primitives géométriques. Les deux modèles de stockage les plus utilisés utilisent essentiellement la notion de point pour accomplir cette tâche :

- ✓ Le modèle maillé (souvent appelé le modèle *raster*) dans lequel chaque objet géographique est représenté par un ensemble de pixels. Chaque pixel est identifié par ses coordonnées cartésiennes et par l'information radiométrique dont il est porteur.
- ✓ Le modèle vectoriel utilise le point, la ligne et le polygone comme primitives géométriques. Le point est défini par ses coordonnées cartésiennes (x,y,z), la ligne est un ensemble de points organisé sous la forme d'une liste orientée ou non et le polygone est considéré comme une ligne fermée. La figure I-2 représente les deux modes utilisés par les SIG actuels.



(a) Modèle maillé (Raster)

(a) Modèle vectoriel

Figure I-2 : modèles de représentation et de stockage de l'information géographique (exemple de données surfaciques)

L'information géographique est donc représentée dans les bases de données "vecteur" d'une manière qui ne permet pas de prendre en compte la définition naturellement floue des frontières et limites des thèmes. En effet, il faut faire appel à la géométrie floue qui utilise à son tour une logique dite "d'ensembles flous" ou vague [Joos 1994; Alesheikh & al. 1999; Molenaar 1998, Worboys 1998] (*fuzzy logic ou vagueness logic*). La limite entre les thèmes doit donc être définie par une fonction d'appartenance (terme anglais: *membership-function*) comme le montre la figure I-3.

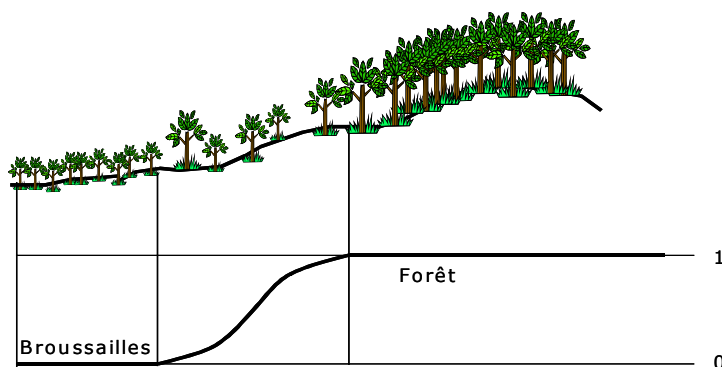


Figure I-3 : exemple d'une fonction d'adhésion pour la limite "floue" entre deux parcelles "forêt" et "broussailles"

Les fonctions d'appartenance sont souvent utilisées dans la manipulation des modèles maillés, mais son adaptation pour les données en mode vecteur est encore difficile à mettre en œuvre [Edwards 1994]. On évoque également le niveau de détail qu'on utilise pour représenter les objets géographiques dans une base de données. Un même objet peut être représenté avec deux densités de points différentes selon l'échelle<sup>2</sup> des données.

Le terrain nominal ne peut donc pas être défini d'une manière unique à cause de la largeur de la zone d'incertitudes de la position des contours des thèmes. [Edwards & Lowell 1996] ont montré que la largeur de la zone d'incertitudes dépend de la différence des textures des zones homogènes ainsi que des informations locales telles que la longueur du contour, le contexte spatial, etc. ce qui donne une marge de liberté au photointerprète lors de la saisie.

En effet, [De Groeve & Lowell 1998] proposent une méthode de construction d'un terrain nominal qu'ils appellent "super terrain nominal" (*Super ground truth*) consistant à croiser plusieurs saisies de la même base de données et de considérer les intersections communes des mêmes thèmes comme base pour définir le terrain nominal (cf. figure I-4). Cette technique a été utilisée pour évaluer la qualité d'un inventaire forestier par l'utilisation des bases à différentes actualités. Les auteurs ont analysé également les parcelles qui ne participent pas à la formation du terrain nominal en imputant les erreurs les plus grossières sur les techniques de classification et proposent l'utilisation des techniques de classification floue pour que le "super terrain nominal" puisse couvrir, le plus possible, l'espace géographique d'une manière correcte.

<sup>2</sup> D'une manière rigoureuse, on ne peut pas simplement parler d'échelle dans les bases de données géographiques numériques. Le terme "échelle" est utilisé par abus pour essayer de donner des équivalences entre les données numériques et les cartes papier. Dire par exemple que l'échelle de la BDCarto® est équivalent à la carte au 1:50.000, parce qu'elle a été saisie à partir des mêmes sources de données qui ont servi à la fabrication de la carte au 1:50.000.

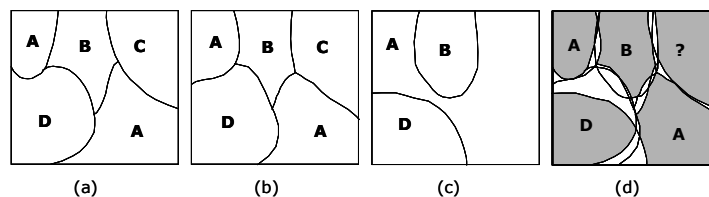


Figure I-4 : construction d'un super terrain nominal [De Groeve & Lowell 1998]  
 (a), (b) et (c) représentent trois interprétations différentes. (d) croisement des différentes interprétations pour la construction du super terrain nominal.

Cette technique paraît facile à mettre en œuvre, si l'on dispose de plusieurs interprétations de la même base de données voire, d'autres bases de données comparables, ce qui n'est pas le cas pour la majeure partie des données géographiques.

L'estimation des écarts d'un jeu de données par rapport au terrain nominal se fait sur cinq composantes [ISO 1994], dont nous donnons les définitions dans la section suivante.

## I.2.2. Composantes de la qualité

Une norme est apparue ces dernières années définissant la qualité de données géographiques sous la forme des cinq composantes suivantes : l'actualité, la généalogie, la cohérence logique, la précision (ou la qualité) géométrique et la précision (ou la qualité) sémantique. Bien que cette liste fasse l'unanimité de presque tous les chercheurs, elle varie encore selon les auteurs qui peuvent distinguer l'exhaustivité de la précision sémantique ou encore la qualité topologique de la cohérence logique ou encore l'ajout de nouvelles composantes telles que la définition, la légitimité, l'accessibilité [Bédard & Vallière 1995]. Nous donnerons, ci-après, la définition des composantes les plus utilisées.

### I.2.2.1. Actualité

Cette composante, appelée aussi qualité temporelle, représente le décalage entre le jeu de données à une date  $T_1$  et le terrain nominal à une date de référence  $T_2$ . L'actualité peut être vue de deux manières différentes. Pour le producteur, l'actualité des données implique une politique de mise à jour décrivant la validité du jeu de données avant la prochaine modification. Pour l'utilisateur, l'actualité des données doit lui permettre de vérifier la validité de ses données au moment de leur utilisation.

L'actualité permet de décrire en quelque sorte « la fraîcheur » des données. L'évaluation de cette composante à une date donnée pourrait se faire par l'utilisateur de données lui-même, en effectuant des mesures rigoureuses de la précision sémantique et de l'exhaustivité à la même date. Or, cette opération (de mesure) très coûteuse est souvent rejetée par l'utilisateur qui veut avoir des informations annexes lui permettant d'évaluer rapidement l'actualité de son jeu de données.

Selon le TC287 [CEN 1999], les indicateurs de la composante d'actualité permettent d'indiquer la date de la dernière mise à jour, la date de validité des données, etc. Cette composante est considérée comme une composante clé, car suite à sa consultation, on peut se faire une idée de la qualité du jeu de données.

### I.2.2.2. Généalogie

Cette composante fournit, d'une manière générale, une expression qualitative décrivant l'histoire des données. Elle fournit également les informations concernant les sources d'acquisition de données, les méthodes utilisées et les opérations appliquées. Il est important de signaler les opérations que les données ont subi, en l'occurrence les transformations de référentiel, les méthodes de passage du mode raster au mode vecteur, les algorithmes de généralisation utilisés si les données résultent d'un autre jeu de données plus détaillé, etc. Cette composante permet également d'indiquer les références complètes des sources (description, date de création, organisme producteur, etc.), des données (zone couverte, classes concernées, objectif de la création, etc.), et des opérations (description, date d'application, équations et paramètres, etc.) [Khagendra & John 1992]

Comme la composante d'actualité, cette composante est aussi considérée comme une information essentielle permettant de juger, rapidement et *a priori*, l'adéquation des données à répondre aux besoins. La quantité de l'information contenue dans cette composante, ainsi que son organisation, sont laissées au choix du producteur de données. A titre d'exemple, la figure I-5 (extraite du [BI 1997]) montre un exemple de modélisation de quelques indicateurs de cette composante.

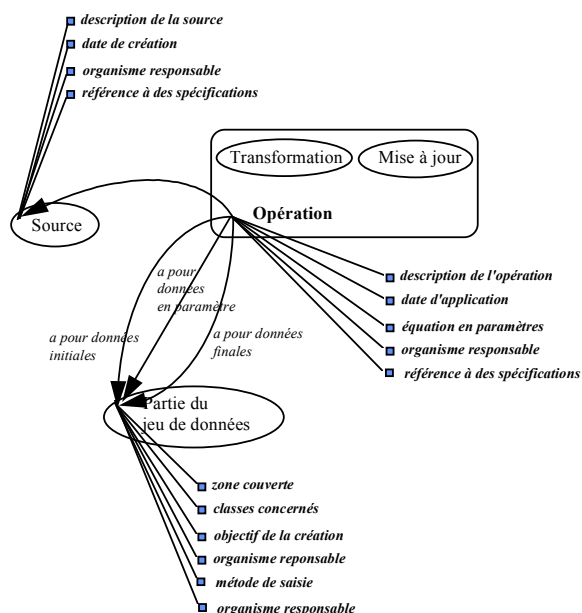


Figure I-5 : Exemple de modélisation des données de la généalogie [BI 1997]

### **I.2.2.3. Cohérence logique**

Cette composante définit le degré de cohérence de données (règles de formatage et contraintes d'intégrité) selon les règles de modélisation et de spécification du jeu de données. La cohérence logique ne dépend pas du terrain nominal, mais uniquement des spécifications, du modèle conceptuel de données et de la connaissance de l'univers physique. Les règles de formatage doivent être impérativement respectées, sinon on n'aura pas accès aux données.

Les contraintes d'intégrité, considérées comme la composante majeure de la cohérence logique, se décomposent en trois règles : [Faiz 1996]

- ✓ Les règles explicites décrites par les spécifications de données [Laurini & Milleret-Raffort 1993], (ex. deux courbes de niveau ne doivent pas se recouper, les arcs d'un polygone doivent former un contour fermé).
- ✓ Les règles explicites liées aux objets. (ex. dans l'occupation du sol de la BDCarto®, il ne doit pas exister de parcelles de moins de 8 hectares pour le bâti).
- ✓ Les règles implicites, telle que par exemple «une rivière coule du haut vers le bas»

A titre d'exemple, pour décrire la cohérence logique on préconise la fourniture des indicateurs suivants [BI 1997]:

- ✓ Description des violations (référence aux règles enfreintes) ;
- ✓ La taille de l'échantillon sur lequel les mesures ont été effectuées ;
- ✓ Le nombre d'irrespect aux règles. Chaque violation des contraintes d'intégrité imposées est comptabilisée. Ainsi on obtient le nombre d'irrespect aux règles.

Cette composante de qualité englobe également la qualité topologique. Cependant, elle se limite à donner le nombre d'occurrences des irrégularités topologiques en les décrivant sans évaluer leur importance ni leur impact sur l'utilisation des jeux de données. La détection des règles enfreintes à la cohérence logique est une tâche aisément automatisable.

### **I.2.2.4. Qualité géométrique**

Cette composante décrit l'écart de géométrie entre l'objet dans le terrain nominal et son homologue dans le jeu de données. Elle se décompose en deux types d'écarts :

- ✓ Précision de position ponctuelle, linéaire et surfacique, qui donne une information sur l'écart probable entre la position planimétrique ou altimétrique des objets dans le terrain nominal et ceux dans le jeu de données à contrôler;



- ✓ Précision de forme qualifiant les éléments géométriques (lignes, polygones et groupes d'objets). La précision de forme donne l'écart de géométrie entre les objets du terrain nominal et ceux du jeu de données.

L'estimation des écarts de position et de forme des primitives géométriques surfaciques sera étudiée en détail dans les chapitres suivants.

#### **1.2.2.5. Qualité sémantique et exhaustivité**

La qualité sémantique décrit la différence entre les valeurs descriptives des éléments du jeu de données et les valeurs de leurs homologues dans le terrain nominal. La qualité sémantique porte essentiellement sur la classification des objets, la codification des attributs et les relations entre les objets [BI 1997]. Il s'agit de voir si les éléments de la base sont correctement identifiés et codés. Souvent, cette composante de la qualité est le plus souvent considérée par les utilisateurs lorsqu'on évoque la question de la qualité des données géographiques.

L'exhaustivité décrit la conformité de la présence ou de l'absence des éléments du jeu de données par rapport au terrain nominal. Elle s'attache aux objets, aux attributs et aux relations. On effectue, donc, des contrôles afin de distinguer, par rapport à la référence, "ce qui manque" et "ce qui est en trop".

L'information sur la qualité sémantique porte alors sur la qualification de l'information factuelle et elle est souvent consignée dans des matrices de confusions entre les classes ou les thèmes contrôlés.

Bien que ces cinq composantes aient été définies séparément, il reste toujours difficile d'isoler les problèmes de qualité dans telle ou telle composante [Chrisman & Lester 1991]. On rencontre le plus souvent des confusions entre la qualité géométrique et la qualité sémantique ou l'exhaustivité [Vauglin 1997]. Cependant, le doute du classement entre les composantes n'est toujours pas résolu d'une manière générique, puisqu'il reste toujours dépendant du choix de l'importance de telle ou telle composante. En effet, il faut mettre en place des mécanismes permettant de faire la distinction entre les composantes afin d'éviter de compter plusieurs fois un problème de qualité.

[Chrisman & Lester 1991], [Norheim 1998] et [Le Men & Jamet 1993] proposent des outils pour automatiser les choix de composantes en identifiant notamment la différence entre la qualité géométrique et la qualité sémantique. Nous reviendrons sur l'explication de ces techniques plus loin.

Les informations contenues dans chacune de ces cinq composantes de qualité sont déterminées dans une optique d'estimation et d'évaluation de la qualité. Le contrôle de qualité est un des principes majeurs liés à la recette d'un jeu de données à l'issue de sa

production. Cependant la qualité d'un jeu de données n'est pas seulement une question soulevée à la fin de sa création mais considérée comme un processus complet qui accompagne le jeu de données depuis sa création. Ce processus est appelé assurance qualité.

### **I.2.3. Assurance qualité**

La définition normative, donnée par l'ISO de l'assurance qualité est la suivante :

*L'ensemble des activités préétablies et systématiques mises en œuvre dans le cadre du système qualité, et démontrées en tant que besoin, pour donner la confiance appropriée en ce qu'une entité satisfera aux exigences pour la qualité.*

D'une manière pratique, il s'agit de l'ensemble des procédures préventives établies à chaque étape de la production afin d'éviter la plupart des fautes grossières, quelles que soient leur origine (défaillances techniques ou humaines). Le but de l'assurance qualité est donc de vérifier que le processus de production mis en place respecte les spécifications.

### **I.2.4. Contrôle qualité**

A l'heure actuelle, il n'existe pas encore de réel consensus sur la définition des indicateurs de chacune des cinq composantes de la qualité interne ni sur la façon de les mesurer. Plusieurs techniques se sont développées ces dernières années pour contrôler les jeux de données tant au niveau du producteur qu'au niveau de l'utilisateur [Chrisman 1982; Hunter & Goodchild 1993; Hunter & al. 1994; Abbas 1994; Aspinall 1996; Leung & Yan 1998; Leung & Yan 1997]. Le contrôle consiste à évaluer les écarts mesurés entre le jeu de données à contrôler et un jeu de données de précision supérieure (voire égale), qu'on appelle par abus "jeu de référence"<sup>3</sup> et dont on suppose qu'il est exempt de tout type d'incertitudes. Par ailleurs, le contrôle de la qualité reste relatif et ne peut pas être fait d'une manière absolue étant donné que le jeu de référence reste entaché d'erreurs. Le contrôle consiste à fixer des seuils empiriques qui servent pour la validation ou le rejet d'un objet en fonction de la mesure qui décrit sa qualité.

La structuration des différentes procédures qui permettent la production des données, l'estimation du terrain nominal et l'évaluation de la qualité par la mesure des écarts entre les données produites et les données nominales est donnée par la figure I-6.

---

<sup>3</sup> Dans la suite de document, nous utilisons les deux termes "jeu de référence" et "terrain nominal" pour désigner la même chose.

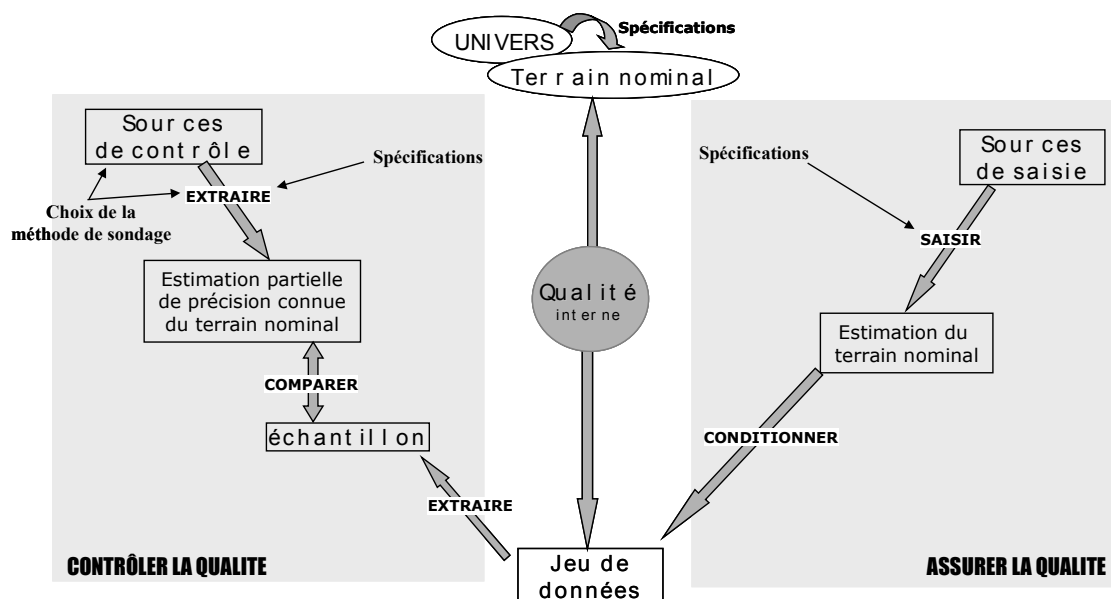


Figure I-6 : Processus d'assurance et de contrôle qualité

La méthode usuelle consiste à reproduire le jeu de données en prenant soin d'éliminer au maximum les possibilités d'avoir des écarts entre le nouveau jeu de données et le terrain nominal. Or, sur le plan pratique, il est difficile et pas toujours possible d'approcher le terrain nominal. En effet, on utilise des données simplement plus précises ou, au pire, de précision comparable, qu'on considère comme des données de "référence". Par conséquent, on parle d'erreur lorsque les données de référence approchent le plus possible<sup>4</sup> le terrain nominal et d'écart lorsque celles-ci sont de précision moindre. Les définitions de l'erreur et de l'écart sont données par les définitions suivantes :

**Erreur :**

L'erreur est définie comme :

*Différence entre une grandeur mesurée et la grandeur nominale correspondante.*

**Ecart :**

L'écart est défini comme :

*Différence entre une grandeur mesurée et une grandeur homologue de référence.*

Les sections précédentes ont montré l'importance de la construction d'une estimation du terrain nominal pour évaluer les indicateurs de qualité ou contrôler la qualité. Cette estimation peut être évitée, en utilisant par exemple les techniques de double saisies [Vauglin 1997]. Or ces techniques, en plus de leur coût, si elles sont

<sup>4</sup> Dans ce cas les données de référence sont appelées quasi-nominales

pratiquées extensivement, ne permettent pas un contrôle satisfaisant. En plus, elles ne permettent pas un contrôle qui prend en compte toutes les composantes de la qualité. En effet, on fait souvent appel aux techniques de sondage. Cependant, le choix de la méthode de sondage a été longuement débattu dans les recherches en la matière, en se posant initialement la question de la pertinence de l'échantillon pour pouvoir généraliser les résultats sur la totalité de la population. Cette question est posée également par [Haining 1990, p. 33] en ces termes :

*The assumption of a "hypothetical universe" of realizations or "superpopulation" is often viewed with considerable skepticism as a model for real spatial data analysis. How might this universe be made susceptible to random sampling and can we be sure the observed map is representative of the universe? What exactly do 5% significance levels or 95% confidence intervals refer to? What if the surface does not satisfy stationarity assumptions?*

Les méthodes de sondage les plus utilisées consistent à tirer au hasard l'emplacement d'un ou plusieurs petits rectangles sur le jeu de données, et à effectuer une saisie exhaustive sur ces régions pour simuler le terrain nominal. La qualité des données sera ensuite évaluée en comparant les données à contrôler avec celles saisies dans les petites régions [Saporitti & al. 1993]. Cette technique est appelée technique des maplets, lorsque la saisie est exhaustive sur les petites zones aléatoires. Cependant, jusqu'à ce jour, la question sur la validité de l'hypothèse de représentativité des maplets reste posée et non résolue.

Les problèmes restent toujours soulevés si l'on se place dans le cadre d'un contrôle qualité en estimant les erreurs entre le jeu à contrôler et l'estimation du terrain nominal. Par ailleurs, une comparaison d'un jeu de données avec un autre jeu de précision supérieure ou de même précision vise à estimer les écarts (sans parler d'erreur) entre les deux jeux. Dans ce cas, on peut procéder à un contrôle exhaustif et mesurer tous les indicateurs de chacune des composantes de qualité.

Les travaux menés dans le cadre de cette thèse se placent dans un contexte d'estimation des écarts entre deux jeux de données surfaciques.

La majeure partie des travaux antérieurs s'est focalisée sur le contrôle des primitives ponctuelles et linéaires, dont nous passerons en revue les majeurs développements dans le §I.3. Le contrôle de la qualité géométrique des primitives ponctuelles et linéaires a mis l'accent d'une manière intensive sur le contrôle de la position des objets dans l'espace géographique en développant des méthodes pouvant accomplir la tâche automatiquement. Cependant, le contrôle de la forme de l'objet est laissé au seul avis "subjectif" de l'opérateur de contrôle.

Les primitives surfaciques ont été contrôlées en les traitant à travers leurs contours, en ramenant la problématique à un simple contrôle linéaire ou bien en réduisant le problème à un simple contrôle ponctuel en mesurant les écarts entre les points de leurs contours ou de leurs centres de masse. En effet, le contrôle de la qualité géométrique des primitives surfaciques sera le problème central de cette thèse auquel nous essayons d'apporter une solution.

### **I.3. QUALITE DES PRIMITIVES GEOMETRIQUES DANS LES BDG**

Nous définissons en détail dans cette section la qualité géométrique. La définition de la composante de la qualité géométrique trace les grandes lignes de cette composante sans donner des détails sur son contenu. Cette composante est trop complexe pour être estimée d'une manière globale. En effet, un certain nombre d'indicateurs doivent être définis pour subdiviser cette composante afin de la rendre quantifiable.

Nous signalons, à ce niveau, que la qualité géométrique de l'information géographique découle d'une agrégation et d'une mise en forme de la qualité des primitives géométriques. En effet, nous désignons par "précision absolue", la précision de l'information géographique, et par "précision relative" la précision des primitives géométriques. Le contrôle linéaire montre un bon exemple de synthèse réussie entre la précision des primitives géométriques et la précision de l'information géographique. Cependant, l'usage de ces deux types de précisions ne rend pas compte de la totalité de la qualité géométrique de l'information géographique. En effet, la précision relative est calculée d'une manière isolée au niveau de chacune des primitives géométriques sans tenir compte de leurs positions dans l'espace les unes par rapport aux autres.

Bien que la qualité géométrique ne soit qu'une composante de la qualité parmi d'autres, elle est souvent confondue dans le langage courant (essentiellement du côté des utilisateurs) avec le terme général de la qualité. Cette composante est une des plus importantes, puisque le problème central dans les travaux sur la qualité de l'information géographique numérique est dû à la difficulté, voire l'impossibilité, d'obtenir des coordonnées parfaites des données acquises.

La position et la forme des objets géographiques dépendent fortement des sources de saisies. Un jeu de données acquis avec soin et respect des spécifications à partir d'une source très précise peut être considéré comme ayant des objets correctement placés dans l'espace. Cependant, on se rend compte, par l'utilisation d'une mesure, que les coordonnées diffèrent légèrement et inévitablement de la position nominale.

Les données sont donc inévitablement entachées d'erreur. Ces erreurs peuvent provenir de plusieurs sources [De Jong 1990] : erreurs relatives aux sources de données, des erreurs de saisie, des erreurs dues aux effets de généralisation, des erreurs générées lors du traitement des données et des erreurs dues aux transformations géométriques (correction géométrique des clichés, passage d'une projection à une autre, etc.).

#### **I.3.1. Indicateurs géométriques**

Comme nous l'avons indiqué plus haut, nous ne pouvons pas déduire directement, de la définition de la composante géométrique, les mesures qui doivent être appliquées pour estimer cette composante. En effet, aucune méthode de calcul des paramètres n'est clairement explicitée par les organisations de normalisation, et il n'existe pas de consensus sur la façon de leur définition [OGC 1999]. Par conséquent les mesures dans

chacun des indicateurs seront librement définies, notamment en fonction des besoins des applications : les indicateurs sont donc dépendants de l'utilisation prévue des données.

A titre d'exemple, nous donnons ici les indicateurs de précision géométrique retenus par l'institut géographique national [BI 1997]. Ce document préconise de séparer la précision géométrique en fonction des primitives géométriques, c'est à dire en précision ponctuelle, linéaire et surfacique. Nous rappelons, dans ce qui suit, la règle courante en matière de mesure d'erreurs.

On définit alors les indicateurs suivants pour la précision ponctuelle. On note que les définitions qui suivent sont données avec les erreurs, qu'on remplace en pratique par les écarts afin d'obtenir une estimation.

- ✓ la moyenne des erreurs

*Moyenne arithmétique des erreurs sur les coordonnées des points, qui donne une estimation du biais.*

- ✓ Biais statistique:

*C'est l'écart entre l'espérance d'une mesure (ou d'une estimation) d'une grandeur et la valeur nominale de cette grandeur.*

Cette définition sous-entend que la valeur nominale est un paramètre d'une loi statistique qu'on estime sur l'échantillon de référence. Plusieurs travaux de recherche ont tenté de formaliser cette loi de probabilité qui régit le comportement des erreurs dans les bases de données [Shi 1994; Vauglin 1997; Tong & al. 2000]. Cette remarque est aussi valable pour les définitions qui suivent.

Pour l'estimation du biais statistique, on évalue la distance entre les objets de l'échantillon à contrôler et les objets homologues du lot de contrôle.

- ✓ La grille de biais régionalisé
- ✓ La grille aux nœuds de laquelle un biais est estimé par la moyenne des erreurs sur les coordonnées des points les plus proches du nœud considéré.
- ✓ L'exactitude
- ✓ La moyenne quadratique des erreurs (EMQ) sur les coordonnées des points
- ✓ La taille de l'échantillon utilisé pour évaluer les indicateurs
- ✓ Le taux de rejet des écarts aberrants : le rapport du nombre d'objets non pris en compte pour l'évaluation des indicateurs à la taille de l'échantillon. Les éléments rejetés de l'échantillon sont ceux qui présentent des écarts aberrants par rapport au terrain nominal. Il s'agit généralement de fautes.

Pour la précision linéaire, on peut conserver les mêmes indicateurs, en utilisant d'autres mesures pour évaluer les écarts. Nous présentons en détail (cf. I.3.2) les travaux antérieurs sur l'évaluation des écarts ponctuels et linéaires.

Il existe par ailleurs d'autres définitions pour des termes analogues, qui servent à estimer la composante de précision géométrique, comme par exemple:

### **Exactitude:**

*L'exactitude mesure les fluctuations des valeurs d'une série de mesures autour de la valeur nominale.*

### **Précision:**

*La précision mesure les fluctuations d'une série de mesures autour de son espérance.*

Ces deux indicateurs prêtent souvent à confusion. Cependant, on fait remarquer que l'exactitude englobe la précision et non pas l'inverse.

A l'heure actuelle, les données surfaciques sont traitées comme des données linéaires en les traitant à travers leurs contours. Un constat, dans ce sens, a été fait par Vauglin [Vauglin 1997, p. 30]:

*Pour l'instant, il n'existe pas de technique réellement surfacique pour évaluer la qualité géométrique des objets surfaciques. Les documents [CNIG 1993; BI 1997] préconisent par ailleurs l'utilisation d'un indicateur "précision de forme" pour qualifier les éléments géométriques construits.*

Ce constat définit parfaitement le contexte des travaux de cette thèse dans laquelle nous essayons d'apporter des solutions aux contrôles de la précision des entités surfaciques, tant au niveau de la position qu'au niveau de la forme.

## **I.3.2. Qualité géométrique du ponctuel et du linéaire**

Les primitives géométriques sont entachées d'erreur, ce qui implique que leur localisation est inévitablement incertaine. La recherche sur la qualité des données géographiques s'est intensivement orientée ces dernières années vers l'étude du comportement statistique des écarts de position en fonction des différentes sources d'erreur.

### **I.3.2.1. Description des erreurs des primitives ponctuelles**

La recherche sur la qualité des primitives géométriques a commencé par le traitement de la plus simple d'entre-elles : le point. La répartition statistique des mesures d'écarts des points a été initialement considérée ou plutôt approchée par une loi uniforme. Ce postulat a été avancé par [Perkel 1954] en proposant le modèle de la bande  $\epsilon$ . La bande  $\epsilon$  propose d'ajouter une valeur de tolérance pour chaque point. Cela signifie que la localisation de l'objet peut varier, au plus, d'une quantité  $\epsilon$ . En d'autres termes, on peut considérer que tout point saisi peut se trouver n'importe où, d'une manière équiprobable, à l'intérieur d'un disque de rayon  $\epsilon$  centré sur ses coordonnées nominales.

[Bolstad & al. 1990] et [Hottier 1997] ont entrepris, plus tard, des études expérimentales consistant à saisir plusieurs pointés et à mesurer les écarts en x et en y. Les deux études s'accordent sur le constat que les écarts ponctuels ont un comportement gaussien. Les erreurs de pointé suivent donc une loi normale. Les résultats de ces deux études s'accordent également sur le fait qu'aucune corrélation significative n'est observée entre les écarts en x et en y, donc l'indépendance entre les deux variables des écarts en x et en y est assurée. D'autres tests ont été engagés par [Vauglin 1997] sur un échantillon de 546 carrefours de Géoroute® et de la BDTopo®; ces tests ont abouti à la même conclusion en modélisant statistiquement les écarts ponctuels par quatre paramètres, en l'occurrence les deux espérances (biais) et les deux écarts types (précision) des écarts observés selon l'axe des abscisses et l'axe des ordonnées.

On signale que la distance utilisée pour mesurer les écarts ponctuels est la distance euclidienne.

### **I.3.2.2. Description des erreurs des primitives linéaires**

L'erreur de restitution commise lors des pointés est liée à l'erreur de position des primitives ponctuelles. Néanmoins, il est très difficile de propager le long de la ligne l'erreur commise sur les coordonnées des points la composant.

Les travaux entrepris par [Amrhein & Griffith 1991] arrivent à la conclusion que, pour les primitives linéaires, l'écart de position est constitué d'un mélange entre les écarts de restitution et les écarts de généralisation. Les écarts de restitution sont essentiellement dus aux écarts de pointés, comme pour les primitives ponctuelles. Les écarts de généralisation sont dus aux écarts introduits lors de la construction de la ligne elle-même en interpolant entre les points. Nous convenons, à ce niveau, de parler d'écarts au lieu d'erreurs du fait qu'il est possible d'avoir des écarts entre les primitives géométriques sans qu'il y ait d'erreurs.

D'une manière générale, si le nombre de points utilisés pour la construction de la ligne augmente, l'effet de l'écart de restitution diminue (cet effet est illustré d'une manière schématique par la figure I-7). Cette constatation est également faite par [Abbas 1994] qui établit une relation entre l'écart-type de loi régissant les écarts et le nombre de sommets de la ligne.



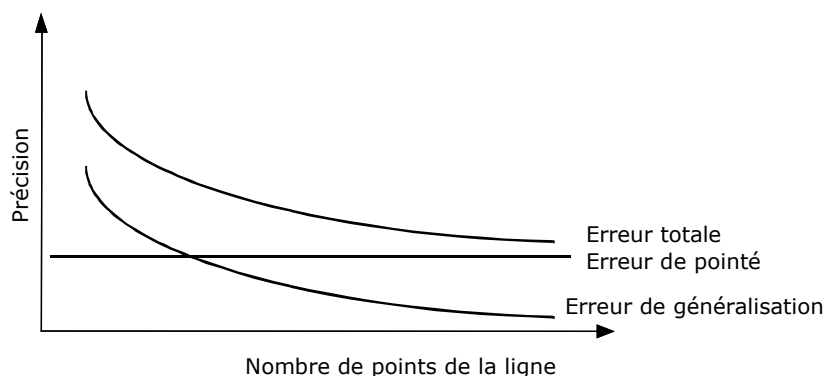


Figure I-7 : Erreur géométrique des primitives linéaires

Bien que ces travaux aient permis de déterminer les deux sources principales d'erreur, ils ont été critiqués à cause de la difficulté de dissocier ces deux types d'erreur. [Veregin 1994, p. 8] en ces termes :

*This model is not easily applied, as there is no simple way to differentiate between the effects of digitizing and generalization error. This is because no clear correspondance exists between the point locations encoded in the database and the locations defining the corresponding real-world entity.*

Une année plus tard, une réponse a été donnée à cette question à travers les travaux de [Vauglin 1997] qui propose un modèle statistique mélangeant une gaussienne et une loi de Laplace. La gaussienne est utilisée pour modéliser les faibles écarts correspondant aux erreurs de pointés. La loi de Laplace<sup>5</sup> est employée pour décrire les grands écarts qui correspondent à ce qu'on a appelé précédemment les erreurs dues aux effets de généralisation. Des tests expérimentaux ont permis de valider l'ajustement de la répartition des écarts mesurés à ce mélange de lois par le test de Kolmogorov-Smirnov au seuil de signification de 1%, ce qui permet d'accepter l'ajustement de la répartition des écarts mesurés par rapport au modèle théorique.

Les travaux précédemment cités ont permis de mieux comprendre le comportement statistique des erreurs des primitives géométriques ponctuelles et linéaires. Pour les primitives surfaciques, il est important de noter que leur incertitude n'est pas restreinte à l'incertitude de leurs contours, puisque leurs intérieurs sont également incertains. L'intérieur d'un polygone est plus homogène que son contour, ce qui rend l'incertitude plus corrélée spatialement pour le contour que pour l'intérieur [Zhou & Lee 1994]. Le modèle d'erreur du linéaire peut ainsi être transposé pour les primitives surfaciques (à travers leurs contours) [Alesheikh & al. 1999]. Les travaux menés dans le cadre de cette thèse ne sont pas orientés vers la recherche d'une modélisation des erreurs des polygones, mais plutôt vers la définition de nouvelles mesures d'écarts géométriques qui prennent en compte les spécificités des données surfaciques. Par ailleurs, les modèles statistiques des écarts sont utilisés pour simuler les bruits afin de tester la robustesse des représentations et des mesures.

<sup>5</sup> Également appelée loi exponentielle symétrique.

### I.3.3. Modèles statistiques : bruitage et simulation

Cette section présente l'utilisation des modèles statistiques pour simuler les types de bruit pouvant affecter les données géographiques. Les techniques de simulation sont souvent utilisées car on n'arrive pas à résoudre d'une manière analytique les problèmes mathématiques courants, que ce soit en science fondamentale, en ingénierie, en économie, en sociologie ou en théorie de la décision. En effet, il est devenu plus efficace de simuler numériquement le comportement d'un système complexe que de l'observer expérimentalement. Il existe deux méthodes de simulation: la méthode dite de "*Monte Carlo*"<sup>6</sup> et la méthode par le calcul différentiel. La méthode la plus utilisée est la méthode dite de "*Monte Carlo*" qui est parfois étendue et formalisée en analyse de sensibilité.

Plusieurs travaux sur la simulation de bruit ont été développés tant dans le domaine du traitement des données "raster" [Goodchild & Cova 1995] que dans le domaine du traitement des données "vecteur" [Fouqué 1999; Bonin 2000; Kiiveri 1997] dans le but de tester le comportement des indicateurs de qualité face aux différents bruits. Ces simulations utilisent des modèles d'erreur qui peuvent être considérés comme des processus stochastiques capables d'ajouter des bruits réels aux jeux de données. Les premiers travaux de simulations ont utilisé un modèle gaussien d'erreur, sans tenir compte de l'auto-corrélation spatiale entre les erreurs des points consécutifs. Bien que ces modèles restent toujours des processus stochastiques, ils ont été mis en cause par [Hunter & Goodchild 1995], puisque le résultat renvoyé reflète partiellement la réalité.

[Kiiveri 1997] présente une méthode de bruitage des primitives géométriques dans les bases de données utilisant le calcul différentiel, en ajoutant un vecteur de déplacement aux coordonnées des points à perturber, et en tenant compte des corrélations qui existent entre les points des primitives à bruite. Le modèle utilisé pour bruite les données est gaussien et utilise deux paramètres, en l'occurrence les deux écarts types de l'erreur selon les deux axes des coordonnées.

Les récents travaux réalisés par [Fouqué 1999] et ceux sur l'utilisation du modèle dit "Gaussienne Exponentielle Symétrique" (GES) [Vauglin 1997] représentent une avancée en matière de simulation des erreurs étant donné que les résultats obtenus s'approchent le plus possible des saisies réelles. Par ailleurs, les résultats obtenus par la simulation révèlent, en pratique, une méconnaissance de la corrélation des écarts, du fait qu'on utilise les paramètres de la précision absolue (§ I.2) pour initialiser le mécanisme de simulation.

Nous présentons succinctement cette méthode de bruitage, puisque nous allons l'utiliser dans le chapitre II afin de tester le comportement des différentes représentations et métriques développées en la présence d'erreurs contrôlées introduites par simulation et suivant une loi GES.

Une loi GES de paramètres  $(\alpha, \mu, \sigma, \lambda)$  est la combinaison convexe d'une loi gaussienne de paramètres  $(\mu, \sigma)$  et d'une loi exponentielle symétrique de paramètres  $(\mu,$

<sup>6</sup> La méthode doit son nom à l'utilisation de nombre au hasard qui évoquent le célèbre casino.

$\lambda$ ) et  $\alpha$  s'interprète comme la mesure mélangeante. La fonction de répartition  $F_{GES}$  peut donc s'écrire :

$$F_{GES} = \alpha F_G + (1 - \alpha) F_{ES} \quad [I-1]$$

Avec, pour tout  $x$  réel,

$$F_G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ et } F_{ES}(x) = \frac{\lambda}{2} \exp(-\lambda|x-\mu|) \quad [I-2]$$

sont respectivement les fonctions de répartition d'une loi gaussienne et d'une loi exponentielle symétrique.

En effet, pour simuler les erreurs utilisant ce modèle, on considère la densité qui modélise la probabilité au point  $M$ , d'avoir un écart  $r$  de la façon suivante:

$$P_M(r) = \frac{\alpha}{\sigma\sqrt{2\pi}} e^{-\frac{r^2}{2\sigma^2}} + (1 - \alpha) \frac{\lambda}{2} e^{-\lambda|r|} \quad [I-3]$$

L'utilisation de cette loi, pour simplement bruitez les sommets des segments des primitives géométriques, ne faisait pas intervenir la corrélation des mesures. Or, le caractère localisé de l'information induit des corrélations non négligeables.

La résolution de ce problème s'est fait par l'utilisation des variogrammes d'écarts afin d'intégrer les corrélations du modèle utilisé pour la simulation des écarts tout en respectant la dépendance spatiale des erreurs. On ne s'est pas contenté de bruitez les sommets des primitives géométriques, mais également d'autres points auxiliaires ajoutés artificiellement sur les primitives. La simulation de l'erreur sur ces points intermédiaires se fait par les techniques de krigeage<sup>7</sup> [Matheron 1963]. Pour en savoir plus sur la méthode, se reporter à [Fouqué 1999].

Les tests de simulation effectués par cette méthode montrent que le bruit généré approche au mieux la réalité. Cependant, une question reste toujours soulevée sur le nombre de points auxiliaires à ajouter à la primitive avant bruitage. En effet, on pourrait penser à échantillonner la primitive selon un pas régulier et effectuer les simulations sur les points intermédiaires générés. Cependant, la méthode renvoie un résultat erroné<sup>8</sup> si les points intermédiaires sont fortement bruités (la primitive bruitée aura une allure en dents de scie).

La section suivante présente les mesures les plus utilisées pour décrire les écarts de géométrie des primitives linéaires ainsi, que les essais de leur transposition pour décrire les écarts géométriques des entités surfaciques.

### **I.3.4. Mesure des écarts entre les primitives géométriques**

Plusieurs techniques pour mesurer la précision géométrique ont été développées, notamment pour mesurer les écarts des primitives géométriques linéaires. Ces

<sup>7</sup> Le krigeage est une méthode d'interprétation linéaire sous contrainte de minimiser la variance des estimateurs.

<sup>8</sup> Le résultat de bruitage n'approche pas les distorsions qui peuvent être causées par un opérateur humain et par les outils de restitution lors de la saisie des données géographiques.

techniques peuvent être classées en deux groupes majeurs : soit par l'utilisation des techniques de bande  $\epsilon$ , soit par l'utilisation des applications de superposition des jeux de données. Pour cela, plusieurs métriques et distances ont été développées pour cerner au mieux les écarts géométriques.

Les techniques de la bande  $\epsilon$  ont été initialement utilisées pour supprimer les effets indésirables lors de la saisie des données géographiques. La valeur  $\epsilon$  agit comme une valeur de tolérance servant à éviter les problèmes "d'overshoot" ou "d'undershoot", ou pour regrouper les nœuds (cf. figure I-8). Ces problèmes sont largement évoqués dans les travaux de [Milenkovic 1989; Pullar 1991; Harvey 1994; Harvey & Vauglin 1997]

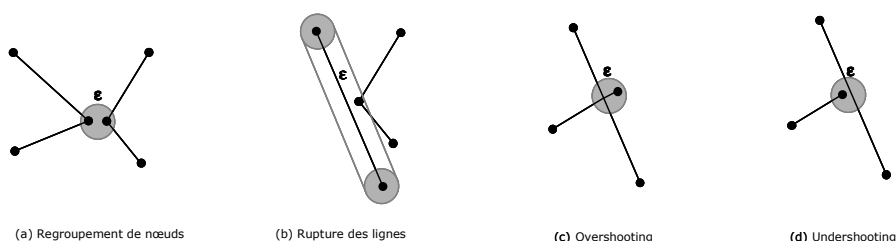


Figure I-8 : exemples d'utilisation de la bande  $\epsilon$

Bien que la bande  $\epsilon$  soit indûment considérée comme un modèle probabiliste traduisant le réalisme des écarts géométriques, elle est souvent utilisée comme outil d'estimation des écarts.

[Goodchild & Hunter 1997; Hunter & al. 1994 ] ont développé un outil utilisant les techniques de la bande  $\epsilon$  pour mesurer les incertitudes des écarts géométriques des primitives linéaires. Cet outil consiste à mesurer le taux d'inclusion de la ligne à contrôler dans la bande  $\epsilon$  de la ligne référence (cf. figure I-9). Le taux d'inclusion est mesuré à différentes valeurs de  $\epsilon$ .

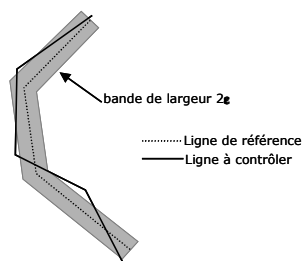


Figure I-9 : bande  $\epsilon$  et mesure de la précision géométrique des primitives linéaires

L'ensemble des mesures de taux d'inclusion constitue un indicateur, qui se présente sous la forme d'une fonction du taux de l'inclusion en fonction de la valeur de  $\epsilon$ . Les auteurs ont montré, par des tests empiriques, la plausibilité de cet indicateur avec un modèle gaussien. Cette démonstration consiste à déterminer la valeur  $\epsilon$  pour une probabilité  $p$  donnée (le pourcentage d'inclusion) en admettant que  $p$  suit une distribution gaussienne. Si la démonstration des auteurs a été faite d'une manière empirique sur un exemple, il nous a été difficile de la reproduire sur d'autres données.

L'exemple de la figure I-10 illustre un essai d'utilisation de cet indicateur sur un couple de polygones extrait de deux saisies d'une couche d'occupation du sol de la BDTopo®.

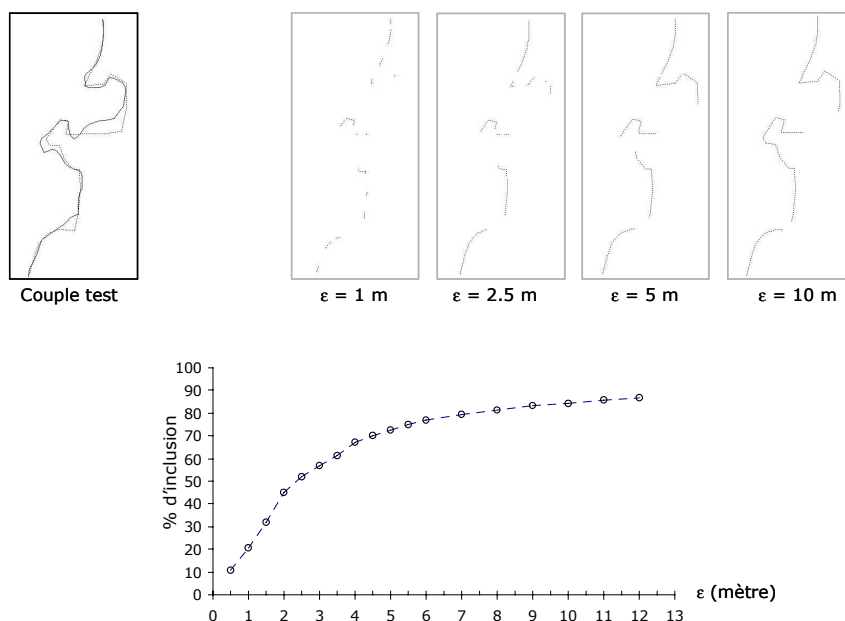


Figure I-10: Indicateur de [Goodchild & Hunter 1997]

La précision géométrique de la polygone "à contrôler" peut être déterminée en prenant la valeur de  $\epsilon$  pour une inclusion entre 90 et 95%. Le couple de lignes (représentant une portion de contours de deux polygones) a été extrait de la BDTopo® dont l'erreur moyenne quadratique est de l'ordre de 1.5 mètres. La valeur de  $\epsilon$  à 90% d'inclusion avoisine les 12 mètres, tandis que pour une inclusion totale (100%) la valeur de  $\epsilon$  est de 24.5 mètres. Par ailleurs, pour une précision à 1.5 mètres, seule 30% de la ligne à contrôler est incluse dans la bande  $\epsilon$ . Notons que la valeur de  $\epsilon$ , à une inclusion à 100%, correspond à la valeur de la distance d'une composante de la distance de Hausdorff. Cet indicateur est très sensible à la présence de forte distorsion locale et renvoie une indication dissymétrique sur l'écart géométrique en fonction du jeu de données qui est utilisé comme référence.

L'utilisation de cet indicateur pour les données surfaciques pourrait se faire en les traitant à travers leurs contours, ce qui exclut le traitement des polygones complexes.

Cet indicateur peut être utilisé comme un filtre de coupure des détails dont on considère qu'ils sont "hors tolérance" avant de procéder à un contrôle de qualité des primitives linéaires par l'utilisation de la distance de Hausdorff (cf. II.4.2.1.). Les portions de lignes, qui ne sont pas incluses dans la bande  $\epsilon$ , seront donc imputées au taux de rejet [Abbas 1994].

Cette méthode est également utilisée pour l'appariement géométrique des primitives linéaires en fixant des seuils sur la valeur de la bande  $\epsilon$ . Par exemple, si la valeur de  $\epsilon$  à 100% d'inclusion est inférieure au seuil fixé, les deux lignes sont considérées comme homologues, sinon elles ne sont pas considérées comme appariables.

D'autres études comparables à celle de [Goodchild & Hunter 1997] engagées par [Tveite 1999] utilisent la même philosophie. Par opposition à l'indicateur de [Goodchild & Hunter 1997] qui utilise l'intersection de la ligne "à contrôler" avec la bande  $\epsilon$  de la ligne de référence, [Tveite 1999] utilise l'intersection entre les deux bandes  $\epsilon$  des deux lignes (cf. figure I-11).

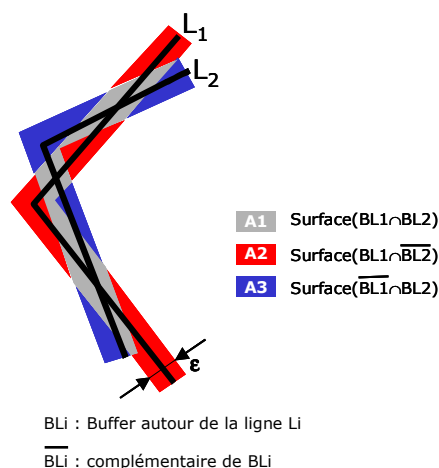


Figure I-11 : évaluation de l'écart géométrique par l'intersection des bandes  $\epsilon$

Pour mesurer les écarts de la géométrie, [Tveite 1999] a défini deux indicateurs : un pour mesurer le déplacement moyen, l'autre pour donner une indication sur le biais.

Le premier indicateur donne le déplacement moyen de la ligne à contrôler par rapport à la ligne de référence. Il est donné par:

$$D_m = 2\epsilon \frac{\text{Surface}(\overline{BL_1} \cap BL_2)}{\text{Surface}(BL_1)} \quad [I-4]$$

Le deuxième indicateur donne une indication sur le biais. L'analyse de cet indicateur, en faisant croître la valeur de la bande  $\epsilon$ , pourrait donner une large indication sur le biais spatial. Cet indice est fortement inspiré de la méthode de comptage des polygones parasites [McAlpine & Cook 1971], et est donné par:

$$I_b = \frac{\#\text{polygones}(\overline{BL_1} \cap BL_2)}{\text{longueur}(L_1)} \quad [I-5]$$

#polygones est donné par le dénombrement de polygones générés suite à l'intersection des bandes  $\epsilon$  et ils appartiennent au "buffer" de la ligne à contrôler (les polygones en bleu sur la figure I-11). A l'instar de l'indicateur utilisant le dénombrement des polygones parasites, cet indicateur ne peut constituer, en aucun cas, une mesure d'écart entre deux saisies.

A côté de ces outils développés autour de la technique de la bande  $\epsilon$ , d'autres outils ont été développés pour doter les SIG de méthodes alternatives de mesure des écarts géométriques des primitives linéaires. Deux outils ont largement bénéficié de

l'attention des chercheurs ces dernières années, en l'occurrence la distance de Hausdorff et la distance de Fréchet (que nous développons plus loin dans les §II.4.2.1 et §II.4.2.2 pour examiner la possibilité de leur utilisation pour la mesure des écarts géométriques des primitives surfaciques).

Si toutes (ou presque) les méthodes d'évaluation des écarts de géométrie linéaire peuvent être étendues aux objets surfaciques en les considérant comme des polygones fermés, elles présentent toujours des faiblesses si l'on se trouve face à des configurations complexes de polygones.

Cependant, il existe des outils et des méthodes d'évaluation spécifiques aux primitives surfaciques. La majeure partie des outils provient du domaine de la reconnaissance des formes, en donnant aux objets une modélisation descriptive qui se rapproche plus de certains indicateurs de qualité géométrique. Ainsi, dans le contexte de la généralisation cartographique, les travaux de [Buttenfield 1991; Plazanet 1996] ont permis de décrire la forme, la largeur, la curvilinéarité, la sinuosité, etc., des objets géométriques linéaires. Des outils ont été également développés pour les objets surfaciques en fonction de la problématique traitée. Pour les tâches de superposition, [Chrisman & Lester 1991] ont développé une série d'indicateurs décrivant la forme des polygones parasites. [Regnauld 1998] a également développé une série d'indicateurs pour analyser la similarité entre les bâtiments afin de les regrouper dans un contexte de généralisation des données géographiques. Ces indicateurs peuvent être divisés en deux catégories:

- ✓ Des indicateurs servant à l'étude de la forme des objets, dans le but de détecter les caractéristiques de chacun.
- ✓ Des indicateurs servant à détecter un caractère typique d'un objet (essentiellement développés sur les bâtiments). Ce type d'indicateurs sert comme mesure de contrôle et permet de vérifier que les objets générés, après un processus de généralisation, conservent toujours les mêmes caractéristiques [Regnauld 1998].

### **Indicateurs simples pour qualifier la forme**

Nous présentons ci-après quelques-uns de ces indicateurs, sans toutefois prétendre à l'exhaustivité de la liste.

Les deux indicateurs les plus utilisés sont les mesures de l'aire et du périmètre du polygone. Ces deux indicateurs donnent une information intrinsèque sur le polygone mesuré, mais ils ne peuvent pas être considérés comme des paramètres pouvant qualifier le polygone d'une manière unique. Leur utilisation pour des besoins de qualification de la forme des polygones s'avère donc impossible, voire donnant un résultat biaisé.

Cependant, la majeure partie des indicateurs a été développée en utilisant les mesures de l'aire et du périmètre.

**Concavité :**

Une partie d'un objet est dite concave si elle est située strictement à l'intérieur de l'enveloppe convexe de l'objet sans appartenir à l'objet lui-même. Pour mesurer cette caractéristique, on rencontre dans la littérature plusieurs définitions pour cet indicateur [Chassery & Montauvert 1991; Regnaud 1998]

Le degré de concavité d'un objet (également appelé, dans sa forme réciproque degré de convexité) est défini comme la déviation de l'objet par rapport à son enveloppe convexe, et donné par:

$$Ic = \frac{\text{Surface(objet)}}{\text{Surface(Enveloppe\_convexe)}} \quad [I-6]$$

Ce paramètre sans dimension est maximal et vaut 1 pour un objet convexe. Il est en général significatif de la présence de fortes concavités lorsque sa valeur est nettement inférieure à 1.

Deux autres indicateurs, traduisant respectivement l'importance des grandes concavités (GC) et des petites concavités (PC), peuvent également être définis par:

$$Gc = \frac{\sum_{i=1}^{Nc} S(i)^2}{Nc} \quad \text{et} \quad Pc = \frac{\sum_{i=1}^{Nc} \frac{1}{S(i)^2}}{Nc} \quad [I-7]$$

Où Nc représente le nombre de concavités de l'objet étudié, et S(i) représente la surface de la concavité de l'indice i. Si la forme ne représente que des concavités de grandes surfaces, l'indicateur Gc est élevé et l'indicateur Pc est faible, alors que dans le cas d'un polygone présentant des petites concavités (contour "dentelé") l'indicateur Gc est faible et l'indicateur Pc est élevé. La figure I-12 illustre un exemple de calcul de ces indicateurs.

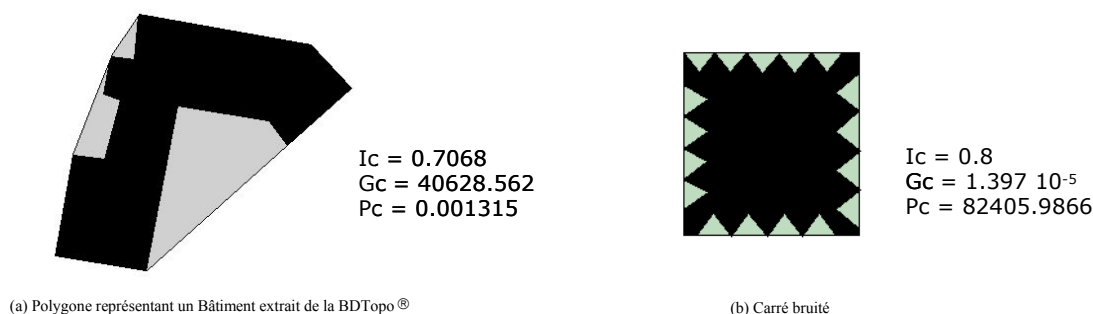


Figure I-12 : indice de concavité

**Allongement:**

D'une manière intuitive, l'allongement d'un polygone varie suivant le rapport entre sa longueur et sa largeur. Cet indicateur peut être calculé par plusieurs méthodes.

Pour les objets convexes, cet indicateur peut être donné par le rapport entre la largeur et la longueur. La longueur de l'objet est donnée par la plus grande section de l'objet orientée suivant l'orientation générale de celui-ci. La largeur est la plus grande



section selon une orientation perpendiculaire à celle de la longueur. En d'autres termes, l'allongement est donné par le rapport entre les côtés du rectangle englobant de l'objet à mesurer. Etant donné que l'orientation générale de l'objet ne peut pas être donnée d'une manière rigoureuse, [Freeman 1975] propose de calculer cet indicateur pour toutes les orientations du rectangle englobant afin de lever toute ambiguïté sur le choix de l'orientation de l'objet. La valeur de cet indice pour l'exemple du polygone de la figure I-12(a) est de 0.6628 (Longueur = 61.75 m et largeur = 40.93 m).

Une autre version de cet indicateur, également définie pour les polygones convexes, est donnée par le rapport entre le rayon du cercle inscrit et le rayon du cercle circonscrit. Si l'utilisation de cette version de l'indicateur est répandue, elle reste toutefois litigieuse du fait qu'elle renvoie une valeur inattendue dans le cas du carré (l'allongement du carré est différent de 1, par cet indice). Cet indice pourrait être classé dans la famille des indices de compacité [Regnauld 1998]. La valeur de cet indice pour l'exemple du polygone de la figure I-12(a) est de 0.1774 (rayon du cercle inscrit = 5.5 m et rayon du cercle circonscrit = 31 m)

Une autre version de cet indicateur dite indice d'allongement géodésique est définie par la formule suivante :

$$I_{ag} = \frac{4}{\pi} \frac{\text{Aire}}{\text{Longueur}^2} \quad [\text{I-8}]$$

La longueur est donnée par la longueur maximale de l'objet, toutes directions confondues. Cet indice évolue entre 0 et 1, et il n'est pas restreint aux seuls objets convexes. La valeur renvoyée par cet indice est difficilement interprétable, puisqu'on ne peut pas savoir si cette valeur donne une idée sur l'allongement réel de l'objet ou si elle relate un fort taux de concavité de l'objet. La valeur de cet indice pour l'exemple du polygone de la figure I-12(a) est de 0.33903.

Les trois versions de l'indicateur de l'allongement évoluent entre 0 et 1, cependant on remarque que les valeurs calculées par ces trois versions pour un même objet ne sont pas comparables. La première version indique que le polygone n'est pas trop allongé, par contre, les deux autres versions indiquent le contraire. La différence entre les valeurs renvoyées peut s'expliquer par la forte concavité de l'objet testé, ce qui confirme encore une fois que ces indicateurs sont souvent confus en la présence d'une forte concavité. Cependant, ces indicateurs ne peuvent pas être considérés comme des indicateurs robustes pour décrire la forme des entités surfaciques.

### **Compacité:**

L'indice de compacité est souvent utilisé pour mesurer la ressemblance d'un polygone par rapport à un cercle (compacité maximale) ou un segment (compacité nulle). Cet indice est également défini de plusieurs façons en fonction de l'utilisation envisagée.

[Coster & Chermant 1989] définissent cet indice par la formule suivante :

$$I_c = 16 \frac{\text{Aire}}{\text{Perimètre}^2}, \text{ pour l'exemple de la figure I-12(a), } I_c = 0.4802 \quad [\text{I-9}]$$

[Chrisman & Lester 1991] définissent cet indice par une autre formule :

$$I_c = 2 \sqrt{\frac{\pi \cdot \text{Aire}}{\text{Perimètre}^2}}, \text{ pour l'exemple de la figure I-12(a), } I_c = 0.614 \quad [\text{I-10}]$$

L'utilisation de ces indicateurs tente de doter les entités surfaciques de plus d'informations permettant de mieux comprendre leurs caractéristiques géométriques. Cependant, leur utilisation pour mesurer les écarts de forme entre les objets surfaciques s'avère très difficile, du fait que ces indicateurs ne peuvent pas qualifier un objet d'une manière unique (deux objets de formes différentes peuvent renvoyer une même valeur pour un indicateur donnée). Par conséquent, leur utilisation reste entachée de confusion pour l'accomplissement de la tâche de contrôle de la qualité de forme. D'autre part, chacun de ces indicateurs représente une caractéristique particulière de l'objet, et donc leur synthèse pour qualifier la géométrie de l'information géographique (qualité absolue) s'avère très difficile.

Ces indicateurs peuvent être utilisés pour des tâches particulières, telle la distinction entre les différentes composantes de la qualité, ou tout simplement le contrôle de la qualité. Nous détaillons dans la section suivante une méthode développée par [Chrisman & Lester 1991] faisant référence. Cette méthode utilise ce genre d'indicateur pour contrôler les jeux de données surfaciques.

### **I.3.5. Approches utilisées pour qualifier la géométrie des objets surfaciques.**

Les approches antérieures utilisées pour évaluer la qualité de la géométrie des objets surfaciques sont fondées sur les techniques de superposition. La superposition est l'une des applications qui permet de mettre en valeur les problèmes liés à la précision géométrique [Vauglin 1997]. Cette application est considérée comme un élément de base de l'analyse spatiale, puisqu'elle permet de reproduire une information en intégrant une multitude de phénomènes géographiques.

Une des opérations, qui nécessite des opérations de superposition, est l'opération de fusion de données. Cette opération consiste à fusionner différents jeux de données superposés. Mais, comme les objets des jeux de données sont issus de différentes saisies, ils n'ont jamais rigoureusement les mêmes coordonnées. Deux problèmes se posent lors de la mise en œuvre de cette opération :

- ✓ Le premier problème consiste à identifier les objets résultants de la fusion des objets des jeux de données superposés.
- ✓ Le deuxième problème découle de la complexité de la mise en correspondance des objets issus des différentes couches : malgré leur géométrie différente, la mise en correspondance (souvent appelée appariement) doit permettre d'identifier tous les objets homologues représentant le même phénomène géographique. Ce problème fait partie intégrante des travaux de cette thèse.

En se servant des opérations de superposition de données, [Chrisman & Lester 1991] ont développé une méthode de contrôle de la qualité des entités surfaciques, tant au niveau de la géométrie qu'au niveau de la sémantique. Cette méthode consiste, après superposition des deux jeux de données, à classer les différents objets générés dans quatre classes : "pas d'erreur", "erreur de position", "erreur sémantique" et "zone de confusion". La méthode proposée est illustrée par la figure I-13, et elle utilise trois indicateurs : l'aire des polygones issus du croisement, leur compacité et un indice de périmètre. L'indice de périmètre est défini spécifiquement pour cette application. Il suppose que le contour d'un polygone issu d'une opération de croisement entre deux jeux de données est composé de deux parties provenant chacune d'un des deux jeux de données. L'indice de périmètre est défini comme suit :

Soit un polygone parasite dont le contour est issu d'une source de données pour une longueur  $l_a$  et d'une autre source pour une longueur  $l_b$ . L'indice de périmètre  $P$  du polygone est défini par:

$$P = \frac{l_a}{l_a + l_b} \quad [I-11]$$

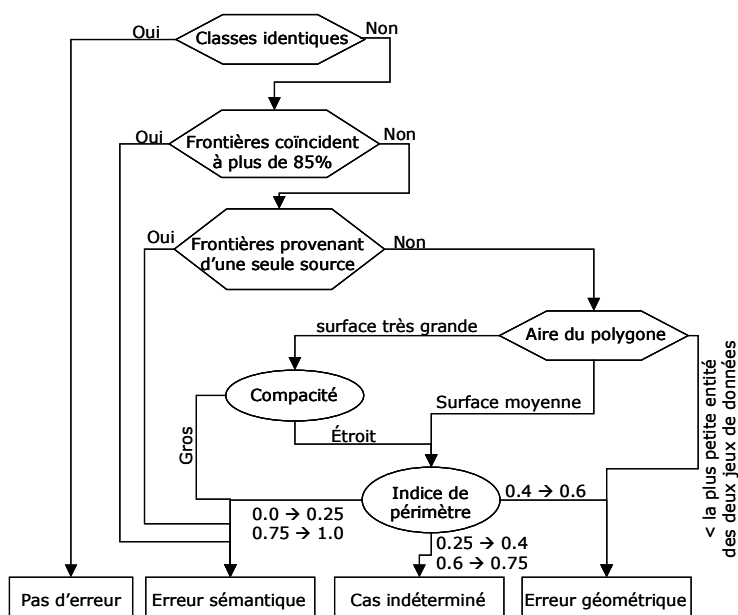


Figure I-13 : test de [Chrisman & Lester 1991]

Les polygones parasites sont utilisés comme fil conducteur pour pouvoir dissocier la qualité géométrique de la qualité sémantique. Ces polygones sont identifiés, soit par leur taille, soit par leur forme, respectivement petite et étroite. Un indice de compacité peut être l'instrument analytique utilisé, avec un indice d'étroitesse et de taille. Cependant, ces critères sont dépendants de la résolution, ce qui peut fausser les résultats. On peut penser à d'autres critères : par exemple, les polygones parasites sont souvent adjacents le long de la vraie ligne. Ce critère topologique est cependant difficile à mettre en œuvre.

Une autre caractérisation du polygone est proposée : un polygone parasite a des arêtes provenant, de façon à peu près équitablement répartie, des deux sources. Dans le cas où les résolutions d'origine seraient égales, les longueurs des périmètres provenant de chaque source seraient à peu près égales. D'où la définition d'un indice de périmètre.

Les tests effectués montrent que cet indice évolue entre les valeurs 0.4 et 0.6. De telles valeurs indiquent une erreur de position. Pour les valeurs proches de 0 ou de 1, on peut penser à une erreur d'exhaustivité, par exemple, toutes les arêtes du polygone proviennent alors d'une seule source, ce qui est considéré par les auteurs comme une erreur sémantique. Il se trouve que certains polygones étroits soient classés en erreur géométrique, alors qu'en réalité ils ne le sont pas.

On retrouve, encore une fois dans ce travail, la décomposition de l'erreur géométrique en deux types : une erreur de saisie et une erreur de généralisation.

La méthode développée par [Chrisman & Lester 1991] a pu démontrer la puissance des techniques de superposition des données géographiques à des fins d'analyse. Cependant, le problème de contrôle de la qualité géométrique n'est pas totalement résolu du fait que les mesures utilisées ne sont pas assez discriminantes pour pouvoir en tirer des conclusions robustes. La méthode permet également de faire la différence (d'une manière partielle) entre les erreurs géométriques et les erreurs sémantiques. Si l'utilisation de cette méthode peut fournir une évaluation de l'erreur sémantique à travers des matrices de confusion, l'évaluation de l'erreur géométrique est appréhendée d'une manière qualitative.

La superposition des données se présente comme un outil puissant pour l'analyse spatiale. Cependant, elle n'est pas assez développée pour analyser des données surfaciques [Harvey & al. 1998].

Les travaux réalisés par [Le Men & Jamet 1993] présentent une étude plus poussée que celle de [Chrisman & Lester 1991] dans laquelle ils dépassent la simple utilisation des opérations de superposition aux processus de mise en correspondance entre les objets des deux jeux de données à analyser. Une analyse de la typologie des erreurs est également réalisée en répondant à la question : jusqu'à quel point une déformation géométrique peut être comptabilisée en erreur géométrique, et, à partir de quand, cette déformation devenant trop importante, faut-il considérer qu'il y a erreur d'exhaustivité. Cependant, les indicateurs utilisés dans cette étude ne sont pas assez discriminants pour pouvoir établir une typologie des erreurs d'une manière nette, d'où le constat fait par [Le Men & Jamet 1993] :

*Des recherches sont indispensables pour qu'au-delà de la normalisation de la typologie intuitive des erreurs, on puisse définir des méthodes standard de calcul qui permettent l'estimation de ces critères, leur donnant un contenu mathématique précis et leur associant un processus de calcul opérationnel.*

## **I.4. APPARIEMENT DES DONNEES GEOGRAPHIQUES**

### **I.4.1. Définition de l'appariement**

L'estimation de la qualité d'une base de données géographiques consiste à la comparer à une base de données de référence dont on connaît la qualité ou dont on estime présenter une qualité meilleure que celle à contrôler. L'estimation se fait en calculant l'écart entre les objets de la base de données et leurs « homologues » dans la base de référence. Donc, avant de procéder à l'estimation de la qualité, une opération de mise en correspondance ou une recherche des objets homologues s'impose. Cette opération, qui vient naturellement en amont de la phase d'estimation, est appelée «phase d'appariement »

*L'appariement des données géographiques est le processus qui consiste à établir des liens de correspondance entre les objets géographiques représentant les mêmes phénomènes du monde réel mais provenant de jeux de données différents.*

### **I.4.2. A quoi ça sert?**

L'appariement des données géographiques est considéré comme un outil puissant pouvant répondre à divers problèmes rencontrés fréquemment dans la gestion et l'analyse des données géographiques. Nous citons dans ce paragraphe quelques champs d'application qui nécessitent *a priori* l'utilisation de l'appariement.

#### **I.4.2.1. Mise à jour de base de données géographiques**

La mise à jour de bases de données peut être envisageable selon deux perspectives : de point de vue du producteur étant responsable de la mise à jour de ses données et du côté de l'utilisateur désirant intégrer les mises à jour fournies par le producteur [Badard 2000].

Du point de vue du producteur : le processus de l'appariement est utilisé pour mettre en correspondance les primitives géométriques entre l'ancienne et la nouvelle base de données afin de fournir des données ne contenant que les modifications sémantiques, géométriques ou topologiques.

Du point de vue de l'utilisateur : l'appariement peut devenir indispensable si l'utilisateur a entrepris des modifications dans l'ancienne base de données en enlevant, modifiant ou créant des objets ou en créant des liens entre les objets de la base fournie par le producteur et ses propres bases. En effet, l'utilisateur doit comparer la nouvelle version de la base avec la base dont il dispose, afin de la mettre à jour en tenant compte des anciennes modifications.

### **I.4.2.2. Unification des bases de données et création des serveurs multi-échelles**

Le but d'une base de données multi-échelles est de pouvoir manipuler des objets géographiques ayant des niveaux d'abstraction différents et de pouvoir retrouver les liens entre un objet à une échelle donnée et un ou un ensemble d'objets à une autre échelle. La construction d'une base de données géographiques multi-échelles fait du processus d'appariement la pierre angulaire pour sa création et sa mise en place [Devogele 1997]. Ce genre d'application soulève des problèmes d'appariement, notamment en ce qui concerne l'établissement de liens multiples lors de la mise en correspondance de bases de données à des échelles différentes.

### **I.4.2.3. Contrôle de la qualité**

Le contrôle et/ou l'estimation de la qualité d'une base de données géographiques nécessite de la comparer à des données de référence. Celle-ci est en général une base de données de précision égale ou supérieure. A l'amont de la comparaison, il est nécessaire de mettre en correspondance les objets des deux bases qui sont censés représenter le même phénomène réel. A titre d'exemple, pour juger la qualité de la BDCarto®, on peut prendre comme référence la BDTopo®. Or, pour cet exemple, on compare deux bases de données n'ayant pas la même échelle ni le même degré de généralisation puisqu'on ne cherche plus à apparier un objet de la BDCarto® avec un objet de la BDTopo®, mais plutôt apparier un objet avec plusieurs. Il s'agit donc de comparer la géométrie d'un objet à celle d'un agrégat qui lui est homologue et de vérifier en même temps la qualité sémantique.

## **I.4.3. Etat de l'art**

Pour la mise en correspondance des entités géographiques des différentes bases de données géographiques, on utilise souvent trois types d'appariement : l'appariement géométrique, l'appariement topologique et l'appariement sémantique.

L'appariement topologique utilise les relations de composition ainsi que les relations topologiques. L'appariement topologique est souvent utilisé pour mettre en correspondance les données de type réseau [Gabay & Doytsher 1994; Dimitrijévic 2000].

L'appariement sémantique est une technique qui consiste à mettre en correspondance les objets, grâce à la valeur des informations sémantiques portées par ceux-ci (noms et valeurs des attributs, nom de la classe, relation entre les classes, etc.). Ce type d'appariement peut avoir lieu si les deux bases de données présentent des schémas de données "proches", à défaut, son utilisation s'avère difficile sans une étape préalable de mise en correspondance des schémas des bases de données à apparier.

L'appariement géométrique repose sur la comparaison de la position et de la forme des objets. Ce type d'appariement est le plus souvent utilisé, mais il reste le plus complexe à mettre en œuvre du fait qu'il est difficile d'expliquer la similarité de deux objets. En effet, il y a des critères qui rentrent en ligne de compte et qui ne sont pas formalisables [Kidner 1996]. L'appariement géométrique devient vite complexe du fait de plusieurs facteurs :

- ✓ Il est difficile de détecter que deux objets occupent la même position dans l'espace sachant que chaque point de l'objet est entaché d'erreurs plus au moins aléatoires;
- ✓ Etant donné qu'un objet d'un jeu de données peut-être apparié avec plusieurs objets de l'autre jeu de données ou même aucun objet, la notion d'objet le plus proche n'est pas suffisante;
- ✓ L'appariement doit tenir compte de l'information contextuelle, en analysant la répartition des objets les uns par rapport aux autres dans l'espace géographique.

La généralisation ajoute une erreur géométrique sur la position des objets. Cette erreur supplémentaire n'est pas négligeable quand on essaie d'apparier des objets provenant de bases de données destinées à des utilisations à des échelles différentes.

Ces techniques d'appariement peuvent être utilisées séparément ou de manière complémentaire.

Les techniques d'appariement sont souvent utilisées pour répondre aux divers problèmes rencontrés lors de la manipulation ou de la gestion des données géographiques. On cite à titre d'exemple l'alignement et la mise en correspondance de bases de données juxtaposées, la mise à jour des bases de données, la création des serveurs multi-échelles, l'intégration de bases de données ou superposition des cartes numériques (*Overlay*, en anglais), le contrôle de la qualité des données géographiques, etc. Il faut noter qu'à chaque application correspond un ou plusieurs types particuliers d'appariement. L'utilisation d'un appariement topologique, par exemple, est fortement recommandée pour apparier des données de type réseaux (routiers, ferrés, etc.).

Plusieurs approches algorithmiques d'appariement de données spatiales ont été mises en œuvre dans les SIG. Ces approches consistent essentiellement en des algorithmes utilisés pour fusionner les données similaires sans aucune action de conservation des étapes de pré-traitement de l'appariement [Pullar 1991; Zhang & Tulip 1990]. La fusion des données par l'utilisation de ces techniques n'est pas tout à fait mise au point, à cause de la modification arbitraire et non contrôlée de la géométrie des primitives fusionnées [Goodchild & Cova 1995].

D'autres approches plus récentes consistent à rechercher des structures identiques d'objets entre deux bases de données afin d'échanger les informations attributaires et d'homogénéiser la géométrie. Ce genre d'approche n'est cependant utilisable que pour les données acquises selon le même modèle de données traitant les mêmes thématiques et de sources différentes [Walter & Fritsch 1999].

La majorité des travaux sur l'appariement s'est focalisée sur les données linéaires. Une des premières approches d'appariement entre des données de modèles différents est développée par le "*Bureau of the Census in Washington DC*" [Rosen & Saalfeld 1985; Saalfeld 1988]. Un système a été développé pour fusionner les données numériques réalisées par le *Bureau* et celles provenant de l'USGS<sup>9</sup>. Il consiste en un algorithme itératif et repose sur l'appariement des nœuds et sur l'utilisation d'une déformation élastique des arcs (*rubber sheeting* – en anglais –) [Gillmann 1985]. Cette approche prend comme hypothèse de départ le fait que les deux jeux de données à appairer sont "isomorphiquement" proches et par la suite ne prend en considération que les appariements de type 1-à-1 (de type objet à objet). Par conséquent, ce genre d'approche n'est applicable que pour des données à des échelles proches.

Cependant, [Brown & Baran 1995] mettent l'accent sur l'importance de la détection des appariements de type n-à-m (correspondance de n objets avec m objets), puisqu'ils estiment que les appariements simples de type 1-à-1 sont insuffisants pour traiter des jeux de données dont l'échelle d'acquisition n'est pas la même.

[Gabey & Doytsher 1994] ont examiné la possibilité d'appariement de deux jeux de données qui diffèrent peu par la géométrie mais beaucoup par les propriétés topologiques, et à des échelles comparables. Leur approche fonctionne d'une manière itérative en cherchant d'abord des candidats à l'appariement en se basant sur les critères de proximités, et en affinant le choix par l'utilisation d'autres critères géométriques et topologiques. Il faut noter que la méthode qu'ils ont développée fonctionne pour les données géographiques linéaires.

On trouve dans [Laurini 1998] un état des problèmes qui peuvent avoir lieu lors de l'intégration de deux bases de données géographiques. Il suggère également quelques-uns des concepts nécessaires à l'accomplissement de l'interopérabilité des bases de données géographiques.

## **I.5. SYNTHÈSE**

Notre travail s'inscrit dans la recherche de nouvelles mesures permettant d'évaluer les écarts entre les primitives surfaciques, en tenant compte des spécificités de ces primitives. Une primitive surfacique doit être traitée comme une partie surfacique de l'espace et non pas comme un simple contour.

Dans un premier temps, nous avons retracé dans ce chapitre le concept général de la qualité des données géographiques en donnant les définitions des composantes qui la définissent, sur lesquelles les intervenants se sont accordés et qui font aujourd'hui l'objet d'une norme. Chaque composante de la qualité est décrite par un ensemble d'indicateurs qui reste à l'heure actuelle défini plus ou moins librement par les utilisateurs et les producteurs des données géographiques. Ces indicateurs ne peuvent pas être

<sup>9</sup> USGS: United States Geological Survey



universalisés, ni employés d'une manière générique, puisqu'à chaque application ou type d'usage de données correspond un type particulier d'indicateurs.

Nous avons porté une attention particulière sur la composante de la qualité géométrique en nous focalisant sur l'étude de l'existant. Ce chapitre a présenté deux volets : la description statistique des erreurs de localisation des primitives géométriques dans les bases de données, et les outils utilisés pour évaluer les écarts entre les primitives. La majeure partie des travaux sur la qualité géométrique s'est intéressée essentiellement à l'étude des erreurs des primitives géométriques ponctuelles et linéaires.

La description des erreurs géométriques des primitives ponctuelles et linéaires par des modèles statistiques a atteint une certaine maturité, ce qui a permis de bien comprendre le comportement de ces erreurs. D'une manière générale, les erreurs géométriques peuvent être décomposées en deux grandes classes : des erreurs générées lors de la saisie des données et qualifiées comme des erreurs de pointé et de généralisation, et des erreurs générées lors de l'interpolation entre les points saisis lors de la construction d'une ligne, par exemple. Le modèle statistique décrivant ces erreurs sera utilisé dans le cadre des travaux de cette thèse pour simuler les bruits et les déformations qui peuvent affecter les primitives géométriques dans une base de données. Cependant, nous ne cherchons pas à décrire les incertitudes de la géométrie des primitives surfaciques. Nous nous contentons d'utiliser le modèle établi pour les primitives linéaires qui a été validé sur des données surfaciques (de type occupation des sols).

L'évaluation des écarts géométriques et le contrôle de la qualité des primitives surfaciques (polygones) sont souvent faits en réduisant le polygone à son contour ou simplement aux points de son contour. Par conséquent, cette démarche revient à ramener la problématique à un simple contrôle linéaire ou ponctuel. Cependant, une primitive surfacique est plus complexe qu'une primitive linéaire, puisqu'elle est définie par un intérieur et un contour [Mark & al. 1999]. Cette spécificité peut être perçue d'une manière simple par la différence entre un cercle et un disque.

L'application des outils initialement développés pour la mesure des écarts des primitives ponctuelles et linéaires sur les primitives surfaciques n'est pas suffisante pour rendre compte de l'écart existant entre les entités. Ces outils ne permettent que l'évaluation de l'écart de position et sont loin d'être pertinents pour traduire les écarts de forme qui sont d'une importance égale à celle des écarts de position pour les primitives surfaciques. L'utilisation de la distance euclidienne pour évaluer l'écart entre les points des contours des polygones nécessite *a priori* l'identification des points homologues entre les deux contours et elle ne permet pas de renvoyer une mesure qui reflète la réalité des écarts entre les entités surfaciques (cf. figure I-14).

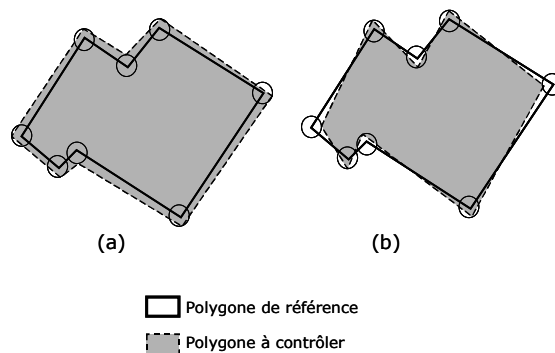


Figure I-14 : insuffisance du contrôle ponctuel pour décrire les écarts entre les polygones

L'exemple de la figure I-14 illustre l'insuffisance du contrôle ponctuel car le résultat obtenu par ce type de contrôle renvoie le même résultat pour les deux entités à contrôler, bien que leurs formes soient manifestement différentes. Toutefois, le résultat de ce contrôle peut être interprété selon la nature de l'application utilisant ces données.

L'évaluation des écarts entre les primitives linéaires est faite par l'utilisation de la distance de Hausdorff, ce qui a permis d'apporter une solution alternative au contrôle ponctuel. Si l'application de cette distance sur les contours des primitives surfaciques permet de donner une indication sur les écarts de position, elle reste toujours incapable de donner une indication fiable sur les écarts de forme. La figure I-15 illustre quelques exemples.

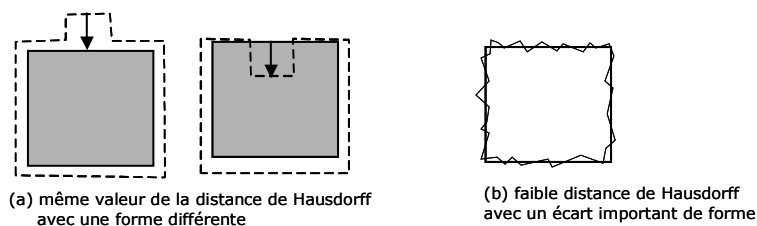


Figure I-15 : insuffisance de la distance de Hausdorff pour mesurer les écarts de forme

Nous avons également analysé dans ce chapitre quelques indicateurs simples pour décrire quelques caractéristiques géométriques (allongement, compacité, etc.) des entités surfaciques. D'une part, ces indicateurs donnent une indication particulière sur la forme de l'objet et ne permettent pas de rendre compte de toutes les caractéristiques géométriques de l'objet à traiter. D'autre part, ces indicateurs montrent une défaillance au niveau de leur robustesse, puisque à une même valeur de l'un de ces indicateurs correspond un ensemble d'entités surfaciques de forme différente.

Nous avons également décrit dans ce chapitre quelques approches utilisées pour étudier les écarts de géométrie entre les entités surfaciques. Ces approches sont fondées sur les techniques de superposition des données et elles utilisent des indicateurs simples, essentiellement pour différencier la qualité géométrique de la qualité sémantique et dire jusqu'à quel point une déformation géométrique peut être comptabilisée comme erreur géométrique et à partir de quel point il faut la considérer comme erreur sémantique. Par ailleurs, ces approches ne permettent pas d'évaluer les erreurs géométriques d'une

manière quantitative. Enfin, aucune information n'est synthétisée concernant la mise en correspondance entre les entités des deux jeux à analyser.

Cette analyse a permis de définir le cadre de notre étude. Nous allons à présent définir de nouveaux espaces de représentation afin de mieux cerner et rendre compte des caractéristiques géométriques des entités surfaciques. Ces nouveaux espaces de représentation doivent être dotés d'outils d'évaluation quantitative des écarts de forme et de position entre les entités surfaciques.

**CHAPITRE II :**  
**REPRESENTATIONS DE LA GEOMETRIE DES**  
**ENTITES SURFACIQUES ET MESURES**

## II.1. INTRODUCTION

Les mesures ou le contrôle de la qualité des primitives géométriques dans les bases de données géographiques ont souvent été abordées avec une attention particulière portée sur l'erreur de la position, sans trop se soucier de l'erreur induite par les différences de la forme des objets [CNIG 1993]. Le contrôle de la position sans contrôle de la forme est souvent synonyme de contrôle ponctuel, et s'il est satisfaisant pour le ponctuel, il l'est moins pour les primitives linéaires, et encore moins pour les primitives surfaciques. Cependant, l'évaluation de la qualité de la forme, pour les primitives surfaciques apparaît nécessaire, voire d'une action d'égale importance que celle de l'évaluation de la qualité de position.

L'étude de la forme des objets géométriques a bénéficié d'une large attention pour des objectifs de reconnaissance des formes ou de vision par ordinateur, voire dans une moindre mesure pour le traitement des images. Cependant, ces études de forme ont bénéficié d'un intérêt particulier dans le domaine de l'intelligence artificielle avec l'objectif de modéliser les connaissances du "sens commun" -*common sense* en anglais- de l'espace géographique afin de construire une nouvelle génération de SIG [Egenhofer & Mark 1995].

En reprenant la définition du SDTS (*Spatial Data Transfer Standard*) [NIST 1992], une surface est définie, d'une manière générique, comme:

*un objet spatial à deux dimensions, borné et continu. Un polygone est décrit comme une surface avec un contour externe et plusieurs contours internes non superposables. Un contour est défini comme un objet uni-dimensionnel fermé, composé par une suite continue de segments. Chaque segment est représenté par une ligne joignant deux points.*

Un objet spatial à deux dimensions ou polygone peut être également vu comme un complexe simplicial de dimension 2 défini dans l'espace euclidien, dans le sens où il peut être décomposé par une triangulation de delaunay contrainte par le contour. Ces polygones sont considérés comme stables par une réunion finie de simplexes géométriques de dimension 2 (triangle) dont les intersections deux à deux sont vides.

Bien que cette définition soit largement utilisée par la plupart des SIG actuels, elle n'est pas la plus adaptée pour des traitements particuliers, tel la traduction des caractéristiques de la forme des entités géographiques. L'exploration de nouvelles représentations peut donc être envisagée.

### II.1.1. Pourquoi de nouvelles représentations?

Lors de la mise en correspondance de deux bases de données, notamment celles dont la résolution n'est pas la même, on aboutit à différents types de cardinalité. Lors de la mise en correspondance entre les objets de deux bases de données, on retrouve plusieurs type de liens qui peuvent être classés comme suit:

- ✓ Liens de cardinalité 1-à-0 ou 0-à-1 : ce type de lien concerne les objets d'une base de données qui n'ont pas d'objets homologues dans l'autre base de données (et réciproquement);
- ✓ Liens de cardinalité 1-à-1 : ce type de lien concerne la correspondance des objets simples entre eux.
- ✓ Liens de cardinalité 1-à-m, n-à-1 et n-à-m : ce type de lien permet de mettre en correspondance un objet ou un agrégat d'objet d'une base de données avec un agrégat d'objets simples de l'autre base de données.

Nous convenons de classer les entités surfaciques en deux catégories :

- ✓ Entité simple ou élémentaire : définie en §II.1.
- ✓ Entité complexe : sont constituées par un agrégat d'entités élémentaires. L'agrégat peut être constitué d'entités élémentaires disjointes ou non (cf. figure II-1). Ces agrégats peuvent être vus comme des agglomérations construites par les entités simples dites membres ou participants [Smith 1999]. La construction d'un agrégat est dictée par le lien auquel ses membres participent dans un contexte donné.

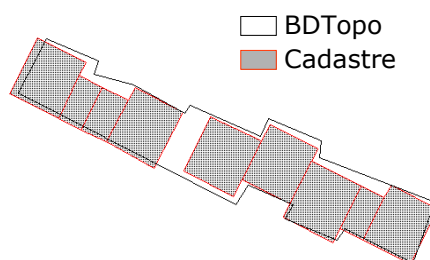


Figure II-1 : exemple d'une entité complexe

*les 9 polygones du cadastre représentés en rouge constituent une entité complexe à part entière du fait qu'ils représentent le même phénomène du monde physique que le polygone (noir) de la BDTopo®*

La classification des entités sous les deux aspects de simple et complexe est fortement dépendante du contexte de l'appariement. Ce qui peut être défini comme une entité simple dans la mise en correspondance de deux bases de données à la même échelle peut contribuer à la définition d'une entité complexe, si les deux bases ne sont pas à la même échelle.

La représentation en mode vecteur (liste chaînée de points) semble difficilement exploitable pour traduire les caractéristiques de la forme d'une entité donnée. En effet, l'information portée par chaque paire de coordonnées (point) est une information trop parcellaire et localisée pour pouvoir donner une indication générale sur la géométrie de l'entité. Chaque point n'est porteur que de l'information de l'endroit "par où passe la ligne", sans donner aucune indication sur la forme. Ce constat a été également établi par

[Fritsch 1997] dans ses travaux de recherche de nouvelles représentations pour les primitives linéaires à des fins de généralisation.

Compte tenu de la spécificité des données surfaciques, notamment si nous voulons traiter le polygone en tenant compte de son intérieur ou bien des polygones simples avec des trous, nous avons tenté de définir des nouvelles mesures et métriques dans l'espace cartésien. Elles s'ajoutent aux mesures existantes, soit pour pallier les faiblesses de ces dernières, soit pour être utilisées d'une manière complémentaire.

Pour des raisons de souplesse à l'implémentation informatique de l'information géométrique dans les BDG, la plupart des SIG actuels représentent l'information géographique sous forme de points, lignes et polygones. D'une part, les lignes sont souvent représentées par une succession de points reliés entre eux par interpolation et les polygones sont des lignes fermées. Ce type de représentation qui se situe dans l'espace cartésien est loin d'être la plus appropriée pour refléter toutes les caractéristiques de l'objet, surtout si l'on veut traiter sa forme [Clementini & Di felice 1997]. D'autre part, les entités complexes sont définies par un agrégat d'entités simples disjointes ou non. Ces deux raisons nous poussent à aller plus loin dans la recherche de nouvelles représentations et de nouveaux espaces de représentation, différents de l'espace cartésien et de la modélisation vecteur sous la forme d'une suite de points. Nous devons également développer un ensemble de métriques et d'indicateurs (que nous traiterons dans le §II.4). Cependant, en préambule à la définition de nouvelles représentations, nous introduisons dans le paragraphe suivant les critères auxquels une représentation doit se conformer pour s'assurer de sa capacité à traduire aux mieux les objectifs visés.

## **II.2. CRITERES DE CHOIX D'UNE REPRESENTATION**

On peut imposer à la représentation de la géométrie d'une entité de respecter certains critères de choix, afin de s'assurer de sa capacité à mieux traduire les caractéristiques géométriques de l'entité. En reconnaissance des formes, [Mokhtarian & Mackworth 1992 ] définissent six critères que nous avons repris pour les représentations et les modélisations et que nous présentons dans les paragraphes suivants. Ces critères sont les suivants :

### **II.2.1. L'unicité**

C'est la définition même de la représentation. Cette propriété assure une bijection ou simplement une injection entre l'ensemble des objets et l'espace des modélisations. Une représentation d'un objet doit être univoque. Elle ne doit représenter que cet objet. Sans cela, on préfère parler de caractérisation, plutôt que de représentation.

### **II.2.2. L'inversibilité**

Ce critère concerne la représentation d'une entité dans un espace en se basant sur sa représentation dans un autre espace, s'il existe une bijection entre ces deux espaces. Cette propriété nous intéresse peu dans le cadre de ce travail, puisque nous ne cherchons pas à faire le chemin inverse. Par ailleurs, elle trouve tout son intérêt dans le domaine de la généralisation des primitives géométriques dans les BDG. La représentation des primitives géométriques dans l'espace des fréquences permet quelques opérations de généralisation telles, par exemple, la suppression des détails insignifiants par élimination de leurs fréquences respectives. En effet, le résultat final est obtenu en effectuant le chemin inverse entre l'espace des fréquences et l'espace cartésien.

### **II.2.3. La stabilité**

La stabilité se traduit par le fait que les représentations des deux objets considérés comme "similaires" doivent être proches. Une stabilité théorique doit être attachée à la représentation elle-même.

Une stabilité algorithmique consiste à limiter au maximum la génération d'artefacts pouvant induire une divergence des résultats numériques. De façon pratique, il faut limiter les opérations de dérivation, d'intégration et de convolution numérique afin d'éviter toute divergence liée aux capacités de calcul des outils informatiques.

Mathématiquement, la stabilité se traduit par un critère de continuité de la représentation.

### **II.2.4. L'invariance**

L'invariance de la représentation de l'objet aux translations, aux rotations et aux homothéties est un critère primordial en reconnaissance des formes. Le respect de ce critère est primordial pour représenter la forme de l'entité, faute de quoi il n'y ait plus unicité de la représentation.

### **II.2.5. L'accès aux propriétés géométriques**

L'accès aux propriétés géométriques impose la possibilité de pouvoir partir de la représentation pour détecter aisément les propriétés géométriques essentielles de l'objet modélisé.



## II.2.6. Efficience et complexité algorithmique

Le critère d'efficience et de complexité algorithmique préconise de minimiser les dérives et les imprécisions dues au calcul ainsi qu'au temps de calcul. Ce critère est souvent recommandé en reconnaissance des formes nécessitant le traitement en temps réel. Ce critère ne constitue pas une priorité en soi dans le cadre de notre étude.

## II.3. REPRESENTATIONS DES OBJETS SURFACIQUES

Le contrôle de la qualité géométrique des entités géographiques dans les bases de données a bénéficié d'une large attention de la part de la communauté de chercheurs en SIG. L'essentiel des travaux met surtout l'accent sur le contrôle de la position d'un objet par rapport à son homologue dans le terrain nominal [Bédard 1987; Vauglin 1997]. La forme de l'objet a été traitée d'une manière subjective et elle a été plutôt laissée aux soins du contrôleur pour en décider l'acceptation ou le rejet [BI 1997].

Cependant, la qualité de la forme de l'objet a pourtant la même importance que sa position par rapport à l'objet nominal. Cela implique que l'accent soit également mis sur le contrôle de la forme des objets au même titre que le contrôle de leurs positions. C'est pourquoi, le contrôle de la géométrie d'un objet doit prendre en compte ces deux aspects. En outre, le contrôle de la forme de l'objet peut être d'une importance majeure, comme par exemple dans la mise à jour, la qualification et l'évaluation d'un processus de généralisation ou encore dans la mise en œuvre des bases de données multi-échelles.

Etant donné que l'espace cartésien est insuffisant pour représenter les formes des objets géographiques, il est nécessaire de chercher de nouveaux espaces. L'objectif est donc de passer d'une représentation dans l'espace cartésien muni d'une distance bornée à une représentation dans l'espace des fonctions variant dans  $[0,1]$ , muni d'une distance également bornée par exemple. Cependant, avant de définir les nouvelles représentations, nous tenterons de donner la définition de quelques espaces de représentations utilisés.

Il est possible de considérer les objets géographiques linéaires comme des ensembles de points particuliers de l'espace (des fermés bornés). Nous travaillerons par la suite sur les polygones fermés du plan, c'est à dire des lignes brisées fermées et bornées. Notons  $\mathcal{F}$  l'ensemble des fermés bornés de l'espace métrique  $(\mathbb{R}^2, d)^{10}$ , et  $\mathcal{F}'$  le sous-ensemble des polygones de  $\mathcal{F}$ .

On note également  $\mathcal{F}_0$  le sous-ensemble des surfaces (avec frontières) de  $\mathcal{F}$ . Pour distinguer les objets surfaciques définis par leurs contours (éléments de  $\mathcal{F}'$ ) des éléments non dégénérés (d'intérieur non vide) de  $\mathcal{F}_0$ , on appelle "polygones" les

<sup>10</sup> A titre d'exemple, on peut munir l'espace  $\mathbb{R}^2$  de la distance euclidienne.

premiers et "entités surfaciques polygonales" les seconds. On note que  $\mathcal{F}'$  est lui-même un sous-ensemble de  $\mathcal{F}_0$ . On a en fait la relation<sup>11</sup> suivante :

$$\mathcal{F}'_0 \subset \mathcal{F}' \subset \mathcal{F}_0 \subset \mathcal{F} \quad [\text{II-1}]$$

Après avoir donné les définitions des entités simples et complexes, ainsi que les définitions des espaces de représentation courants, cette section sera composée de trois parties.

La première partie est dédiée aux nouvelles représentations, dans laquelle seront détaillés des nouveaux espaces de représentations des éléments de  $\mathcal{F}'$  en les modélisant soit par les angles que forment les segments (cf. §II.3.1.2.), soit par les distances radiales (cf. §II.3.1.3.), soit par les fréquences (cf. §II.3.1.4.). Les éléments de  $\mathcal{F}_0$  seront à leur tour représentés par un ensemble fini de valeurs de leurs moments mathématiques (cf. §II.3.2.2.).

La deuxième partie de ce chapitre (cf. §II.4.) sera consacré à la définition de nouvelles métriques et de nouveaux indicateurs de similarités à associer à chaque espace de représentation.

Des tests ont été menés dans la troisième partie (cf. §II.5.) afin de tester la robustesse ainsi que le comportement des métriques et des indicateurs face à différents types de bruit.

### II.3.1. Pour les objets simples -représentation du contour-

Dans cette section, nous présentons les représentations utilisées pour représenter l'information géométrique des entités surfaciques en nous limitant à l'utilisation de leurs contours. Ces représentations ne sont valides que pour les éléments de  $\mathcal{F}'$  que nous désignons par "simples" ou "élémentaires". Nous commençons par la représentation la plus évidente utilisée actuellement par les SIG actuels.

#### II.3.1.1. Représentation cartésienne

Les objets surfaciques sont représentés dans l'espace cartésien soit par leurs contours (l'espace  $\mathcal{F}'$ ), soit par leurs intérieurs (l'espace  $\mathcal{F}_0$ ). La représentation via les contours se présente sous la forme d'une liste chaînée de points appartenant à  $\mathbb{R}^2$  ou  $\mathbb{R}^3$ . C'est la représentation la plus utilisée actuellement par la plupart des outils SIG. Elle est le plus souvent sollicitée pour le stockage des primitives géométriques dans les BDG. Une entité surfacique qui est représentée par son contour peut avoir autant de points que l'opérateur veut<sup>12</sup> saisir, sans toutefois changer sa forme (point intermédiaire se situant de manière colinéaire).

<sup>11</sup>  $\mathcal{F}_0$  est l'ensemble des surfaces de  $\mathcal{F}$ ,  $\mathcal{F}'$  est l'ensemble des polygones,  $\mathcal{F}'_0$  est l'ensemble des ensembles finis de points de disjoints.

<sup>12</sup> "veut": Nous avons utilisé ce terme pour caricaturer la situation, bien que les opérateurs de saisie soient contraints de suivre les spécifications de saisie, ils tombent souvent dans le piège inévitable de la "sur-qualité" par la génération de l'information inutile.

A cet espace de représentation, nous pouvons associer toutes les mesures qui s'articulent sur la distance euclidienne, à savoir les distances de Hausdorff [Abbas 1994] et de Fréchet [Alt & al. 1993; Devogele 2000] dont nous donnons les définitions plus loin.

### II.3.1.2. Représentation angulaire

La fonction angulaire est initialement définie par [Arkin & al. 1991] comme une fonction qui décrit l'entité surfacique à travers les angles formés par les segments qui composent son contour avec une demi-droite horizontale orientée selon l'axe des abscisses. Notons par  $\mathcal{L}$  l'espace de ces fonctions définies sur l'intervalle  $[0,1]$  et évoluant dans  $\mathbb{R}$ .

*La fonction angulaire notée  $\theta(s)$ , est définie sur  $[0,1]$  dans  $\mathbb{R}$ . Elle renvoie la tangente, dans le sens trigonométrique, et est mesurée à partir d'un point de référence  $O$  situé sur le contour en fonction de l'abscisse curviligne  $s$  (cf. figure II-2). Ainsi,  $\theta(0)$  donne l'angle  $v$  formé par la tangente au point de référence  $O$  avec une direction de référence associée au polygone.  $s$  est une abscisse curviligne qu'on normalise par le périmètre du polygone. Les distances entre le point de référence et les sommets parcourus évoluent entre 0 et 1.*

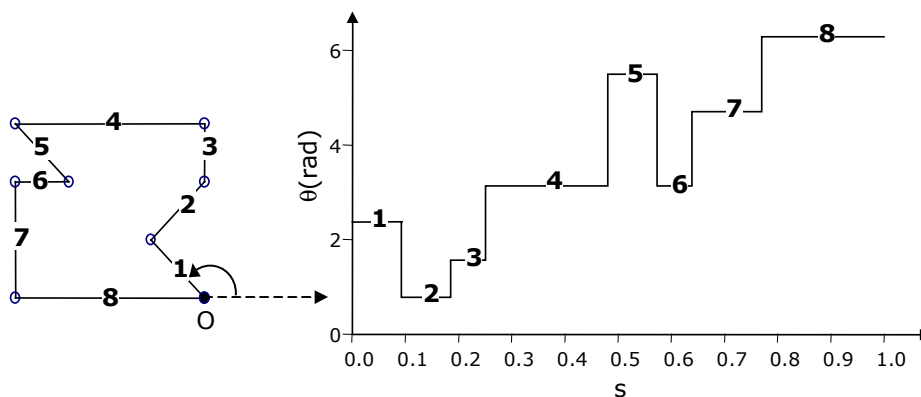


Figure II-2 : fonction angulaire d'un polygone simple

Cette modélisation respecte les critères cités dans le § II.2, à savoir :

- ✓ A un élément de l'espace  $\mathcal{F}$  des polygones peut correspondre plusieurs fonctions de l'espace  $\mathcal{L}$  des fonctions angulaires variant de  $[0,1]$  dans  $\mathbb{R}$ . Ceci est dû au choix du point "Origine des mesures". Cependant, la reconstruction du polygone à partir de sa fonction angulaire sera possible à condition de disposer des coordonnées du point de départ des mesures et de la longueur du contour du polygone. Ceci permettra, en quelque sorte, de répondre au critère de l'inversibilité (cf. §II.2.2). Cependant, dans le cadre de ce travail, nous ne cherchons pas à reconstruire le polygone à partir de sa fonction angulaire, mais plutôt à s'assurer que l'espace ( $\mathcal{F}$ ) des polygones et celui des fonctions angulaires ( $\mathcal{L}$ ) restent équipotents.

- ✓ La fonction angulaire est invariante par translation et par homothétie. La rotation du polygone d'un angle  $\alpha$  se traduit par l'ajout de la valeur de cet angle aux valeurs prises par la fonction angulaire.
- ✓ La façon de choisir le point de départ "origine des mesures" est aléatoire, donc à chaque choix du point correspond une fonction angulaire. Cependant, deux fonctions déterminées à partir de deux points différents peuvent être vues comme égales après la correction du déphasage. Le déphasage est égal à la distance normalisée entre les deux points origines des mesures.

Cette représentation n'est valable que pour les polygones simples. Les polygones à trous et les agrégats ne peuvent pas être représentés par la fonction angulaire.

### II.3.1.3. Représentation par les signatures polygonales

La modélisation d'un polygone par une fonction à distances radiales (que nous convenons d'appeler signature polygonale) consiste à mesurer les distances qui séparent le centre de masse du polygone aux points composant son contour (en le parcourant dans le sens trigonométrique direct), puis de les reporter sur une représentation graphique en fonction de l'abscisse curviligne normalisée par le périmètre [Bel Hadj Ali 1997].

La construction de la signature polygonale nécessite le choix d'un point de départ qu'on appellera "point origine des mesures". Notons par  $\mathcal{B}$  l'espace des signatures polygonales évoluant entre  $[0,1]$  sur  $\mathbb{R}^+$ .

*La signature polygonale notée  $SP$  du polygone  $P$  est définie comme suit:*

$$SP : [0,1] \rightarrow \mathbb{R}^+$$

$$s \rightarrow SP(s)$$

Avec  $SP(s) = \sqrt{(xc - x(s))^2 + (yc - y(s))^2}$  ;  $xc$  et  $yc$  désignent les coordonnées du centre de masse du polygone et  $x(s)$ ,  $y(s)$  les coordonnées du point courant du contour à l'abscisse curviligne  $s$ . On note également que la signature polygonale est une fonction périodique de période égale à l'unité.

La figure II-3 illustre un exemple de construction de la signature polygonale :

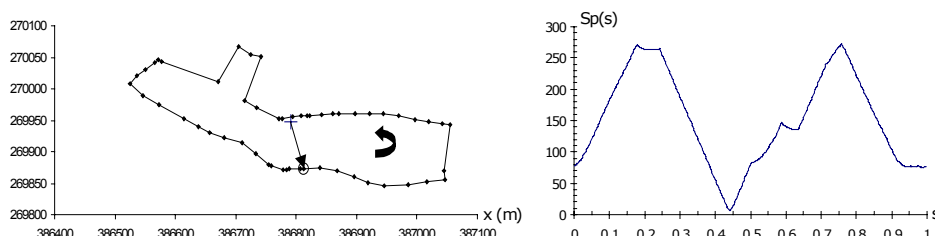


Figure II-3 : Signature polygonale d'un polygone représentant une parcelle de bois extrait de la couche de l'occupation du sol de la BDTopo © sur la région d'Angers

Le choix du pas d'échantillonnage utilisé pour segmenter le contour en des tronçons de distance égale est déterminant pour l'établissement de la signature polygonale. Ce choix doit respecter le théorème de Nyquist-Shannon dans la mesure où le pas minimal utilisable doit être inférieur ou égal à la demi-longueur du plus petit tronçon qui participe à la formation du contour du polygone [Bel Hadj Ali 2000]. Cependant, plus le pas d'échantillonnage est fin, plus la fidélité de représentation de la signature polygonale est élevée. Le choix du pas d'échantillonnage influe énormément sur la construction de la signature polygonale. Une quantification de cette influence est donnée dans [Bel Hadj Ali 1997] et illustrée par la figure II-4.

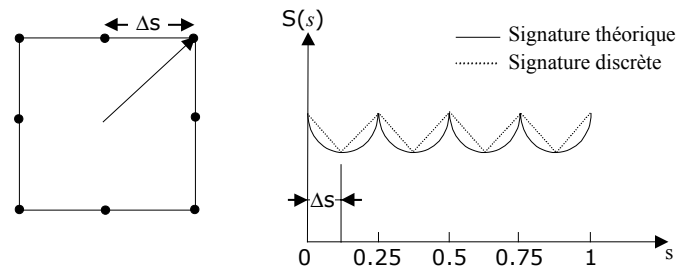


Figure II-4 : Influence du pas d'échantillonnage sur la construction de la signature du polygone

Cette représentation présente les propriétés suivantes :

- ✓ Selon le choix de point de départ "origine des mesures", on peut faire correspondre à un élément de l'espace  $\mathcal{F}$  des polygones plusieurs fonctions de l'espace  $\mathcal{B}$  des signatures polygonales variant de  $[0,1]$  sur  $\mathbb{R}^+$ . Cependant, la reconstruction du polygone à partir de sa signature polygonale sera possible à condition de disposer des coordonnées du point de départ des mesures ainsi que celles du centre de masse du polygone. Ceci permet en quelque sorte de répondre au critère de l'inversibilité (cf. §II.2.2.). Le choix du point de départ "origine des mesures" influe sur la construction de la signature polygonale. Ainsi, pour un polygone donné on peut avoir, par exemple, deux signatures polygonales qui sont intrinsèquement les mêmes avec la présence d'un déphasage égal à la distance qui sépare les deux points de départ, normalisée par le périmètre du polygone.
- ✓ La signature polygonale respecte quelques-unes de ces règles d'invariance (cf. §II.2.4.), à savoir, l'invariance par translation et par rotation. Par ailleurs, la transformation d'un polygone  $P$  par une homothétie "isotrope" de facteur  $k$  se traduit par la multiplication de toutes les valeurs de sa signature polygonale par le même facteur. Le choix du point de départ "origine des mesures" influe également sur la construction de la signature polygonale. La correction de ces effets est présentée plus loin (cf. §II.4), essentiellement lors de la définition des indicateurs de similarité ou des distances.
- ✓ Si le polygone est symétrique par rapport à l'un des axes de représentation, on voit la période de la signature polygonale se réduire de l'unité à 0.5. S'il est symétrique par rapport aux deux axes de représentation, la signature

polygonale aura une période de 0.25. Cette caractéristique permet de traduire une des propriétés géométriques du polygone qui est la symétrie de sa forme.

On note que ce type de représentation ne concerne que les éléments appartenant à  $\mathcal{F}$ .

On trouve dans la littérature d'autres approches semblables à celle que nous proposons, telle la représentation utilisant la transformation dite "*tangentiel axis transform*" proposée par [Edwards 1997].

### II.3.1.4. Représentation dans l'espace des fréquences -Descripteurs de Fourier-

La représentation dans l'espace des fréquences est souvent utilisée dans le domaine du traitement de signal et de la parole. Elle est également utilisée dans le domaine de la reconnaissance des formes et de la vision robotique [Ezer & al. 1994; Kauppinen & al. 1995; Tello 1995]. Cette représentation consiste à décomposer toute fonction périodique satisfaisant des critères de continuité en une somme infinie de fonctions harmoniques arithmétiquement croissantes et de fréquences multiples de la fréquence de la fonction représentée. Cette fréquence est appelée "la fondamentale".

Cette représentation est dédiée aux éléments de  $\mathcal{F}$ . Cependant, au lieu d'être définis dans  $\mathbb{R}^2$ , les points du contour seront définis dans l'espace des complexes ( $\mathbb{C}$ ). Notons donc  $\mathcal{F}_c$  l'espace de représentation des éléments de  $\mathcal{F}$  dont les points du contour sont définis dans l'espace des complexes.

Etant donné un élément  $\Gamma \in \mathcal{F}$  de longueur  $L$  fermé et orienté, identifié par un point de départ  $Z(0) = (x(0), y(0))$  et une abscisse curviligne  $s$  permettant de représenter le point courant  $Z(s) = (x(s), y(s))$ , deux approches peuvent être envisageables. Ces deux approches permettant d'associer un mode de représentation à cette courbe : dans un premier cas, une représentation bi-dimensionnelle dans le champ complexe et dans le deuxième cas une représentation mono-dimensionnelle en utilisant la signature polygonale. Nous ne retiendrons que la représentation dans l'espace complexe  $\mathcal{F}_c$  que nous détaillerons dans le paragraphe suivant.

#### Représentation dans le champ complexe :

*L'élément  $\Gamma \in \mathcal{F}$  peut être décrit par des descripteurs de Fourier sur la base d'une représentation dans le champ complexe des points qui le composent. Disposant des coordonnées  $(x(s), y(s))$  du point courant en fonction de son abscisse curviligne, on définit l'élément  $\Gamma_c \in \mathcal{F}_c$  par la séquence complexe  $u(s) = x(s) + jy(s)$ . Cette suite est périodique, de période égale à la longueur de la courbe et elle peut se représenter en série de Fourier. Cette série qu'on appelle souvent "les descripteurs de Fourier" est donnée par :*

$$a_n = \frac{1}{L} \int_0^L u(t) e^{-j(2\pi/L)nt} dt \quad \text{avec} \quad u(s) = \sum_{k=-\infty}^{k=+\infty} a_n e^{j(2\pi/L)nl} \quad [\text{II-2}]$$

Cette représentation en descripteurs de Fourier permet une reconstruction spatiale aisée de la courbe. Par ailleurs, en admettant que  $u(s)$  ne contienne pas de discontinuité, on peut espérer avoir une représentation compacte, c'est à dire, avoir plus rapidement des coefficients dont le module converge vers 0 quand  $n$  tend vers l'infini.

Soit pour un polygone donné défini par  $N$  sommets  $V_0, \dots, V_{m-1}$ , les descripteurs de Fourier sont donnés par :

$$a_n = \frac{1}{L \left( \frac{n2\pi}{L} \right)^2} \sum_{k=1}^{k=n} (b_{k-1} - b_k) e^{-jn(2\pi/L)k} \quad \text{avec } l_0 = 0 \quad \text{et } l_k = \sum_{i=1}^k |V_i - V_{i-1}| \quad \text{pour } K > 0$$

et  $b_k = \frac{V_{k+1} - V_k}{|V_{k+1} - V_k|}$  [II-3]

Se basant sur cette définition, l'établissement d'une métrique entre les descripteurs s'avère très coûteux, du fait d'une part que les segments du contour ne sont pas uniformes, et d'autre part que la séquence  $\{b_k\}$  est fortement dépendante du choix de point de départ "origine des mesures". Pour donner une réponse à ce problème, on se propose d'échantillonner le contour avec un pas régulier, ce qui permet de revoir la définition des descripteurs comme suit :

Soit une courbe  $C$ , définie par une séquence de points  $z(i)$  avec le point courant  $z(i) = x(i) + jy(i)$ , avec  $i = 0, \dots, N_B - 1$ , les descripteurs de Fourier discrets seront donnés par :

$$Z(k) = \sum_{n=0}^{N_B-1} z(n) e^{-j2\pi nk / N_B} = M(k) e^{j\theta(k)} \quad \text{[II-4]}$$

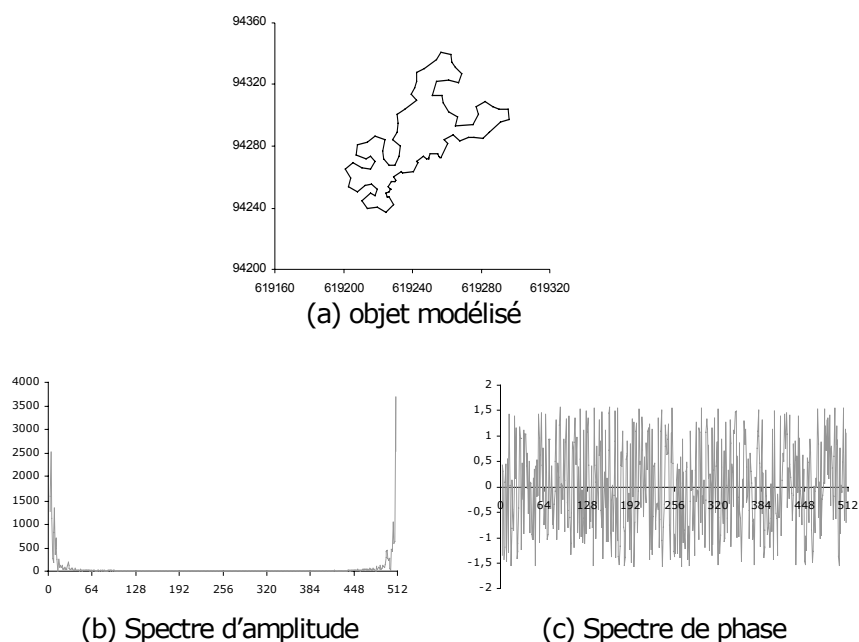
avec  $k = 0, \dots, N_B - 1$ .

Pour la mise en œuvre de cette représentation, on a utilisé la transformée de Fourier rapide (plus connue sous son appellation anglaise FFT pour *Fast Fourier Transform*). Pour cela on effectue une interpolation de la suite des points du contour du polygone pour se ramener à une suite codée sur  $2^p$  points (cf. figure II-5(b)).

Le choix du pas d'échantillonnage doit également répondre aux critères de Nyquist-Shannon, à savoir qu'il soit au plus égal à la demi longueur du plus petit segment du contour du polygone (même critère utilisé pour établir les signatures polygonale §II.3.1.3.). L'utilisation de ce critère suppose la définition d'un pas d'échantillonnage "personnalisé" pour chacun des polygones. Cependant, on peut s'affranchir de ce critère en utilisant un pas égal au tiers de l'erreur moyenne quadratique de la base de données à contrôler [Abbas 1994]. Le choix du pas d'échantillonnage conditionne donc l'interpolation du contour du polygone et, par la suite, le nombre de fréquences le représentant. Le facteur  $p$  ( $2^p$ ) sera alors conditionné par le rapport du périmètre du polygone sur le pas d'échantillonnage.

Les descripteurs associés aux indices proches du rang 0 (modulo  $N$ ) correspondent aux fréquences fondamentales associées aux variations globales du contour. La figure II-5 illustre également le module de la transformée des descripteurs de Fourier ordonnés

du rang  $(-N/2 + 1)$  au rang  $(N/2 + 1)$  et, par conséquent, centrés sur la fondamentale associée au rang 0.



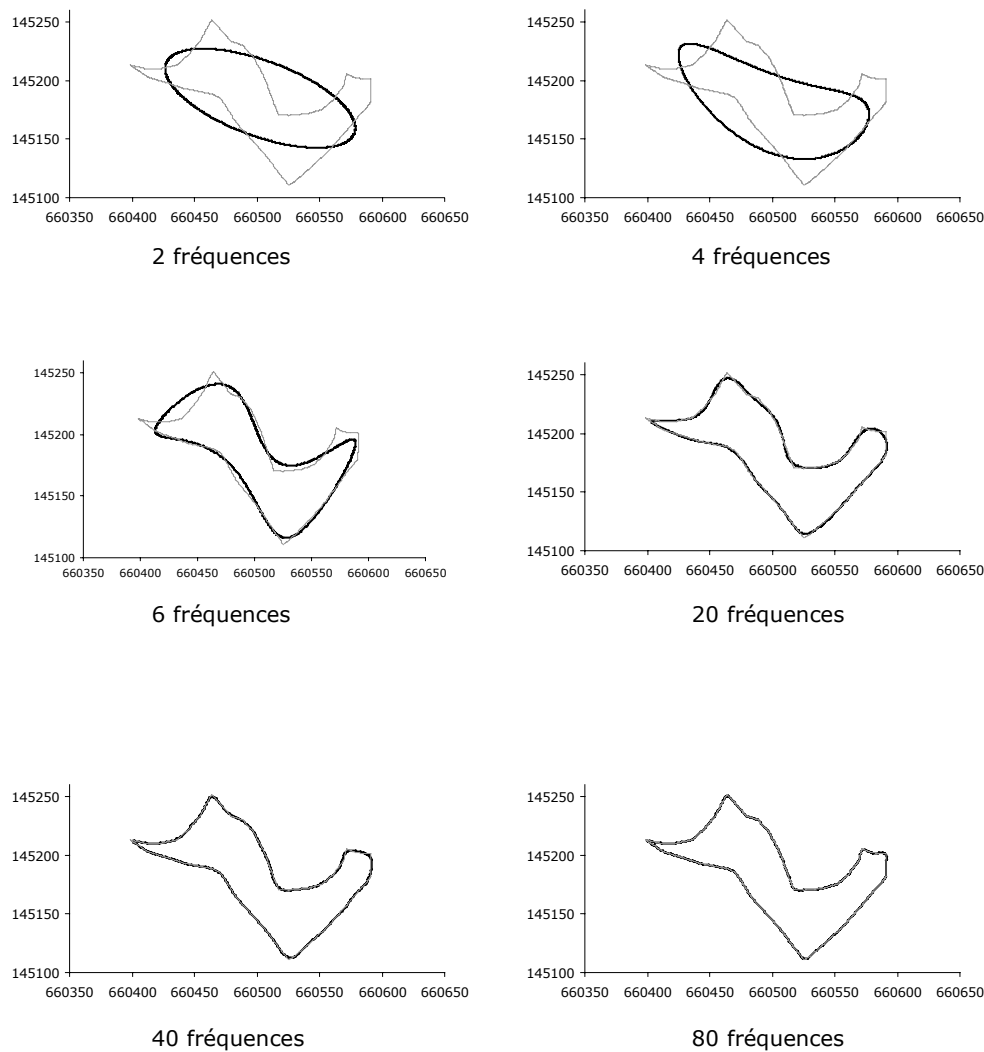
**Figure II-5 : Descripteurs de Fourier**

(a) Polygone test échantillonné en 512 points et représentation des modules des descripteurs de Fourier (b). Pour des raisons de clarté de la figure et pour faire apparaître les autres harmoniques, on a mis la fondamentale à zéro. (c) Spectre de phase des descripteurs de Fourier.

Par analogie avec le domaine du traitement du signal, on peut démontrer que les basses fréquences (celles qui sont les plus proches de la fondamentale) sont porteuses de l'information sur la forme générale de l'objet. En se limitant à l'utilisation de deux fréquences, la transformée inverse de Fourier permet la régénération de l'ellipse caractéristique de l'objet. Les hautes fréquences sont porteuses de bruit et des petites variations pouvant se trouver sur le contour. Cette analyse est mieux illustrée par la figure II-6, dans laquelle on a essayé de reconstruire le polygone en se servant de quelques-uns de ses descripteurs de Fourier.

La représentation par les descripteurs de Fourier peut être également réalisée en considérant l'entité surfacique comme une image binaire. Nous ne présentons pas cette méthode, puisque nous démontrons le lien existant entre cette représentation et celle de la représentation par les moments que nous exposerons plus loin (cf. §II.3.2.2.3.).





**Figure II-6 : Reconstruction d'un polygone à partir de ses descripteurs de Fourier.**

*Initialement le polygone est représenté par 512 descripteurs de Fourier. La reconstruction montre que l'utilisation de quelques-uns de ces descripteurs est largement suffisante pour représenter le polygone, cela prouve que la représentation est bien compacte. Dans cet exemple, on voit très bien que l'utilisation de 80 fréquences parmi 512 permet une reconstruction complète du polygone.*

Les modélisations présentées précédemment ne peuvent décrire que les objets simples appartenant à l'espace  $\mathcal{F}^1$ . Elles ne sont toutefois pas applicables dans les cas où l'on se trouve face à des configurations complexes d'objets surfaciques. Dans la section suivante §II.3.2, nous présentons de nouveaux espaces de représentation dont les techniques permettent la prise en compte des objets complexes et servent également à représenter les polygones simples.

## II.3.2. Pour les objets complexes (représentation par l'intérieur de l'entité surfacique)

Dans le paragraphe précédent, nous avons présenté des espaces de représentation pouvant modéliser les entités surfaciques d'une manière autre que celle du mode vecteur dans le plan cartésien. Par ailleurs, ces représentations ne peuvent servir que pour les entités élémentaires appartenant à  $\mathcal{F}^1$ , sans pour autant être utilisables pour les entités complexes. Dans les paragraphes suivants, nous présentons de nouveaux espaces de représentation pouvant traiter à la fois les entités simples et les entités complexes. Nous commençons par présenter, la représentation cartésienne en mode maillé, puisqu'elle va nous servir de base pour le reste des modélisations que nous introduirons par la suite.

### II.3.2.1. Représentation cartésienne (mode maillé)

La représentation cartésienne en mode maillé ou représentation en mode "raster"<sup>13</sup> est une représentation qui remonte aux premiers systèmes d'information géographique. Elle est essentiellement utilisée pour coder les images spatiales et les photographies aériennes. L'utilisation de ce mode de représentation oblige une "discrétisation" de l'espace cartésien sous la forme d'une matrice à pas régulier. Dans le contexte d'un contrôle de qualité des données géographiques, le choix du pas de maillage influe sur le résultat du contrôle. Cette question a été abordée par Abbas [Abbas 1994] en démontrant que le choix d'un pas de maillage égal au plus à  $1/3$  de l'erreur moyenne quadratique de la base de données traitée est optimal pour ne pas altérer les résultats des mesures. Nous adoptons ce résultat essentiellement pour le calcul de la distance de Hausdorff entre entités complexes. La représentation en mode maillé suppose un passage de l'espace  $\mathcal{F}^1$  à l'espace  $\mathcal{F}_0$ .

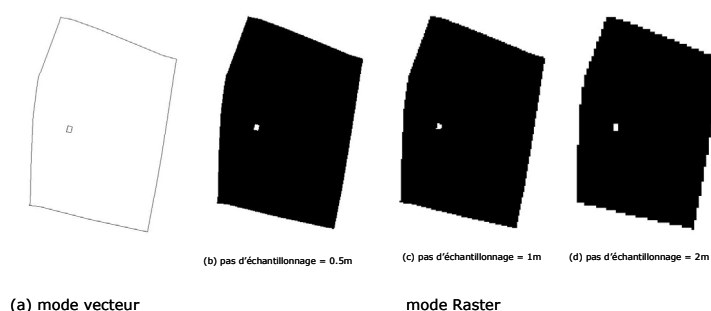


Figure II-7 : passage du mode vecteur au mode raster

La représentation en mode maillé respecte le critère du choix énuméré dans le §II.2, à savoir : l'unicité, puisque les espaces  $\mathcal{F}^1$  et  $\mathcal{F}_0$  sont équipotents. Cependant, l'inversibilité reste dépendante du choix du pas d'échantillonnage. Ce problème est largement abordé dans les techniques de vectorisation [Suan & al. 2000; Behzad & David 1985].

<sup>13</sup> Connue aussi sous l'appellation de représentation naturelle ou rétinienne [Milgram 1993]

En se basant sur cette représentation, nous proposons un changement de l'espace de représentation en introduisant une nouvelle modélisation fondée sur l'utilisation des moments mathématiques. Dans cette nouvelle représentation, l'entité surfacique sera représentée par un ensemble fini de valeurs de ses moments. Cette représentation pourrait être considérée comme une application entre l'espace  $\mathcal{F}_0$  dans  $\mathbb{R}^n$  ou bien dans  $\mathbb{C}^n$  selon le type de moments utilisés (cf. §II.3.2.2.).

### II.3.2.2. Représentation par les moments mathématiques

Les représentations des objets surfaciques par les moments présentent un large spectre d'utilisation dans le domaine d'analyse d'images, tels que la reconnaissance des formes [Milgram 1993; Adoram & Lew 1999; Pejnovic & al. 1992], la vision par ordinateur [Belkasim & al. 1990], la correction géométrique d'images [Dai & Jing 1999; Greenberg & al. 1996 ; Flusser & Suk 1994] ou pour la reconnaissance des signatures [Nassery & Faez 1996; Hai & Hai 1996], etc. Un ensemble de moments caractérisant une entité représente généralement les caractéristiques globales de la forme de l'entité et donne aussi, en général, une large indication sur la position de l'entité. Les techniques qui s'appuient sur les moments ont largement montré leur performance dans plusieurs applications liées notamment à la vision par ordinateur et à la robotique.

Les propriétés des moments d'une entité présentent une forte analogie avec ceux utilisés en statistique et en mécanique. Par exemple, les moments centrés géométriques d'ordre 0 et 2 d'une fonction de densité de probabilité représentent respectivement la probabilité totale et la variance; en mécanique, ces moments donnent la masse totale et les valeurs des moments d'inertie.

Donc, en considérant une entité comme une distribution bi-dimensionnelle d'intensité constante et en quantifiant l'espace de représentation sous la forme d'une matrice à pas régulier, les moments géométriques d'ordre 0 et 2 donnent d'une manière similaire l'aire totale de l'image et son orientation.

Les moments d'ordres 0 à 3 sont généralement utilisés pour représenter les détails prépondérants des objets composant l'image. Les moments d'ordre supérieur à 3 représentent les détails les plus fins. Ils sont fortement sensibles au bruit [Chassery & Montauvert 1991].

Dans cette section, nous présenterons les définitions relatives aux moments (géométriques et orthogonaux). Aussi, nous détaillerons la nature de l'information qui peut être dégagée des moments géométriques, aussi bien que leur signification physique pour décrire les formes des objets.

#### II.3.2.2.1. Définition générale des moments

*Une entité surfacique donnée pourrait être considérée comme une distribution bi-dimensionnelle de densité  $f$ .  $f$  représente l'intensité du pixel de coordonnées  $(x,y)$ . Soit  $\xi$  la région qu'occupe l'entité dans le plan cartésien, c'est à dire le support de la fonction  $f$  et qui*

représente son domaine de définition. La définition générale des moments  $\Phi_{pq}$  d'ordre  $n$  de la fonction  $f$  est donnée par :

$$\forall n \in \mathbb{N}, \forall p \in \mathbb{N}, \forall q \in \mathbb{N} / n = p+q;$$

$$\Phi_{pq} = \iint_{\xi} \Psi_{pq}(x, y) f(x, y) dx dy, \text{ avec } p, q = 0, 1, 2, \dots, \infty \quad [\text{II-5}]$$

$\Psi_{pq}$  est une fonction polynomiale continue sur  $\xi$  souvent appelée noyau, sur lequel s'appuie la définition des moments. Les indices  $p$  et  $q$  représentent les degrés des polynômes en  $x$  et en  $y$  au sein de la fonction  $\Psi$ .

Pour une entité donnée, la fonction d'intensité  $f$  est bornée et à support compact<sup>14</sup> sur  $\xi$ . Ainsi, l'intégrale donnée par l'équation II-5 aura des valeurs finies. Ceci amène également à dire que la "masse totale" de la distribution est positive. La masse totale de la distribution est donnée par :

$$|f| = \iint_{\xi} f(x, y) dx dy \quad [\text{II-6}]$$

Cependant, on peut rencontrer dans la littérature une variété de définitions des moments. La définition des moments donnée par l'Equation II-5 peut être exprimée autrement selon le système d'axes utilisé. Par exemple, l'utilisation des coordonnées polaires  $(r, \theta)$  nécessite de redéfinir les moments en fonction de la représentation polaire :

$$\Phi_{pq} = \iint_{\xi} r^{p+q+1} \Psi_{pq}^{r, \theta}(\theta) f(r, \theta) dr d\theta, \text{ avec } p, q = 0, 1, 2, \dots, \infty \quad [\text{II-7}]$$

### II.3.2.2. Représentation par les moments géométriques

La fonction noyau à partir de laquelle sont définis les moments géométriques s'appuie sur les coordonnées des pixels de l'image dans le plan cartésien. Le calcul des moments géométriques est le plus simple à mettre en œuvre comparativement aux moments définis par d'autres fonctions, en l'occurrence les fonctions complexes [Belkasim & al. 1996]. Ils sont parfois appelés moments cartésiens ou moments réguliers.

*La représentation d'une entité surfacique par un ensemble de valeurs de ses moments géométriques est définie comme une application de l'espace  $\mathcal{F}_0$  dans l'espace  $\mathbb{R}^\infty$*

*Les moments géométriques sont définis par l'ensemble des monômes  $\{x^p y^q\}$ ; le  $(n)^{\text{ième}}$  moment géométrique, noté  $m_{pq}$  est défini par :*

$$\{m_{pq}, p+q=n / m_{pq} = \iint_{\xi} x^p y^q f(x, y) dx dy, \text{ avec } p, q = 0, 1, 2, \dots, \infty \} \quad [\text{II-8}]$$

$\xi$  est la région sur laquelle la fonction de l'intensité de l'image  $f$  est définie. D'après l'équation II-8, les moments géométriques se présentent sous la forme d'une projection de la fonction  $f(x, y)$  sur le monôme  $x^p y^q$ . Il est à noter que la base complète composée par l'ensemble  $\{x^p y^q\}$  n'est pas orthogonale [Borowski & Borwein 1989], ce

<sup>14</sup> fonction numérique nulle en dehors d'un intervalle compact.

qui implique l'existence d'une redondance au niveau de l'information portée par chacun de ces moments. Cependant les moments géométriques :

**Existent** : La fonction d'intensité  $f$  est continue par morceaux et bornée sur son domaine de définition  $\xi$ . Donc, tous les moments  $m_{pq}$  existent et ont des valeurs finies.

**Et sont uniques** : La fonction d'intensité  $f$  est continue par morceaux et bornée dans son domaine de définition  $\xi$ . Donc, la séquence des moments  $\{m_{pq}\}$  est déterminée d'une manière unique (cf. §II.2.1.) par la fonction d'intensité  $f$  et inversement (cf. §II.2.2.). Cependant, l'inversibilité n'est pas évidente à mettre en œuvre, étant donné la complexité algorithmique qui lui incombe.

### II.3.2.2.3. Relation entre moments géométriques et la transformée de Fourier

La transformée de Fourier bi-dimensionnelle  $F$  de la fonction d'intensité  $f$  aussi appelée fonction caractéristique, est définie par :

$$\forall (u, v) \in \mathbb{R}^2, \forall A \in \mathcal{F}_0 \text{ défini par la fonction } f, \quad [II-9]$$

$$F(u, v) = \iint_{\xi} e^{i(ux+vy)} f(x, y) dx dy$$

$u$  et  $v$  représentent les coordonnées dans l'espace de fréquences. En développant le terme exponentiel en terme de séries et en utilisant la définition des moments géométriques, nous obtenons :

$$F(u, v) = \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \frac{i^{p+q}}{p!q!} u^p v^q m_{pq} \quad [II-10]$$

En dérivant la fonction caractéristique  $F$  au  $pq^{\text{ième}}$  ordre et en annulant  $u$  et  $v$ , il est facile de trouver :

$$\left[ \frac{\partial^p \partial^q F(u, v)}{\partial u^p \partial v^q} \right]_{u=v=0} = i^{(p+q)} m_{pq} \quad [II-11]$$

D'une manière analogue à la fonction caractéristique, la fonction génératrice des moments est définie par :

$$M(u, v) = \iint_{\xi} e^{(ux+vy)} f(x, y) dx dy \quad [II-12]$$

Et si on dérive  $M$  d'une manière similaire à  $F$ , on obtiendra :

$$\left[ \frac{\partial^p \partial^q M(u, v)}{\partial u^p \partial v^q} \right]_{u=v=0} = m_{pq} \quad [II-13]$$

Le développement en série du terme exponentiel de l'équation II-12 donne l'équation suivante :

$$M(u, v) = \sum_{p=1}^{\infty} \frac{1}{p} \sum_{r=0}^p C_r^p m_{p-r, r} u^{p-r} v^r \quad [II-14]$$

Les équations II-11 à II-14 montrent donc la relation qui lie les moments géométriques à la transformée de Fourier bi-dimensionnelle.

#### II.3.2.2.4. Implémentation et mode de calcul

On peut trouver dans la littérature quelques variations dans la définition des moments géométriques donnée par l'Equation II-8, et ceci, selon le domaine d'application. Ces définitions varient en fonction de leur mode d'implémentation, tout en s'accordant sur la même définition de base.

##### Calcul des moments par l'utilisation de l'intérieur :

Appelé aussi moments par la silhouette : ce sont les moments calculés à partir d'une image binaire. Dans ce cas, la fonction d'intensité ne peut avoir que deux valeurs 0 ou 1. Les pixels qui contribuent à la composition de l'objet auront pour intensité la valeur 1 et tous les autres auront pour intensité la valeur 0. Ainsi, les moments sont calculés par une double sommation sur les coordonnées du pixel de l'objet élevé à l'ordre du moment à calculer. Ce mode de calcul est détaillé dans le §II.3.2.2.6.

##### Calcul des moments par l'utilisation du contour :

Ces moments sont calculés en n'utilisant que les points du contour de l'objet. La forme d'un objet est souvent représentée par son contour, et par conséquent, seuls les points du contour contribueront au calcul de ces moments. Dans ce cas, la région  $\xi$  est réduite aux seuls pixels qui composent son contour. Une définition alternative est donnée par [Chen 1993 ] qui calcule les valeurs des moments d'une image binaire en utilisant les pixels composant le contour de l'objet. Par ailleurs, on propose une méthode pour le calcul des moments en utilisant les contours vectoriels des objets (annexe A). Elle ne peut être applicable que pour les éléments appartenant à  $\mathcal{F}^1$ .

Dans la suite de notre travail, nous utiliserons le mode d'implémentation des moments par l'intérieur. Le choix d'adopter ce mode de calcul est justifié par le fait qu'il est valable pour les éléments de  $\mathcal{F}_0$  et de  $\mathcal{F}^1$ . Par contre, le mode qui utilise uniquement les contours pour le calcul des moments, n'est toutefois pas utilisable pour les éléments de  $\mathcal{F}_0$ .

#### II.3.2.2.5. Signification physique des moments géométriques

Chacun des moments géométriques (selon son ordre) représente une caractéristique spatiale de la distribution de l'objet dans l'espace. Un ensemble de moments peut donc former un descripteur global de la forme de l'objet. L'interprétation physique de quelques moments est donnée dans cette section.

Par définition, le moment d'ordre 0 ( $m_{00}$ ) représente l'intensité totale de la fonction  $f$  représentant l'entité. Dans le cas d'une entité surfacique, ce moment donne la mesure de son aire.

Les moments de premier ordre  $m_{10}$  et  $m_{01}$  donnent une indication sur la répartition de la fonction  $f$ , respectivement, selon l'axe des  $x$  et selon l'axe des  $y$ . Par la suite, ils participent à la détermination du centre de masse de l'entité donné par :

$$x_0 = \frac{m_{10}}{m_{00}}; y_0 = \frac{m_{01}}{m_{00}} \quad [\text{II-15}]$$

Soit "y = c" l'axe de giration (de rayon c) du polygone autour de l'axe des x et parallèle à l'axe des abscisses. On peut donc écrire la relation suivante :

$$m_{02} = \iint_{\xi} c^2 f(x, y) dx dy \quad [\text{II-16}]$$

Et par la suite, on aura :

$$c = \sqrt{\frac{m_{02}}{m_{00}}} \quad [\text{II-17}]$$

d'une manière analogue, le rayon de giration du polygone autour de l'axe des ordonnées est donné par  $c = \sqrt{\frac{m_{20}}{m_{00}}}$

Afin de rendre les moments insensibles à la translation, il est judicieux d'utiliser les moments centraux. Ces moments sont définis comme suit :

$$\{\mu_{pq}, p+q=n / \mu_{pq} = \iint_{\xi} (x-x_0)^p (y-y_0)^q f(x, y) dx dy, \text{ avec } p, q = 0, 1, 2, \dots, \infty\} \quad [\text{II-18}]$$

En utilisant l'équation II-18, nous aurons :

$$\mu_{00} = m_{00}; \mu_{10} = \mu_{01} = 0 \quad [\text{II-19}]$$

Les moments centraux de second ordre sont aussi appelés moments d'inertie du polygone autour d'un système d'axe parallèle au système d'axe initial et qui passe par le centre de masse.

Les moments principaux d'inertie notés I<sub>1</sub>, I<sub>2</sub> peuvent être exprimés en fonction des moments centraux du second ordre de la manière suivante :

$$I_1 = \frac{(\mu_{20} + \mu_{02}) + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2};$$

$$I_2 = \frac{(\mu_{20} + \mu_{02}) - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{2} \quad [\text{II-20}]$$

On note que si  $\mu_{11} = 0$ , on a  $I_1 = \mu_{20}$  et  $I_2 = \mu_{02}$ .

L'angle d'orientation  $\theta$  que fait l'axe principal majeur d'inertie avec l'axe des abscisses est donné par l'équation suivante :

$$\mu_{11} \tan^2(\theta) + (\mu_{20} - \mu_{02}) \tan(\theta) - \mu_{11} = 0 \text{ et par la suite}$$

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \quad [\text{II-21}]$$

Les équations II-20 et II-21 peuvent être utilisées pour définir une ellipse caractéristique qui aura les mêmes moments d'inertie, ainsi que les mêmes directions des axes principaux du polygone d'origine. Le grand et le petit axe de l'ellipse sont donnés par les équations suivantes :

$$a = 2\sqrt{\left(\frac{I_1}{\mu_{00}}\right)}; b = 2\sqrt{\left(\frac{I_2}{\mu_{00}}\right)} \quad [\text{II-22}]$$

La figure II-8 illustre quelques-unes des caractéristiques citées précédemment.

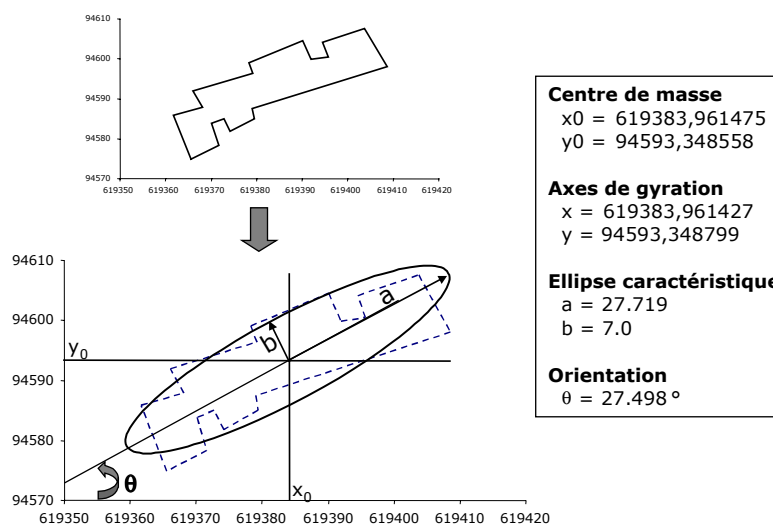


Figure II-8 : Moments géométriques et caractérisation d'un polygone

L'ellipse caractéristique donne une idée sur la forme fondamentale de l'objet. Les paramètres  $a$ ,  $b$  et  $\theta$  sont appelés les **descripteurs elliptiques** de la forme du polygone.

D'autres indicateurs peuvent être utilisés et ils découlent des équations précédentes tels que le terme  $(I_1 + I_2)/m_{00}^2$  donnant une indication sur l'étendue du polygone, ou le terme  $(I_2 - I_1)/(I_2 + I_1)$  donnant une idée sur l'élongation de la forme.

Les moments du troisième ordre  $\mu_{30}$  et  $\mu_{03}$  donnent respectivement une indication sur la symétrie selon l'axe des abscisses et selon l'axe des ordonnées. Si le polygone est symétrique autour de l'axe :  $x = x_0$ , alors la valeur de  $\mu_{30}$  est nulle. On peut, en effet, considérer  $\mu_{30}$  comme une mesure de symétrie autour de l'axe "moyen"  $x = x_0$ . Étant donnée que la valeur de  $\mu_{20}$  est toujours positive, on peut diviser le terme  $\mu_{30}$  par  $(\mu_{20})^{3/2}$  pour obtenir une quantité sans dimension. Cette quantité est appelée le "coefficient de symétrie" autour de l'axe des abscisses. Le coefficient  $\mu_{03}/(\mu_{02})^{3/2}$  peut être également calculé pour mesurer la symétrie autour de l'axe des ordonnées.

Nous venons de donner les définitions des moments géométriques, la signification physique de quelques-uns d'entre eux et ce qu'ils présentent comme contribution à la description des différentes caractéristiques de la forme d'une entité surfacique. Pour être utilisés comme descripteurs de forme, il est nécessaire que les moments géométriques soient invariants aux transformations que peut subir l'objet, telles que le changement d'échelle (homothétie<sup>15</sup>), la translation et la rotation (figure II-9).

<sup>15</sup> Pour la simplicité de calcul, nous assimilons un changement d'échelle à une homothétie. Cette hypothèse n'est pas toujours vérifiée, puisqu'on peut aboutir à un changement d'échelle sous l'effet d'une dilatation non isotrope.



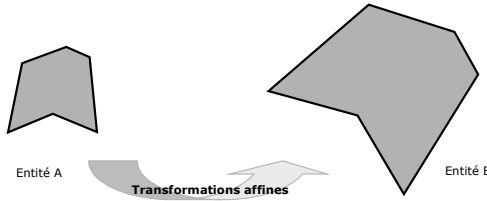


Figure II-9 : Invariance de forme sous l'effet des transformations affines

Nous allons chercher à transformer la définition des moments pour obtenir des moments invariants. Ceux-ci seront constants pour un objet et son image par les transformations précédemment citées. Pour rendre les moments géométriques invariants par les transformations affines, on se reporte à [Hu 1962] et à [Rothe & al. 1996]

Pour le calcul des moments, nous avons choisi la méthode passant par la quantification de l'entité surfacique par un maillage régulier. Avec cette méthode, le choix du pas d'échantillonnage ou de quantification paraît délicat puisqu'il conditionne les valeurs des moments, et, par la suite, le résultat de la qualification. Dans la section suivante, nous présenterons les problèmes liés au choix de ce pas et les valeurs optimales proposées.

### II.3.2.2.6. Problèmes liés au pas d'échantillonnage

Partant de l'équation II-8, définie dans un espace continu, pour calculer les moments géométriques d'une entité surfacique, il est nécessaire de "discrétiser" l'espace de représentation. Ceci revient à échantillonner la fonction  $f$  en une image  $f(x_i, y_j)$  de  $M$  lignes et  $N$  colonnes. L'intégrale double de l'équation II-8 peut donc s'écrire sous la forme d'une double sommation, tout en gardant à l'esprit que le couple  $(x_i, y_j)$  représente les coordonnées du centre du pixel au point  $P_{ij}$ . La formule de cette sommation s'écrit comme suit :

$$\tilde{M}_{pq} = \sum_{i=1}^M \sum_{j=1}^N x_i^p y_j^q f(x_i, y_j) \Delta x \Delta y \quad [\text{II-23}]$$

avec  $\Delta x = x_N - x_{N-1}$  et  $\Delta y = y_M - y_{M-1}$ , représentant respectivement les pas d'échantillonnage selon l'axe des abscisses et selon l'axe des ordonnées. Il est clair que la variation de l'une de ces deux valeurs (ou les deux ensemble) influe sur le résultat. Ceci est du essentiellement à la formulation de l'équation II-23 qui ne donne qu'une valeur approximative des moments. Cela résulte du fait qu'on néglige d'intégrer le produit de la fonction  $f$  et du noyau de la fonction "moments" à l'intérieur du pixel lui-même (cf. figure II-10). Par ailleurs cette formulation n'est pas la plus appropriée au calcul des moments notamment quand nous cherchons à évaluer les moments d'ordre élevé. Ceci nous amène à la remplacer par la formulation suivante, plus précise en terme d'approximation de valeurs des moments :

$$\hat{M}_{pq} = \sum_{i=1}^M \sum_{j=1}^N h_{pq}(x_i, y_j) f(x_i, y_j) \quad \text{avec} \quad [\text{II-24}]$$

$$h_{pq}(x_i, y_j) = \int_{x_i - \frac{\Delta x}{2}}^{x_i + \frac{\Delta x}{2}} \int_{y_j - \frac{\Delta y}{2}}^{y_j + \frac{\Delta y}{2}} x^p y^q dx dy \quad [\text{II-25}]$$

$h_{pq}(x_i, y_j)$  représente l'intégration du monôme  $x^p y^q$  à l'intérieur du pixel (i,j).

Par ailleurs, la relation entre  $\tilde{M}_{pq}$  et  $\hat{M}_{pq}$  est établie et donnée par l'équation suivante :

$$\hat{M}_{pq} = \tilde{M}_{pq} + \frac{q(q-1)}{24} (\Delta y)^2 \tilde{M}_{p,q-2} + \frac{p(p-1)}{24} (\Delta x)^2 \tilde{M}_{p-2,q} + O((\Delta x \Delta y)^2) \quad [\text{II-26}]$$

La démonstration de l'équation II-24 est donnée dans [Liao & Pawlak 1996].

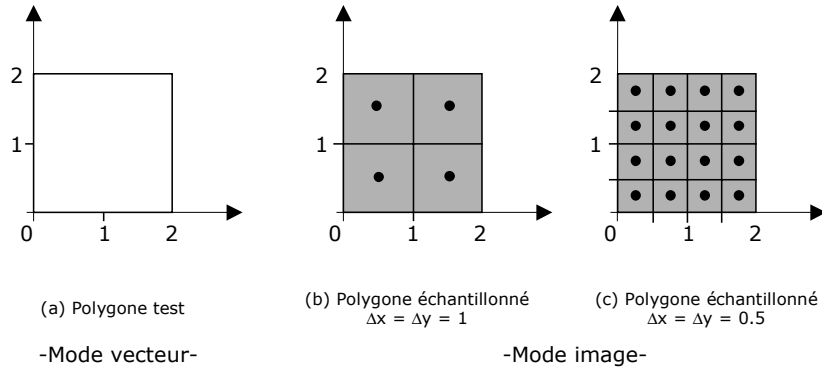


Figure II-10 : influence du pas d'échantillonnage sur le calcul des moments

Calculons à titre d'exemple le moment d'ordre 2,  $M_{20}$ , du carré (figure II-10(a)). La valeur exacte est donnée par l'utilisation de la double intégration (équation II-8) :

$$M_{20} = \int_0^2 \int_0^2 x^2 dx dy = \int_0^2 \left[ \frac{x^3}{3} \right]_0^2 dy = \frac{16}{3} \cong 5.33333$$

Le calcul de ce moment par l'utilisation de la formule approximative (équation II-23) renvoie respectivement pour les figures II-10(b) et II-10(c) les valeurs 5 et 5.25. Cependant, l'utilisation de l'équation II-26 corrige l'influence du choix du pas d'échantillonnage et renvoie la valeur exacte, c'est à dire, 5.333333 pour les deux pas d'échantillonnage (figures II-10(b) et II-10(c)).

L'exemple de la figure II-10 ne constitue qu'un cas "vraiment" particulier des polygones que nous pouvons rencontrer lors des traitements des jeux de données réels. Il est très rare de trouver des polygones dont les côtés sont parallèles aux axes des abscisses et des ordonnées. D'autre part, la division de la longueur ou la largeur du polygone par la dimension du pas d'échantillonnage ne donne souvent pas un nombre entier.

Pour s'affranchir de ces difficultés supplémentaires, des tests empiriques ont été engagés afin de déterminer le pas d'échantillonnage optimal à utiliser afin de se rapprocher le plus possible des valeurs exactes des moments. La démarche suivante a été adoptée :

1- Calculer les moments géométriques d'un polygone simple, en utilisant son contour, et ce, par l'utilisation du théorème de Green-Riemann (Annexe A, pour la

méthode de calcul). Cette méthode de calcul est considérée comme la plus précise pour le calcul des moments dont l'ordre n'est pas trop élevé [Mukundan & Ramakrishnan 1998].

2- Calculer les mêmes moments géométriques du polygone "discrétisé" avec divers pas d'échantillonnage, par l'utilisation de l'équation II-26.

3- Calculer les différences entre les valeurs des moments obtenues en (1) et celles obtenues en (2) et analyser les résultats pour le choix d'un pas d'échantillonnage optimal. Pour mesurer ces différences, nous avons utilisé un indicateur défini par Mukundan [Mukundan & Ramakrishnan 1998] comme le pourcentage moyen de l'erreur noté  $\varepsilon$ , qui est donné par :

$$\varepsilon = \frac{1}{6} \sum_{p=0}^2 \sum_{q=0}^p \frac{|m_{pq}^{\text{calculé}} - m_{pq}^{\text{exact}}| * 100}{m_{pq}^{\text{exact}}} \quad \text{[II-27]}$$

Plusieurs tests ont été réalisés en utilisant des polygones simples de la BDTopo® représentant des bâtiments [Delattre 2000]. Nous présentons dans la figure II-11 un exemple des résultats de la méthode que nous allons analyser par la suite. Quelques autres exemples sont présentés en annexe B, montrant la même tendance de l'évolution de l'erreur en fonction de l'accroissement de la taille du pas d'échantillonnage.

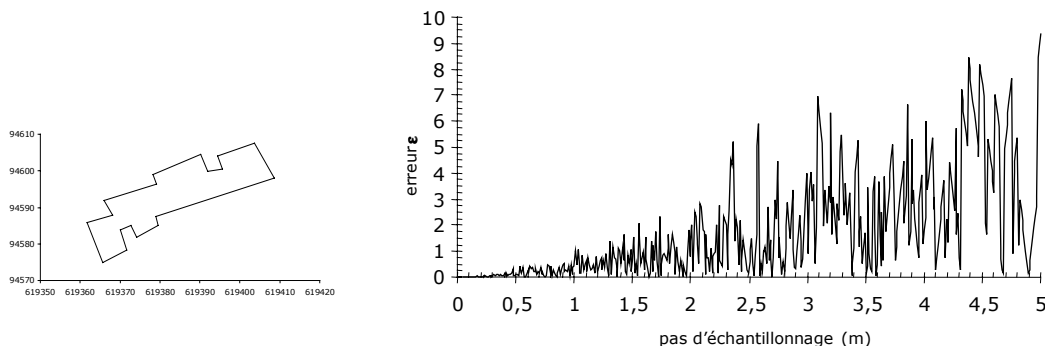


Figure II-11 : Différence entre moments exacts et moments calculés.

Nous remarquons, d'après la figure II-11, que les différences entre les valeurs exactes et les valeurs calculées des moments sont presque constantes quand les valeurs du pas d'échantillonnage sont inférieures à 0.5m : elles varient autour de la valeur de 0.2%. Au-delà de la valeur 0.5m, l'erreur commence à s'accroître d'une manière significative avec une forte fluctuation au moment où l'entité est sous-échantillonnée. Une analyse plus fine de cette tendance en utilisant la moyenne et la variance mobile (cf. figure II-12) renforce cette constatation, puisque nous pouvons remarquer que pour les valeurs élevées du pas d'échantillonnage la variance évolue avec la moyenne.

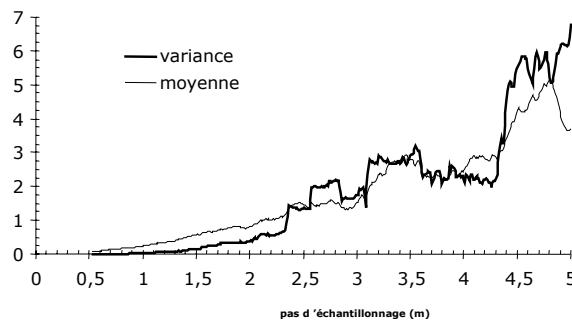


Figure II-12 : moyenne et variance mobile d'un échantillon mobile de 10 individus de l'erreur relative au pas d'échantillonnage

Dans la suite de ce travail et pour tous les tests que nous avons entrepris, la valeur 0.5m est utilisée comme la taille du pas optimal d'échantillonnage. Il est à noter que cette valeur n'est pas universelle, mais relative aux données de type "bâtiments" de la BDTopo® (d'autres exemples sont illustrés en Annexe B). Donc, pour toute autre base de données qui n'a pas les mêmes caractéristiques que la BDTopo®, le test précédent doit être reconduit pour définir le pas d'échantillonnage.

Les moments géométriques sont vus comme une projection de la fonction d'intensité  $f$  sur l'ensemble des monômes  $x^p y^q$ , qui ne forme pas une base orthogonale de représentation. Cette non-orthogonalité de la base rend l'information portée par les moments redondante et la reconstruction du polygone à partir de ses moments géométriques très difficile et très coûteuse. Par conséquent, il est difficile de savoir l'ordre optimal des moments pouvant donner une description totale de l'entité surfacique.

Vers les années 1980, [Teague 1980] suggère l'utilisation des polynômes orthogonaux comme noyau pour les fonctions moments, afin de pallier aux problèmes rencontrés avec les moments géométriques. Nous présentons dans les sections suivantes deux types de moments qui se basent sur l'utilisation des polynômes orthogonaux : les moments de Legendre et les moments de Zernike.

### II.3.2.2.7. Moments de Legendre

Les moments de Legendre sont définis à partir des polynômes du même nom. Ils sont définis dans le carré unité  $[-1,1] \times [-1,1]$ , ce qui oblige à normaliser l'objet dont on veut calculer ses moments.

Le polynôme de Legendre d'ordre  $n$  est donné par:

$$\forall x \in [-1,1], \forall n \in \mathbb{N}, P_n(x) = \frac{1}{2^n n!} \frac{d^n (x^2 - 1)^n}{dx^n} \quad [\text{II-28}]$$

Les polynômes de Legendre  $\{P_n(x)\}$  forment une base complète et orthogonale sur le domaine de définition  $[-1,1]$ :

$$\forall (x,y) \in [-1,1]^2, \forall (m,n) \in \mathbb{N}^2, \int_{-1}^1 P_m(x) P_n(y) dx dy = \frac{2}{2m+1} \delta_{mn} \quad [\text{II-29}]$$

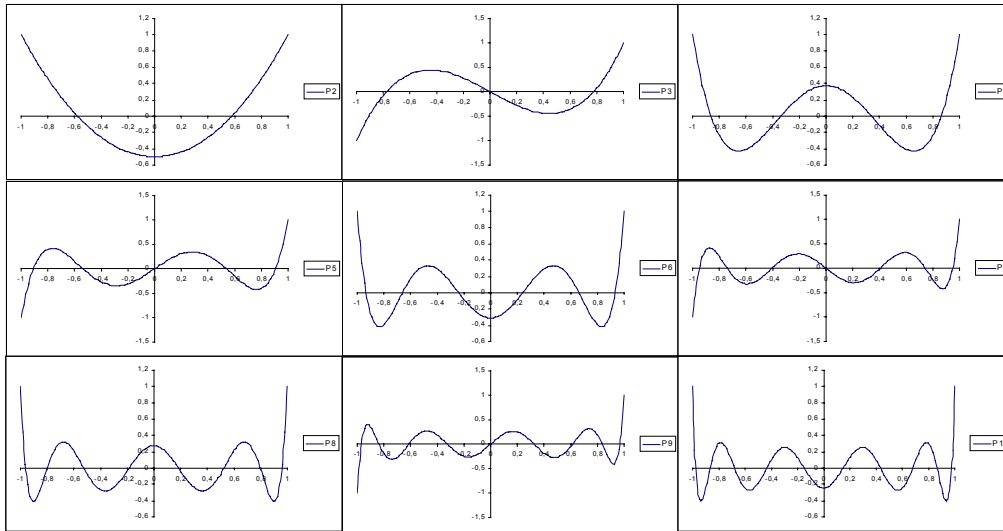
$\delta_{mn}$  représente la fonction de Kronecker.

Les moments de Legendre d'ordre  $N$  sont donc donnés par:

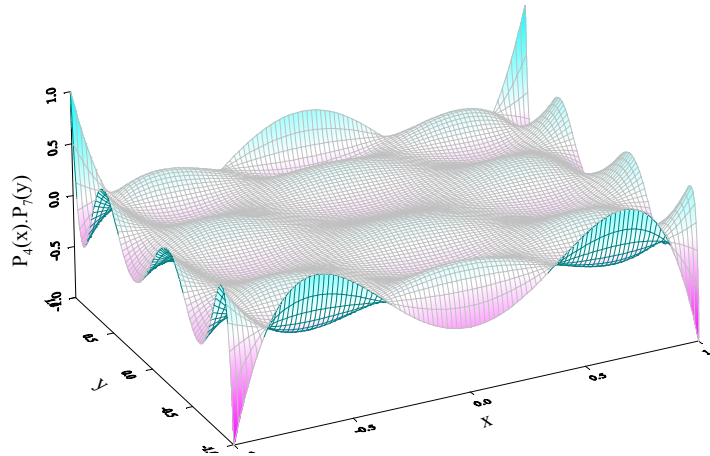
$$\forall (x,y) \in [-1,1]^2, \forall (p,q) \in \mathbb{N}^2 / N = p+q,$$

$$L_{pq} = \frac{(2p+1)(2q+1)}{4} \int_{-1}^1 \int_{-1}^1 P_p(x) P_q(y) f(x,y) dx dy \quad [\text{II-30}]$$

La figure II-13 illustre quelques-uns des polynômes de Legendre.



(a) Illustration des polynômes de Legendre à une variable, de l'ordre 2 à l'ordre 10



(b) Illustration du produit des deux polynômes de Legendre  $P_4(x).P_7(y)$

**Figure II-13 : Polynômes de Legendre**

Les moments de Legendre héritent de la propriété d'orthogonalité des polynômes de Legendre. Ainsi il n'existe plus de redondance de l'information véhiculée.

A partir de l'équation II-30, on peut générer une infinité de moments de Legendre. Plusieurs études sur la reconnaissance des formes ont démontré que l'utilisation des moments de Legendre de bas ordre (jusqu'à l'ordre 3) est suffisante pour représenter la forme globale de l'entité donnée [Shen & Shen 1996]. Cependant, notre problématique

n'est pas la même que celle de la reconnaissance des formes, puisque nous ne cherchons pas à identifier l'entité modélisée par rapport à des entités stockées dans une quelconque base. Notre objectif est de savoir quel est le nombre optimal de moments de Legendre pouvant représenter le maximum de caractéristiques géométriques de l'entité modélisée, permettant ainsi sa reconstruction totale.

### II.3.2.2.8. Recherche de l'ordre optimal des moments de Legendre

Pour chercher l'ordre optimal des moments de Legendre modélisant au mieux une entité surfacique, nous allons utiliser principalement la propriété d'orthogonalité des moments. Cette propriété permet la reconstruction d'une entité donnée à partir de l'ensemble de ses moments de Legendre [Parademetriou 1992]. Par l'utilisation du théorème de Fourier, nous pouvons approcher la fonction  $f$  par  $\hat{f}$  en fonction des moments de Legendre calculés jusqu'à un ordre donné  $N_{\max}$ , tout en forçant les autres moments d'ordre supérieur à  $N_{\max}$  à zéro. En utilisant le théorème de Fourier, la fonction reconstruite est donnée par :

$$\forall (x, y) \in \xi, \xi = [-1, 1]^2, f(x, y) = \iint_{\xi} L_{pq} P_p(x) P_q(y) dx dy \quad [\text{II-31}]$$

La version approximative de cette fonction par l'utilisation des moments de Legendre jusqu'à l'ordre  $N_{\max}$  est donnée de la manière suivante :

$$\hat{f}_{N_{\max}}(x, y) = \sum_{j=0}^{N_{\max}} \sum_{k=0}^j L_{j-k, k} P_{j-k}(x) P_k(y) \quad [\text{II-32}]$$

La méthode adoptée pour rechercher l'ordre optimal des moments ( $N_{\max}$ ) à utiliser, consiste à reconstruire l'entité à partir de ses moments jusqu'à un ordre donné et à mesurer l'erreur de reconstruction en comparant l'entité reconstruite à celle d'origine. Cette erreur peut être quantifiée selon [Teh & Chin 1988] par :

$$e^2(\hat{f}_{N_{\max}}, f) = \frac{\iint_{\xi} [f(x, y) - \hat{f}_{N_{\max}}(x, y)]^2 dx dy}{\iint_{\xi} [f(x, y)]^2 dx dy} \quad [\text{II-33}]$$

Il est à noter que chacun des moments de Legendre véhicule une information concernant la géométrie de l'entité modélisée. Nous pouvons isoler la contribution des moments d'ordre  $i$  dans le processus de reconstruction afin d'analyser la part d'information qu'ils représentent. La contribution des moments d'ordre  $i$  dans le processus de reconstruction peut être mesurée par la façon de voir à quel point la fonction  $f_i$  est proche de la fonction  $f$  comparée à  $f_{i-1}$ . Cette contribution, qu'on note  $C(i)$ , est donnée selon [Khotanzad & Hong 1990] par :

$$C(i) = e^2(\hat{f}_{i-1}, f) - e^2(\hat{f}_i, f) \quad [\text{II-34}]$$

Une valeur positive élevée de  $C(i)$  indique que les moments calculés jusqu'à l'ordre  $i$  véhiculent une information importante sur la géométrie de l'entité. Une faible valeur positive ou une valeur négative de  $C(i)$  indique que l'information portée par les moments correspondants est faible, voire non significative.

Plusieurs tests ont été réalisés pour reconstruire des entités surfaciques à partir de leurs moments de Legendre. Une analyse de ces reconstructions a permis de déterminer l'ordre optimal des moments de Legendre qui permet de mieux représenter l'entité surfacique. La figure II-14 illustre l'erreur de reconstruction (Eq. II-33), ainsi que la contribution  $C(i)$  (Eq. II-34) en fonction de l'ordre des moments d'une entité extraite de la BDTopo®.

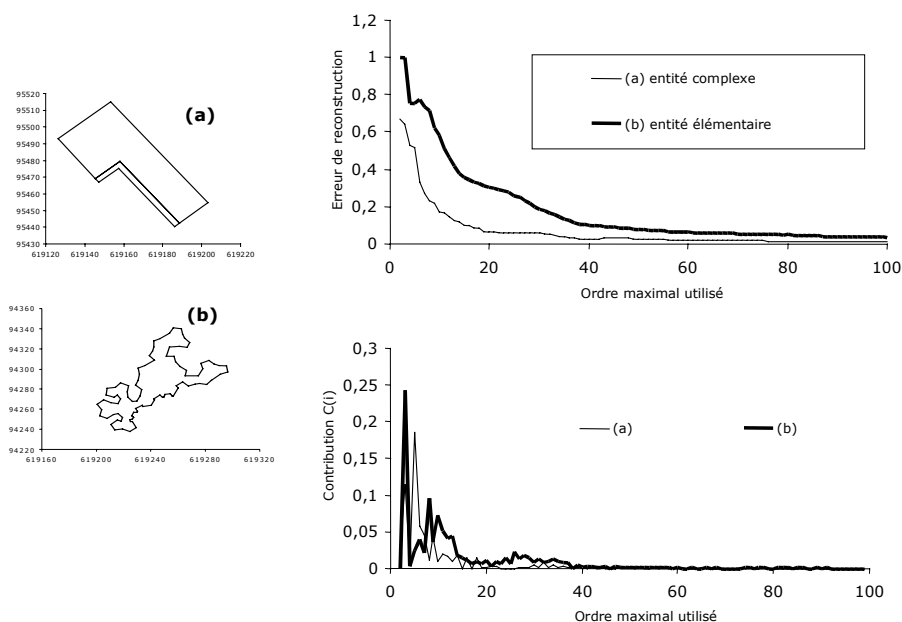


Figure II-14 : Erreur de reconstruction d'une entité surfacique à partir de ses moments de Legendre

D'après la figure II-14, on note que l'erreur de reconstruction commence à décroître doucement et à se stabiliser à partir des moments d'ordre 35 (soit un total de 666 moments). Cela nous permet de conclure qu'il faut utiliser des moments jusqu'à un ordre compris entre 35 et 40. Cette conclusion est confirmée par l'analyse des courbes de contribution qui montrent qu'au-delà de l'ordre 40, les valeurs de la contribution sont non significatives, et donc, les moments dont l'ordre est supérieur à 40 ne sont pas vraiment porteurs d'une information significative à propos de la géométrie de l'entité. Une confirmation visuelle de cette conclusion est illustrée par la figure II-15.

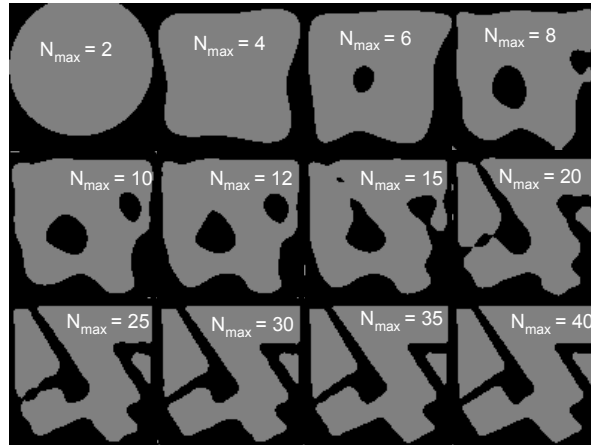


Figure II-15 : Reconstruction d'un agrégat de polygones par l'utilisation de ses moments de Legendre

On note que les moments de Legendre ne sont pas invariants par les transformations affines. Donc, leur utilisation à des fins de qualification de forme nécessite une normalisation afin de les rendre insensibles à ce genre de transformations. Cependant, le fait de normaliser l'entité surfacique dans le domaine  $[-1,1][[-1,1]$  annule l'effet d'une homothétie et d'une translation. L'annulation de l'effet de rotation sera traitée dans le § II.4.6. Ces propriétés répondent au critère du choix cité dans le §II.2.4.

D'une manière générale, les moments de Legendre sont définis sur l'espace  $\mathbb{R}^\infty$ . Cependant nous avons pu démontrer que l'utilisation d'un ensemble fini de moments de Legendre permet de reconstruire l'entité surfacique, répondant ainsi au critère de l'inversibilité (cf. §II.2.2.). De plus, nous pouvons même parler de bijection entre les espaces  $\mathcal{F}$  et l'espace des ensembles finis des valeurs des moments de Legendre.

La représentation des entités surfaciques permet également l'accès aux propriétés géométriques de l'entité. Leurs liens avec les moments géométriques sont donnés dans le §II.3.2.2.11. Donc, les propriétés géométriques pouvant être révélées par les moments géométriques peuvent l'être également par les moments de Legendre.

Dans le paragraphe suivant, nous présenterons une autre modélisation par un autre type de moments orthogonaux. Cette modélisation s'appuie sur une représentation polaire de l'entité modélisée et elle renvoie des valeurs dans l'espace des complexes. Ces moments utilisent les polynômes de Zernike [Teague 1980] comme noyau de leur fonction.

### II.3.2.2.9. Moments de Zernike

Les moments de Zernike sont définis à partir des polynômes de Zernike. Les polynômes de Zernike constituent une multitude de polynômes orthogonaux identifiés par un indice  $q$ , souvent appelé "répétition", et un ordre  $p$ .

*Pour tout  $0 \leq r \leq 1$ ,  $0 \leq \theta \leq 2\pi$ ,  $\forall p \in \mathbb{N}^+$ ,  $\forall q \in \mathbb{N} / p - |q|$  est pair et  $|q| < p$ , la valeur du polynôme de Zernike d'ordre  $p$  et de répétition  $q$  au point de coordonnées  $(r, \theta)$  notée  $V_{pq}(r, \theta)$ , elle est donnée par:*



$$V_{pq}(r, \theta) = R_{pq}(r)e^{iq\theta} \quad [\text{II-35}]$$

avec  $i = \sqrt{-1}$ , et

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s! \left(\frac{p+q-2s}{2}\right)! \left(\frac{p-q-2s}{2}\right)!} r^{p-2s} \quad [\text{II-36}]$$

avec  $p$  un entier positif et  $q$  un entier tel que  $p-|q|$  est pair et  $|q| < p$ .

Il existe  $\frac{(p+1)(p+2)}{2}$  polynômes linéairement indépendants de degré inférieur à  $p$ , et autant de moments d'ordre  $p$ .

Avec ces notations, les moments de Zernike sont donnés par :

$$Z_{pq} = \frac{p+1}{\pi} \int_0^1 \int_0^{2\pi} V_{pq}^*(r, \theta) f(r, \theta) r dr d\theta \quad [\text{II-37}]$$

L'originalité des moments de Zernike par rapport à ceux de Legendre tient à ce qu'ils sont invariants par rotation. Les amplitudes des moments de Zernike restent inchangées, si l'entité surfacique subit une quelconque rotation. Cependant, l'information concernant la rotation relative entre deux entités surfaciques peut être aisément extraite à partir de leurs moments de Zernike respectifs. Dans ce sens [Kim & Kim 1999] proposent une méthode originale pour estimer l'angle de rotation relatif entre deux entités afin de pallier l'ambiguïté de la méthode de détermination des angles par l'utilisation des moments géométriques. Nous exposons brièvement cette méthode que nous utilisons pour détecter les différences d'orientation entre les entités surfaciques.

### Détermination de l'angle de rotation relatif entre deux entités surfaciques

Soit une entité  $A \in \mathcal{F}$ , décrite par la fonction  $f(r, \theta)$  avec  $0 \leq r \leq 1$ ,  $0 \leq \theta \leq 2\pi$ , soit  $B \in \mathcal{F}$ , représentant l'image de l'entité  $A$  : par une rotation d'angle  $\alpha$  autour de son centre de masse, on notera  $f^\alpha(r, \theta)$ , la fonction décrivant l'entité  $B$  et qu'on peut écrire de la manière suivante :

$$f^\alpha(r, \theta) = f(r, \theta + \alpha) \quad [\text{II-38}]$$

Les moments de Zernike de l'entité  $B$  peuvent s'écrire en fonction de ceux de l'entité  $A$  de la manière suivante :

$$Z_{pq}^\alpha = Z_{pq} \exp(iq\alpha) \quad [\text{II-39}]$$

A partir de l'équation II-38, on peut dire que la rotation d'une entité par un angle  $\alpha$  ( $0 \leq \alpha < 2\pi$ ) induit un décalage de phase  $\theta_{pq}$  pour tous les moments de Zernike d'ordre  $p$  et de répétition  $q$  de valeur  $\theta_{pq} = q\alpha$ . Donc, l'angle  $\alpha$  peut être estimé à chaque ordre  $p$  et répétition  $q$  de la manière suivante :

$$\hat{\alpha}_{pq} = \frac{\theta_{pq}}{q}, q \neq 0 \quad [\text{II-40}]$$

L'estimation de l'angle  $\alpha$  se fait par la détection du maxima de la fonction de densité de probabilité  $P(\hat{\alpha})$ . Cette fonction de densité est obtenue en sommant toutes les fonctions de densités de probabilité  $P(\hat{\alpha}|p, q)$  pondérées par les moments de Zernike

respectifs. La fonction de densité, à un ordre  $p$  et une répétition  $q$ , est déterminée par une convolution d'un peigne de Dirac à  $q$  impulsions (solution de l'équation II-39) avec une fonction gaussienne de moyenne nulle et d'écart type égal à  $\pi/4q$  (pour en savoir plus sur la méthode se reporter [Kim & Kim 1999]). La figure II-16 illustre graphiquement la méthode ainsi décrite.

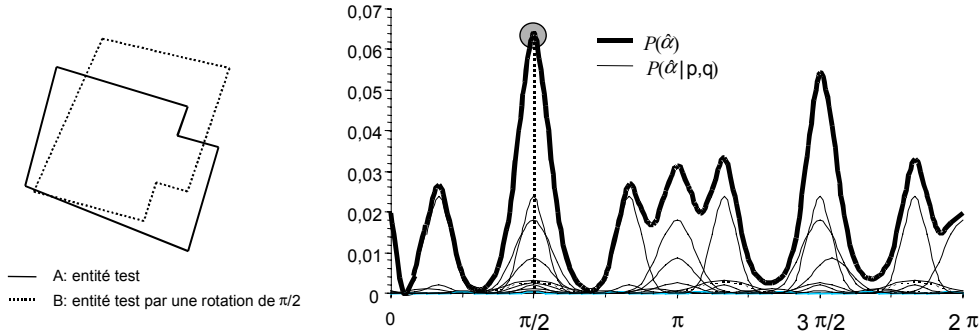


Figure II-16 : Estimation de l'angle de rotation relatif entre deux entités surfaciques par l'utilisation des moments de Zernike

En se basant sur la propriété d'orthogonalité des polynômes de Zernike et en utilisant le théorème de Fourier, nous pouvons reconstruire les entités surfaciques en utilisant leurs moments de Zernike. Etant donné que nous ne pouvons pas décrire une entité surfacique par un ensemble infini de moments de Zernike, nous limitons cette représentation à l'application reliant l'espace  $\mathcal{F}$  à l'espace  $\mathbb{R}^{(P_{\max}+1)(P_{\max}+2)/2}$ . Comme nous l'avons déjà fait pour les moments de Legendre, nous utilisons la même méthode pour déterminer l'ordre optimal ( $P_{\max}$ ) des moments de Zernike pour lequel les dits moments sont porteurs du maximum d'informations concernant la géométrie des entités.

### II.3.2.2.10. Recherche de l'ordre optimal des moments de Zernike

Dans ce paragraphe, et à l'instar de ce que nous avons fait pour les moments de Legendre, nous essayons de donner une réponse à la question suivante : comment peut-on représenter une entité surfacique par un ensemble fini de moments de Zernike avec une approximation convenable ? En d'autres termes, quel est l'ensemble minimal des moments de Zernike qui peut donner une description correcte des caractéristiques géométriques de l'entité représentée ? Pour répondre à cette question, nous utilisons la méthode décrite en §II.3.2.2.8, en reconstruisant l'entité et en évaluant l'erreur de reconstruction.

La reconstruction d'une entité à partir de ses  $P_{\max}$  premiers moments de Zernike se fait de la manière suivante :

On suppose que tous les moments de Zernike  $Z_{pq}$  d'une entité  $f(x,y)$  sont connus jusqu'à un ordre  $P_{\max}$  et que tous les autres, dont l'ordre est supérieur à  $P_{\max}$ , sont nuls. La fonction  $f$  peut alors être estimée par la fonction  $\hat{f}$  définie de la manière suivante :

$$\hat{f}(x, y) = \sum_{p=0}^{p_{\max}} \sum_q Z_{pq} V_{pq}(r, \theta) \quad [\text{II-41}]$$

où  $q$  a les mêmes contraintes qu'à l'équation II-35.

Les tests réalisés, similaires à ceux réalisés sur la reconstruction à partir des moments de Legendre, montrent que l'utilisation des moments de Zernike jusqu'à l'ordre 20 ( $p=20$ ) est largement suffisante pour donner une indication satisfaisante sur la géométrie de l'entité modélisée. La figure II-17 représente un exemple de reconstruction d'une entité surfacique par ses moments de Zernike.

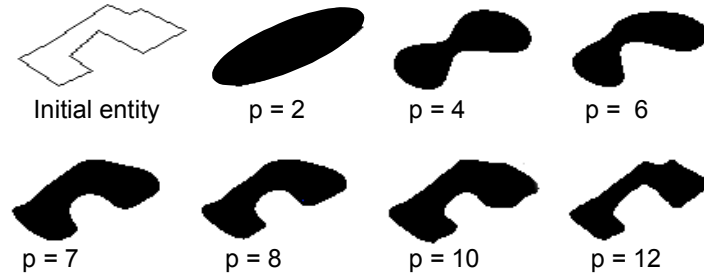


Figure II-17 : Reconstruction d'un polygone par ses moments de Zernike jusqu'à l'ordre 12.

### II.3.2.2.11. Relations entre moments

Bien que définis par des noyaux différents, les différents types de moments présentés peuvent être écrits les uns en fonction des autres.

Les moments de Legendre peuvent être exprimés en fonction des moments géométriques de la façon suivante :

$$L_{pq} = \frac{(2p+1)(2q+1)}{4} \sum_{i=0}^p \sum_{j=0}^q a_{pi} a_{qj} m_{ij} \quad [\text{II-42}]$$

avec  $a_{pi}$  représentent les coefficients du polynôme de Legendre d'ordre  $p$  à la valeur  $x^i$ . Il faut noter que cette relation n'est vraie que si les moments géométriques sont calculés en normalisant l'entité surfacique dans le domaine  $[-1,1][[-1,1]$ . L'équation II-42 montre que les moments de Legendre dépendent généralement des moments géométriques du même ordre ou bien d'ordre inférieur, et vice et versa.

Les moments de Zernike (définis par l'équation II-37) peuvent être écrits de la manière suivante:

$$Z_{pq} = \frac{p+1}{\pi} \sum_{k=q}^p B_{pqk} \int_0^{2\pi} \int_0^1 r^k e^{-iq\theta} f(r, \theta) r dr d\theta \quad [\text{II-43}]$$

$$\text{avec } B_{pqk} = \frac{(-1)^{(p-k)/2} \left(\frac{p+k}{2}\right)!}{\left(\frac{p-k}{2}\right)! \left(\frac{k+q}{2}\right)! \left(\frac{k-q}{2}\right)!} \quad [\text{II-44}]$$

L'équation II-43 peut être exprimée sous la forme cartésienne de la manière suivante :

$$Z_{pq} = \frac{p+1}{\pi} \sum_{k=q}^p B_{pqk} \int_x \int_y (x-iy)^p (x^2+y^2)^{(k-q)/2} f(x,y) dx dy \quad [\text{II-45}]$$

L'intégrale de l'équation II-45 est facilement développable sous la forme d'une série de moments géométriques. Ainsi, la relation entre les moments de Zernike et les moments géométriques est établie. Il faut noter au passage que cette relation n'est valide que si les moments géométriques sont définis en normalisant l'entité surfacique dans le cercle unité.

Il existe également une relation entre les moments de Zernike et ceux de Legendre, puisque la partie réelle des polynômes de Zernike (équation II-36) peut être exprimée de la manière suivante :

$$\begin{aligned} R_{pq}(1) &= 1, \\ R_{pp}(r) &= r^p, \\ R_{00}(r) &= 1, \\ \text{et } R_{2p,0}(r) &= P_p(2r^2 - 1) \end{aligned} \quad [\text{II-46}]$$

avec  $P_p$  représente le polynôme de Legendre d'ordre  $p$ .

### II.3.3. Synthèse sur la représentation par les moments

Nous partons du principe qu'une entité surfacique n'est pas simplement un contour délimitant une partie de l'espace, mais plutôt une partie de l'espace délimitée par un contour. L'intérêt majeur de ce principe est de permettre de discerner le cercle du disque, mais aussi de traiter les entités que nous avons qualifié de complexes (entités surfaciques composées d'agrégat de polygones et entités surfaciques avec trous).

La représentation des entités surfaciques par les techniques des moments répond aux critères énumérés dans le §II.2, à savoir :

L'unicité : il existe une bijection entre les espaces  $\mathcal{F}^1$  et  $\mathbb{R}^\infty$  pour les moments géométriques et les moments de Legendre et entre les espaces  $\mathcal{F}^1$  et  $\mathbb{C}^\infty$  pour les moments de Zernike. Cependant, les entités surfaciques ne sont représentées que par des ensembles finis de valeurs des moments. Donc, les espaces de représentation par les moments ont été réduits aux ensembles  $\mathbb{R}^{666}$  (pour les moments de Legendre) et  $\mathbb{R}^{231}$  (pour les moments de Zernike).

L'inversibilité est assurée pour tous les types de moments testés. Cependant, cette propriété est facile à démontrer pour les moments de Legendre et de Zernike puisqu'ils sont définis dans des bases orthogonales. L'inversibilité n'est pas évidente à mettre en œuvre pour les moments géométriques, à cause de la non orthogonalité des monômes utilisés comme noyau et à cause de la redondance de l'information portée par les valeurs des moments géométriques.

La propriété de l'invariance est respectée pour tous les types des moments utilisés. La définition des moments géométriques est modifiée en utilisant le théorème général des invariants défini initialement par [Hu 1962] afin de rendre ce type de moments

invariants par les transformations affines. Les moments de Legendre sont invariants par translation et changement d'échelle, grâce à la normalisation opérée dans le carré unité. L'invariance à la rotation doit être effectuée *a priori* par la détection de l'angle de rotation relatif entre les deux entités à mesurer. Les moments de Zernike sont invariants par translation et changement d'échelle grâce également à la normalisation dans le disque unité. L'invariance à la rotation est aussi assurée puisque, par définition, les normes des moments de Zernike sont invariantes par rotation.

Aussi, nous avons mis l'accent sur le fait que la géométrie des entités complexes sont mal exprimées en mode vecteur par une série de points et de tronçons de lignes. La représentation en mode vecteur s'articule essentiellement sur la notion du point, ce qui permet de ne fournir qu'une information localisée et isolée au niveau du point, sans traduire les caractéristiques géométriques de l'entité, ni décrire sa forme.

La recherche de nouvelles modélisations s'est donc avérée nécessaire afin de résoudre les problèmes soulevés dans ce constat. Les modélisations présentées dans ce chapitre, et notamment celles qui s'appuient sur les moments mathématiques, montrent très bien leur capacité à traduire avec fidélité les caractéristiques géométriques des entités surfaciques.

Dans les paragraphes suivants, nous allons établir des distances et des indicateurs entre les différentes représentations précédentes afin de doter les espaces de représentation correspondants de métriques adaptées à la comparaison de formes des polygones.

## **II.4. METRIQUES ET DISTANCES**

Cette section est consacrée à la présentation des métriques et d'indicateurs que nous avons définie et utilisée pour qualifier la position et la forme des entités surfaciques issues de deux bases de données géographiques différentes. Nous faisons remarquer à ce niveau, que l'établissement d'une distance – au sens mathématique du terme – n'est pas toujours trivial et que nous contournons cette difficulté par la mise en place d'indicateurs de similarité, lorsqu'il n'est pas possible de définir et d'utiliser des distances.

Cependant, il est préférable de définir des distances, puisque la propriété de l'inégalité triangulaire permet de définir une échelle de mesures dans l'espace de représentation.

Nous commençons, dans un premier temps, par un rappel mathématique de quelques définitions, avant de passer à l'établissement des métriques proprement dites.

## II.4.1. Rappels mathématiques

### II.4.1.1. Distance ou métrique

Soit  $\Delta$  l'espace des représentations. Une fonction  $f : \Delta \times \Delta \rightarrow \mathbb{R}^+$  est une distance sur l'espace  $\Delta$  entre deux entités  $\delta_A$  et  $\delta_B$  si et seulement si elle respecte les propriétés suivantes :

- (i) La propriété de symétrie :  $\forall (\delta_A, \delta_B) \in \Delta^2, f(\delta_A, \delta_B) = f(\delta_B, \delta_A)$
- (ii) La propriété de positivité :  $\forall (\delta_A, \delta_B) \in \Delta^2, f(\delta_A, \delta_B) \geq 0$
- (iii) L'inégalité triangulaire :  $\forall (\delta_A, \delta_B, \delta_C) \in \Delta^3, f(\delta_A, \delta_B) \leq f(\delta_A, \delta_C) + f(\delta_C, \delta_B)$
- (iv) La définition :  $\forall (\delta_A, \delta_B) \in \Delta^2, f(\delta_A, \delta_B) = 0 \Leftrightarrow \delta_A = \delta_B$

L'espace  $\Delta$  muni de la métrique  $f$  est appelé espace métrique.

Si la fonction ne remplit que les deux conditions (i) et (iii), elle sera désormais appelée une semi-distance.

La distance la plus utilisée, par exemple, dans le plan cartésien, est la distance euclidienne définie de la manière suivante :

Soient  $P_1$  et  $P_2$  deux points définis par leurs coordonnées respectives  $(x_1, y_1)$  et  $(x_2, y_2)$  dans le plan cartésien, la distance euclidienne entre ces deux points est donc donnée par :

$$d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad [\text{II-47}]$$

Cette distance est souvent considérée comme la distance "naturelle"

Il existe d'autres distances dans le plan cartésien, des exemples seront donnés plus loin.

### II.4.1.2. Indicateur de similarité normalisé

Soit  $\Delta$  l'espace des représentations. Une fonction  $f : \Delta \times \Delta \rightarrow [0, 1]$  est dite indice de similarité normalisé sur l'espace  $\Delta$  entre deux entités  $\delta_A$  et  $\delta_B$  si et seulement si elle respecte les propriétés suivantes [Chavent 1992] :

- La propriété de symétrie :  $f(\delta_A, \delta_B) = f(\delta_B, \delta_A)$
- La propriété de positivité :  $f(\delta_A, \delta_B) \geq 0$
- La propriété de normalisation :  $f(\delta_A, \delta_B) = 1 \Leftrightarrow \delta_A = \delta_B$

## II.4.2. Distances associées à l'espace cartésien

### II.4.2.1. Distance de Hausdorff

La distance de Hausdorff est une définition classique et déjà ancienne [Hausdorff 1937], que [Abbas 1994] a repris pour en faire l'axe central de ses travaux de thèse. La distance de Hausdorff est énoncée comme suit :

Soient deux éléments  $C_1$  et  $C_2$  de l'espace  $\mathcal{F}$ . On appelle distance de Hausdorff des deux éléments  $C_1$  et  $C_2$  la valeur maximale des deux quantités suivantes : la première est le maximum des plus courtes distances euclidiennes des éléments de  $C_1$  à l'ensemble des éléments de  $C_2$ , et la seconde est le maximum des plus courtes distances euclidiennes de l'ensemble des éléments de  $C_2$  à l'ensemble des éléments  $C_1$ .

La définition mathématique de cette distance est la suivante :

$\forall (C_1, C_2) \in \mathcal{F}^2$  avec  $C_1 = \{P_i\}$  et  $C_2 = \{P_j\}$

$$d_{12} = \sup_{P_i \in C_1} \left( \inf_{P_j \in C_2} d(P_i, P_j) \right), \quad d_{21} = \sup_{P_j \in C_2} \left( \inf_{P_i \in C_1} d(P_j, P_i) \right), \quad d_H = \max(d_{12}, d_{21}) \quad [\text{II-48}]$$

$d(P_i, P_j)$  désigne la distance euclidienne des points  $P_i$  et  $P_j$ .  $d_{12}$  et  $d_{21}$  sont souvent appelées les composantes de la distance de Hausdorff.

L'illustration graphique de cette distance est donnée par la figure II-18.

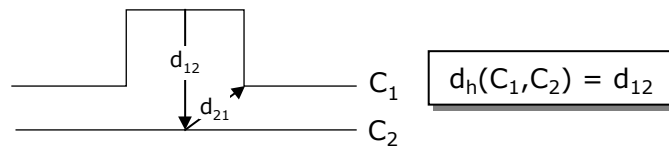


Figure II-18 : distance de Hausdorff

Telle qu'elle est utilisée par Abbas la distance de Hausdorff utilise la distance euclidienne pour mesurer les écarts entre deux éléments de  $\mathcal{F}$ , mais rien n'empêche de définir une autre distance de Hausdorff en utilisant la distance de Manhattan<sup>16</sup> (aussi appelée distance maximale). La distance de Hausdorff remplit les conditions d'une distance au sens mathématique du terme (cf. § II.4.1.1.), à savoir la symétrie, la positivité et l'inégalité triangulaire.

A partir des deux composantes de la distance de Hausdorff, Abbas a également défini un indice de généralisation comme suit :

$$i = 2 \frac{d_{21} - d_{12}}{d_{21} + d_{12}}$$

prenant 1 pour indice du jeu de données et 2 pour indice du jeu de référence.

<sup>16</sup> On rencontre souvent cette terminologie dans la littérature anglaise, pour désigner ce qu'on appelle en français la distance maximale. Ce terme vient en fait de l'architecture des ruelles de la ville de Manhattan qui font un angle droit entre-elles. Donc, par opposition à la distance euclidienne qu'on pourrait se qualifier comme une distance en vol d'oiseau entre deux points, la distance de Manhattan est la distance entre ces deux points en suivant les ruelles.

Si l'indice de généralisation est voisin de 2, on peut déduire qu'il y a l'oubli d'un détail pertinent par rapport à la référence. Dans le cas inverse (si  $i$  tend vers -2), on peut déduire l'existence d'une sur-information par rapport à la référence. Cet indice peut être généralisé pour l'ensemble d'un jeu de données en moyennant tous les indices de généralisation de tous les objets le composant.

Par l'utilisation de la distance de Hausdorff, on a pu montrer à l'aide de ces deux composantes qu'un bon modèle (la carte, par exemple) est plus proche de la réalité, mais que la réalité reste éloignée du modèle, car plus complexe.

Pour la mise en pratique du calcul de la distance de Hausdorff, [Abbas 1994] a développé dans le cadre de sa thèse une méthode de calcul utilisant l'algorithme de la boule à rayon variable. Il a proposé également une autre méthode de calcul qui se base sur l'échantillonnage du contour. Cependant, cette méthode n'est valide que pour les éléments de  $\mathcal{F}^1$ , puisqu'elle ne peut pas traiter les polygones à trous et les agrégats de polygones (entités complexes) éléments de  $\mathcal{F}_0$ . En effet, dans le cadre de notre travail, nous gardons la même définition de la distance de Hausdorff, mais en considérant les entités surfaciques comme étant des images binaires (elles prennent la valeur 1 si le pixel est à l'intérieur de l'entité et 0 si le pixel est à l'extérieur).

On trouve également dans [Le Men 1994] la déclinaison de cette méthode pour obtenir les écarts de géométrie entre deux éléments de  $\mathcal{F}$  en fonction de l'abscisse curviligne. Pour cela, il suffit de mémoriser les calculs d'écarts à partir de chaque point courant de la polyligne :

*On parcourt chaque polyligne et on cherche au point courant, le point le plus proche de l'autre polyligne. Réciproquement, on procède de la même manière en intervertissant les deux polygones. On obtient donc un ensemble de mesures d'écart de géométrie entre les deux polygones dans un sens ou dans l'autre (cf. figure II-19).*

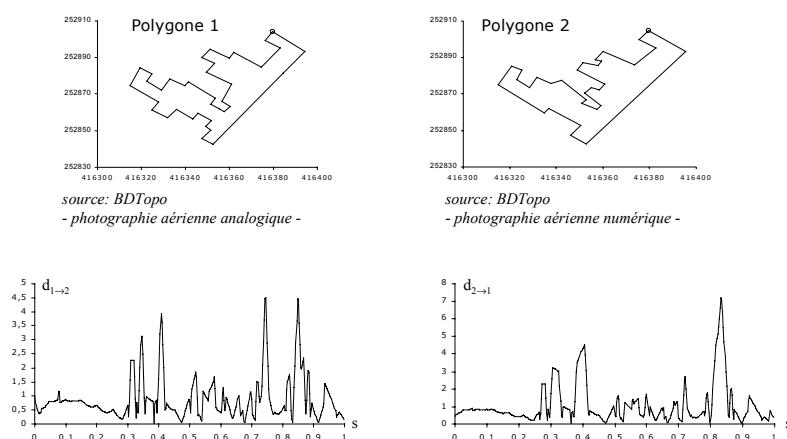


Figure II-19 : Variation des deux composantes de la distance de Hausdorff en fonction de l'abscisse curviligne

La variation des composantes de la distance de Hausdorff a été retenue comme un outil principal pour examiner en détail les incertitudes géométriques des différentes parties des polygones considérées [Vauglin 1997]. Ces courbes de variations peuvent être également considérées comme un outil robuste, par le fait que deux couples



d'entités proches ont des fonctions de variations proches l'une de l'autre (au sens de la norme infinie, par exemple). Les petites variations rencontrées sur les courbes (cf. figure II-19, pour  $s < 0.2$ ) sont dues essentiellement aux effets de l'échantillonnage. L'évaluation de la distance de Hausdorff ne se fait pas d'une manière exacte et reste fortement dépendante du choix du pas d'échantillonnage. [Alt & al. 1993] proposent une méthode exacte d'évaluation de la distance de Hausdorff fondée sur la propriété fondamentale des diagrammes de Voronoï généralisés<sup>17</sup> [Okabe & al. 1992]. Soient P et Q deux polygones de  $\mathcal{F}^1$  composés respectivement de p et q points et soit Vor(P) le diagramme de Voronoï généralisé du polygone P. La composante de la distance de Hausdorff de Q à P,  $D_h(Q \rightarrow P)$  s'appuie soit sur un sommet de Q, soit sur un point d'intersection de Q avec Vor(P), et réciproquement. La figure II-20 illustre un exemple.

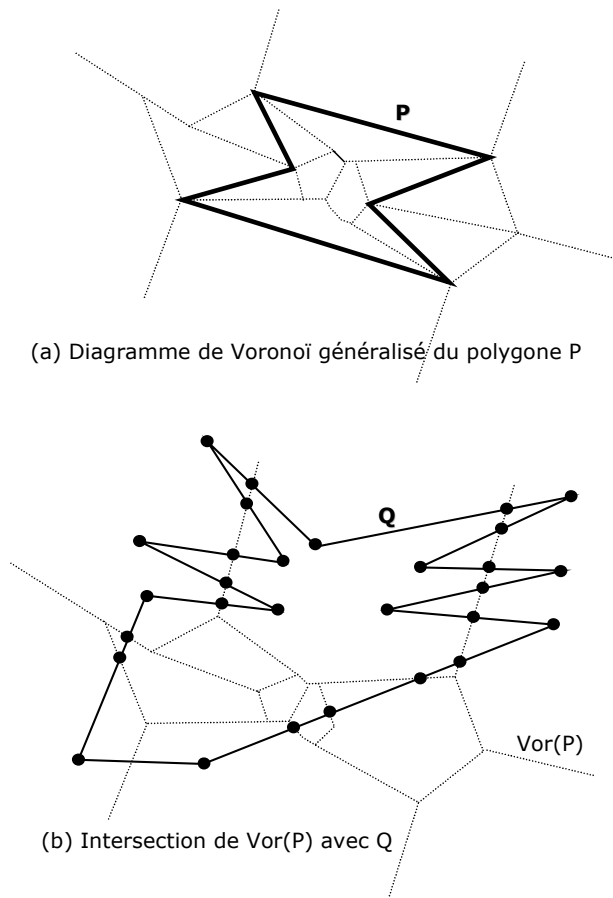


Figure II-20 : Calcul de la distance de Hausdorff par l'utilisation des diagrammes de Voronoï

Pour les méthodes de calcul, présentées ci-dessus, nous remarquons que la distance de Hausdorff est atteinte sur l'un des points du contour. Or, comme nous l'avons introduit, une entité surfacique n'est pas simplement un contour. De plus, ces algorithmes présentent une limitation : ils ne peuvent pas être utilisés pour évaluer la distance de Hausdorff entre deux entités surfaciques, appartenant à l'espace  $\mathcal{F}_0$ . L'exemple de la figure II-21 illustre cette limitation.

<sup>17</sup> Pour en savoir plus sur leur définition et les méthodes de leur construction, il faut se reporter aux travaux de J.F. Hangouët [Hangouët, 1998]

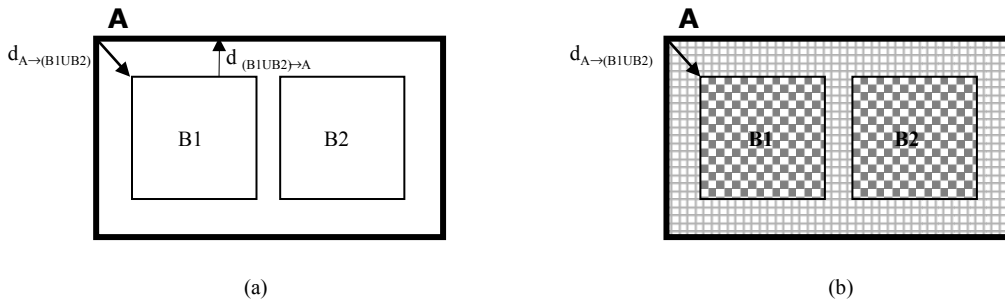


Figure II-21 : Distance de Hausdorff entre entités surfaciques

En s'appuyant sur les contours pour calculer la distance de Hausdorff (a), l'algorithme renvoie la valeur maximale des deux quantités  $d_{A \rightarrow (B1 \cup B2)}$  et  $d_{(B1 \cup B2) \rightarrow A}$ . Par ailleurs, si on tient compte de l'intérieur des entités surfaciques, on remarque clairement que la composante  $d_{(B1 \cup B2) \rightarrow A}$  s'annule et que la distance de Hausdorff sera égale à  $d_{A \rightarrow (B1 \cup B2)}$ .

Bien que la distance de Hausdorff utilise tous les points composant le contour de l'entité surfacique ou son intérieur, son résultat renvoie une valeur fondée sur une mesure locale. La distance de Hausdorff peut parfois donner une fausse indication sur la forme de l'entité (cf. figure II-22), puisqu'elle n'utilise que l'ensemble des points qui composent l'entité, sans pour autant tenir compte de ce qui se passe entre eux (parcours). La distance de Fréchet, que nous présentons en §II.4.2.2., semble être une bonne alternative pour remédier à ce genre de problème pour mesurer les écarts de formes entre les éléments de  $\mathcal{F}$ .

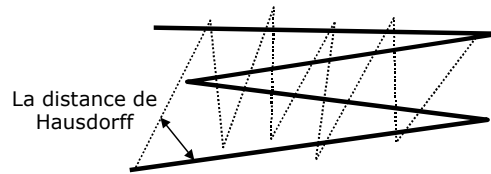


Figure II-22 : Distance de Hausdorff et mesure de forme

### II.4.2.2. Distance de Fréchet

La distance de Fréchet est définie comme une distance entre deux lignes orientées. Elle s'appuie sur la propriété suivante : toute polyligne orientée est équivalente à une application continue  $f : [a,b] \rightarrow V$  où  $a$  et  $b$  deux réels avec  $a < b$  et  $V$  l'espace vectoriel. La distance de Fréchet  $d_F$  est la suivante :

Soient  $f : [a,a'] \rightarrow V$  et  $g : [b,b'] \rightarrow V$  deux polygones et  $\|\cdot\|$  la norme usuelle,

$$d_F(f, g) = \inf_{\substack{\alpha: [0,1] \rightarrow [a,a'] \\ \beta: [0,1] \rightarrow [b,b']}} \max_{t \in [0,1]} \|f(\alpha(t)) - g(\beta(t))\| \quad \text{[II-49]}$$

Une illustration intuitive de la distance de Fréchet, donnée par [Alt & al. 1993], est la suivante :

Un maître et son chien suivent deux chemins : ils avancent ou s'arrêtent à volonté, indépendamment l'un de l'autre, mais ils ne peuvent pas revenir sur leur pas. La distance de

*Fréchet entre ces deux chemins est la longueur minimale de la laisse qui permet de réaliser un cheminement de concert satisfaisant ces conditions.*

Il s'agit bien en effet d'une distance (au sens mathématique), mais la démonstration de l'inégalité triangulaire n'est pas triviale. La distance de Fréchet a l'avantage d'être très proche d'une distance maximale "visuelle" [Devogele 2000]. Cependant, elle présente un inconvénient majeur, voire handicapant qui consiste au fait qu'elle est d'une complexité algorithmique élevée. Pour les éléments de  $\mathcal{F}$ , cette complexité est de l'ordre de  $O(N.M.\log^2(N.M))$  [Alt & al., 1993], avec N et M les nombres respectifs des segments dans chaque polyligne.

Tout en conservant les mêmes propriétés de la distance de Fréchet (équation II-49), [Eiter & Mannila 1994] propose une version discrète de cette distance qui a le mérite d'être simple à mettre œuvre avec une complexité algorithmique de l'ordre de  $O(n.m)$ , avec n et m les nombres respectifs des extrémités des segments des deux polygones. Cette méthode utilise les techniques de la programmation dynamique<sup>18</sup>, qui s'approche d'une méthode consistant à mettre en correspondance des contours extraits d'une paire d'images [Ohta & Kanade 1985].

La distance de Fréchet discrète peut être calculée de la manière suivante :

*soit  $(L_1, L_2)$  un couple de polygones chacune composées d'une suites ordonnées des extrémités des segments  $\langle L_{1.1} \dots L_{1.n} \rangle$  pour  $L_1$  et  $\langle L_{2.1} \dots L_{2.m} \rangle$  pour  $L_2$ . La distance de Fréchet discrète entre  $L_1$  et  $L_2$  ( $d_{Fd}(L_1, L_2)$ ) peut se calculer récursivement à partir de la formule suivante :*

$$d_{Fd}(L_1, L_2) = \max \left( \begin{array}{l} d_E(L_{1.n}, L_{2.m}) \\ \min \left( \begin{array}{l} d_{Fd}(\langle L_{1.1} \dots L_{1.n-1} \rangle, \langle L_{2.1} \dots L_{2.m} \rangle) \text{ si } (n \neq 1) \\ d_{Fd}(\langle L_{1.1} \dots L_{1.n} \rangle, \langle L_{2.1} \dots L_{2.m-1} \rangle) \text{ si } (m \neq 1) \\ d_{Fd}(\langle L_{1.1} \dots L_{1.n-1} \rangle, \langle L_{2.1} \dots L_{2.m-1} \rangle) \text{ si } (n \neq 1, m \neq 1) \end{array} \right) \end{array} \right)$$

*$\langle L_{1.1} \dots L_{1.n-1} \rangle$  et  $\langle L_{2.1} \dots L_{2.m-1} \rangle$  étant des polygones, il est possible d'appliquer récursivement la distance de Fréchet discrète avec comme paramètre l'une de ces deux lignes. La procédure récursive s'arrête quand les deux lignes sont réduites à deux points. La distance de Fréchet discrète est alors égale à la distance euclidienne.*

*Intuitivement, cette procédure se traduit de la manière suivante. Pour aller jusqu'au point  $L_{1.n}$  et  $L_{2.m}$ , la longueur de la laisse du chien doit permettre de relier  $L_{1.n}$  et  $L_{2.m}$ . Elle est donc supérieure ou égale à la distance entre  $L_{1.n}$  et  $L_{2.m}$ . Cette laisse doit aussi permettre d'aller en  $L_{1.n}$  et  $L_{2.m}$  en partant de  $L_{1.1}$  et  $L_{2.1}$ . Pour cela trois déplacements sont possibles :*

- Le chien reste en  $L_{2.m}$  et le maître passe de  $L_{1.n-1}$  à  $L_{1.n}$*
- Le maître reste en  $L_{1.n}$  et le chien passe de  $L_{2.m-1}$  à  $L_{2.m}$*
- Le maître passe simultanément de  $L_{1.n-1}$  à  $L_{1.n}$  et de  $L_{2.m-1}$  à  $L_{2.m}$ .*

<sup>18</sup> La programmation dynamique est un moyen d'obtenir la meilleure mise en correspondance possible entre deux séquences d'éléments semblables mais pas identiques.

Nous devons noter par ailleurs, que cette distance n'est applicable que pour les entités élémentaires. Son usage pour les entités complexes nécessite la gestion des sauts au niveau du calcul pratique de la distance.

### II.4.2.3. Probabilité d'association

La probabilité d'association est une mesure qui permet de décrire le type d'association pouvant avoir lieu entre deux entités surfaciques. Cette mesure peut être exprimée en termes ensemblistes, puisqu'elle s'appuie sur un croisement<sup>19</sup> entre les éléments de  $\mathcal{F}_0$ .

Le croisement des jeux de données géographiques, en général, induit des relations topologiques distinctes entre les ensembles des points formant les contours et les intérieurs des polygones, et éventuellement l'extérieur des entités [Egenhofer & Herring 1990; Egenhofer & Franzosa 1991; Jen & Boursier 1994]. Dans le cadre de cette thèse, nous ne nous intéressons qu'aux entités surfaciques appartenant à  $\mathcal{F}_0$ . En effet, lors de croisement des entités surfaciques, nous pourrions ne tenir compte que des intérieurs des entités. La description des liens topologiques résultats du croisement entre les entités surfaciques est donnée par [Venn 1881] et est connue depuis le temps d'Aristote sous la forme suivante :

*Soient deux classes A et B, on peut établir 5 relations entre ces deux classes de la manière suivante: "tout A est tout B", "tout A est une partie de B", "une partie de A est tout B", "une partie de A est une partie de B", "aucune partie de A ne correspond à aucune partie de B".*

D'une manière quantitative, ces 5 types de relations ensemblistes entre deux entités surfaciques peuvent être exprimés par des probabilités conditionnelles (probabilité d'association) en utilisant les mesures des aires des entités de la manière suivante :

$P(B|A) = \text{surface}(A \cap B) / \text{surface}(A)$  et  $P(A|B) = \text{surface}(A \cap B) / \text{surface}(B)$ .  
Ces 5 configurations sont illustrées par la figure II-23.

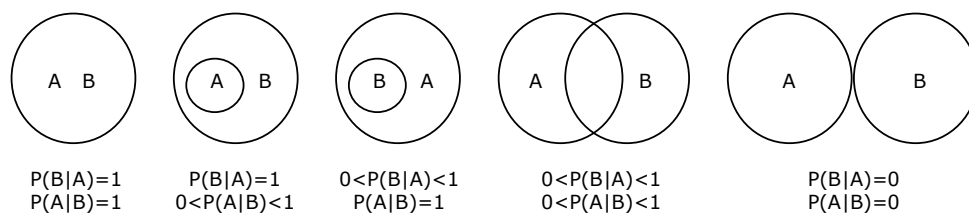


Figure II-23 : Les cinq configurations possibles décrites par Venn [Venn, 1881]

Cette probabilité d'association a été utilisée par [Phalakarn 1991] pour évaluer la qualité d'un processus de segmentation d'images par la mise en correspondance à une référence. Elle a été reprise par [Lemarié 1996] pour établir des liens d'associations entre deux bases de données pour des fins d'appariement dont nous avons critiqué

<sup>19</sup> Plus connu par le terme anglais "overlay"

l'emploi pour en proposer une fonction d'inclusion définie à partir de la probabilité d'association [Bel Hadj Ali 1997]. La fonction d'inclusion est définie comme suit :

*Soient A et B deux entités surfaciques éléments de  $\mathcal{F}_0$  pour lesquelles l'adhérence de l'intérieur est égale aux entités elles mêmes. Soient  $S(A)$  et  $S(B)$  les mesures respectives des aires des entités A et B, avec  $S(A) \neq 0$  et  $S(B) \neq 0$ . la fonction d'inclusion est alors donnée par :*

$$Fi(A, B) = \frac{S(A \cap B)}{\text{Min}[S(A), S(B)]} \quad [\text{II-50}]$$

Avec  $S(A \cap B)$  la mesure de l'aire commune entre les deux entités A et B. Cette fonction peut servir également comme une mesure de ressemblance entre les entités surfaciques; si deux entités sont très ressemblantes, leur fonction d'inclusion tend vers l'unité.

#### II.4.2.4. Distance surfacique

La distance surfacique est introduite par [Vauglin 1997], qui l'a définit de la manière suivante :

*Soient A et B deux entités surfaciques éléments de  $\mathcal{F}_0$  pour lesquelles l'adhérence de l'intérieur est égale aux entités elles mêmes. Soient  $S(A)$  et  $S(B)$  les mesures respectives des aires des entités A et B, avec  $S(A) \neq 0$  et  $S(B) \neq 0$ . On note  $\Delta$  l'opérateur de la différence symétrique ( $A\Delta B = A \setminus B + B \setminus A$ , avec  $A \setminus B$  représente le complémentaire de B dans A). La distance surfacique est alors donnée par :*

$$Ds(A, B) = \frac{S(A\Delta B)}{S(A \cup B)} \quad [\text{II-51}]$$

Nous notons que cette mesure est une distance au sens mathématique du terme dont les valeurs évoluent dans l'intervalle [0,1]. Si les deux entités sont totalement disjointes, leur distance surfacique est donc nulle. Par ailleurs s'il existe une égalité parfaite entre les deux entités, on aura une distance surfacique égale à l'unité.

Cette distance a montré l'essentiel de son intérêt pour les objets géographiques, surtout dans son utilisation pour la décision de la qualité des liens d'appariement dans une première ébauche d'un algorithme d'appariement des données surfaciques [Bel Hadj Ali 1997], et pour en établir des cartes de qualité de la géométrie des entités surfaciques par l'utilisation des valeurs de cette distance [Bel Hadj Ali & Vauglin 1999].

#### II.4.3. Distance entre fonctions angulaires

Dans cette section, l'espace  $\mathcal{L}$  des fonctions angulaires sera doté d'une métrique pour évaluer la différence de forme entre deux polygones de l'espace des éléments de  $\mathcal{F}$ .

On note que le changement du point O (point de départ pour les mesures, cf. figure II-2) fait subir à la fonction angulaire un décalage sur l'axe des abscisses.

La distance utilisée pour comparer les formes des polygones en se basant sur leurs fonctions angulaires repose sur la métrique  $L_2$  dans l'espace  $\mathcal{L}$ . Cette métrique est définie de la façon suivante :

Soient  $(A, B) \in \mathcal{F}^2$  et  $(\theta_A, \theta_B) \in \mathcal{L}^2$  leurs fonctions angulaires respectives. Le degré de similarité entre les deux polygones A et B peut être mesuré par une métrique  $L_2$  entre les deux fonctions  $\theta_A$  et  $\theta_B$ . La distance  $\delta_2$  fondée sur la norme  $L_2$  entre les fonctions  $\theta_A$  et  $\theta_B$  est définie par:

$$\delta_2(A, B) = \|\theta_A - \theta_B\|_2 = \left( \int_0^1 |\theta_A(s) - \theta_B(s)|^2 ds \right)^{\frac{1}{2}} \quad [\text{II-52}]$$

$\delta_p$  présente quelques propriétés indésirables dans notre contexte telle que la sensibilité à la rotation de l'un des polygones A ou B. Le choix du point origine des mesures influe également sur le résultat de la métrique. Pour éviter ces effets, on définit une autre métrique robuste à ces transformations.

On suppose que le point origine des mesures est décalé d'une quantité t le long du contour du polygone "A" et on suppose également que le polygone "A" a subi une transformation par rotation d'un angle  $\alpha$ , ce qui se traduit par la définition d'une nouvelle fonction angulaire pour le polygone "A":  $\forall s \in [0, 1], \theta_A(s+t) + \alpha$ . La nouvelle métrique est alors calculée de manière à être minimale pour l'ensemble des transformations de paramètres t et  $\alpha$ .

La nouvelle métrique est définie comme suit :

$$d_2(A, B) = \left( \min_{\alpha \in \mathbb{R}, t \in [0, 1]} \int_0^1 |\theta_A(s+t) - \theta_B(s) + \alpha|^2 ds \right)^{\frac{1}{2}} \quad [\text{II-53}]$$

On note que  $d_2(A, B)$  est une distance.

Bien qu'elle soit indispensable pour la propriété de l'invariance, la normalisation du contour reste un handicap pour cette distance, notamment dans le cas où les objets présentent une ressemblance locale.

La distance entre les fonctions angulaires des objets polygonaux a été initialement utilisée pour chercher le ou les objets homologues d'un objet nouveau, dans une base de modèles. Dans l'établissement de cette tâche bien particulière, [Alt & al. 1993; Cohen & Guibas 1997] préconisent un lissage du contour pour minimiser la sensibilité de l'indicateur face aux bruits parasites. Cette recommandation n'est toutefois pas applicable dans le cadre de cette étude, du fait que le but visé par l'utilisation de cet indicateur est la qualification de la différence de forme entre les objets géographiques. Donc, un lissage *a priori* fausserait le résultat.

D'autre part, [Arkin & al. 1991] démontrent d'une manière empirique que, si la valeur de la métrique entre les fonctions angulaires fondée sur la norme  $L_2$  est inférieure à 0.5, on peut dire que les deux objets comparés ont une forme semblable.

Les objets géographiques présentent plus de complexité par rapport aux objets géométriques simples (carré, triangle, carré bruité, etc.) et le seuil fixé par [Arkin & al. 1991] pourrait ne pas correspondre à la réalité des objets représentant de l'information géographique. L'estimation d'une valeur appropriée par la réalisation de tests empiriques s'impose, elle est présentée dans le §II.5.2.

#### II.4.4. Distance entre signatures de polygones

Soient deux polygones  $A$  et  $B$  éléments de  $\mathcal{F}$  avec  $S_A$  et  $S_B$  éléments de  $\mathcal{B}$  leurs signatures polygonales respectives (cf. §II.3.1.3). Pour confirmer ou infirmer la similarité entre ces deux polygones, on peut utiliser l'indicateur suivant :

$$\delta_2(A, B) = \|S_A - S_B\|_2 = \left( \int_0^1 |(S_A(s) - S_B(s))^2| ds \right)^{\frac{1}{2}} \quad [\text{II-54}]$$

où  $\|\cdot\|_2$  désigne la norme 2.

On note que cet indicateur est très sensible au choix du point de départ "origine des mesures", ainsi qu'à l'effet d'homothétie qui peut affecter l'entité surfacique. Donc, une re-définition de l'indicateur s'impose pour minimiser l'effet de ces deux opérations.

On suppose que le point origine des mesures est décalé d'une quantité  $t_{AB}$  le long du contour du polygone "B" et on suppose également que le polygone "B" a subi une transformation par homothétie de facteur  $k_{AB}$  ce qui se traduit par la définition d'une nouvelle signature polygonale pour le polygone "B":  $\forall s \in [0,1], k_{AB}S_A(s+t_{AB})$ .

Le nouvel indicateur est défini comme suit :

$$\forall (A, B) \in \mathcal{F}^2, \forall (S_A, S_B) \in \mathcal{B}^2, \forall t_{AB} \in [0,1]$$

$$d_2(A, B) = \left( \underset{k \in \mathbb{R}^+, t \in [0,1]}{\text{Min}} \int_0^1 |(S_A(s) - k S_B(s+t))|^2 ds \right)^{\frac{1}{2}} \quad [\text{II-55}]$$

En pratique, les données réelles ont une probabilité très faible de présenter deux objets qui sont en parfait rapport d'homothétie. En conséquence, la relation  $k_{AB} = 1/k_{BA}$  n'est pas tout à fait respectée, et par la suite,  $d_2(A, B)$  sera légèrement différente de  $d_2(B, A)$ . On choisit de prendre pour indicateur  $d$  la plus grande des deux quantités  $d_1$  et  $d_2$  :

$$d(A, B) = \max(d_2(A, B), d_2(B, A)) \quad [\text{II-56}]$$

#### Annulation de l'effet de l'homothétie :

Pour calculer en pratique l'une des composantes de l'indicateur de similarité, il est nécessaire de minimiser l'intégrale pour tout  $k_{AB}$  réel positif [Bel Hadj Ali 2000]. Soit la fonction  $h$  définie de la manière suivante :

$\forall (S_A, S_B) \in \mathcal{B}^2, \forall k_{AB} \in \mathbb{R}^{+*}, \forall t_{AB} \in [0,1],$  ( $t_{AB}$  représente le décalage entre les deux points de départ "origines des mesures" pour les deux entités A et B)

$$h(k_{AB}, t_{AB}) = \int_0^1 |(S_A(s) - k_{AB} S_B(s + t_{AB}))|^2 ds \quad [\text{II-57}]$$

Minimiser l'effet de l'homothétie pour les valeurs de l'indicateur revient à annuler la dérivée première en k de la fonction  $h(k,t)$ , et à choisir la solution qui correspond à un minimum :

$$\frac{\partial h(k_{AB}, t_{AB})}{\partial k_{AB}} = 0 \quad [\text{II-58}]$$

Ceci donne la valeur optimale de k, écrite de la manière suivante :

$$k_{AB}^*(t_{AB}) = \frac{\int_0^1 S_A(s) \cdot S_B(s + t_{AB}) ds}{\int_0^1 S_B^2(s + t_{AB}) ds} \quad [\text{II-59}]$$

En reprenant la définition de la première composante de l'indicateur de similarité (Equation II-53) et en remplaçant la variable k par sa valeur optimale, il ne reste qu'à minimiser l'effet du choix du point de départ "origine des mesures". La définition de la première composante de l'indicateur de similarité s'écrit alors de la manière suivante :

$$d_2(A, B) = \left( \underset{t \in [0,1]}{\text{Min}} \int_0^1 |(S_A(s) - k_{AB}^*(t_{AB}) S_B(s + t_{AB}))|^2 ds \right)^{\frac{1}{2}} \quad [\text{II-60}]$$

Les mêmes opérations de minimisation sont effectuées pour la deuxième composante de l'indicateur. Dans le cas d'une homothétie parfaite, on trouve que  $K_{AB} = 1/K_{BA}$ .

Une fois les deux composantes calculées, en minimisant les effets de k et de t, la valeur de l'indicateur de similarité est égale à la valeur maximale de ces deux composantes.

D'autres indicateurs peuvent être définis dans l'espace des signatures polygonales et ce par l'utilisation de la programmation dynamique. En cherchant la meilleure mise en correspondance entre la signature polygonale de l'objet de référence et la signature polygonale de l'objet à contrôler, la mesure de similarité sera donnée par le coût d'appariement entre les deux signatures polygonales. Cette technique est potentiellement utilisée pour réaliser la comparaison de deux signaux en reconnaissance de parole [Cheung & Eisenstein 1978] ou pour mesurer les coûts de transformations pour passer d'une chaîne de caractères codant des chromosomes à une autre [Milgram 1993]. Cette technique peut toutefois être adaptée pour comparer les signatures polygonales (cf. Figure II-24). Par ailleurs, faute de temps, nous n'avons pas pu l'étudier en détail afin de voir l'intérêt de son application sur des données géographiques réel. Cependant elle reste une piste intéressante à explorer.



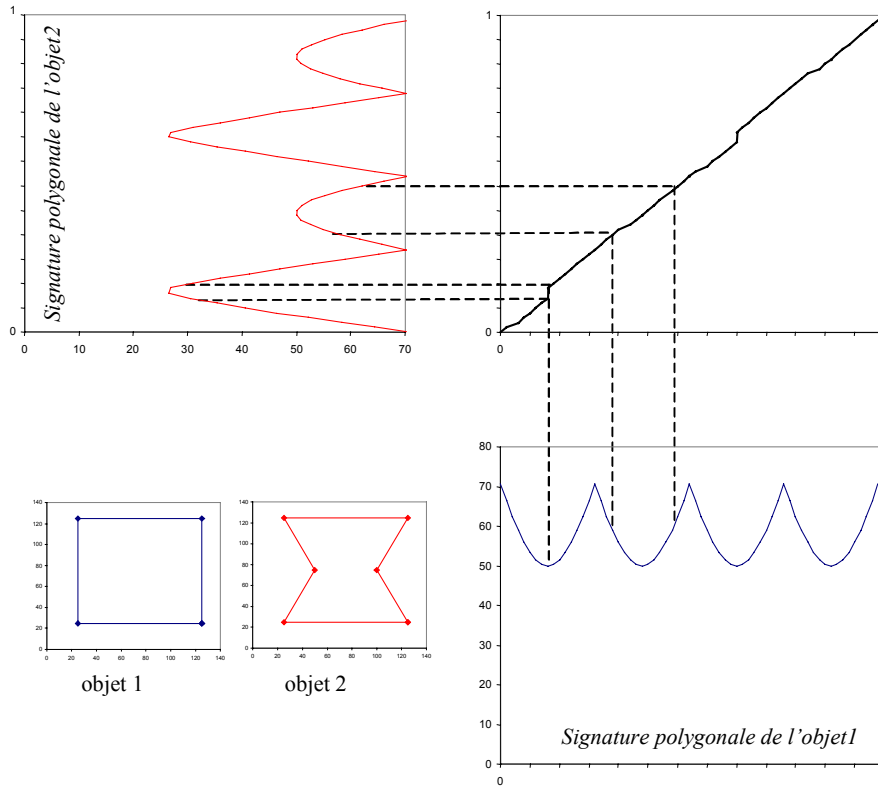


Figure II-24 : détermination du coût de mise en correspondance entre les signatures polygonales par l'utilisation de la programmation dynamique

### II.4.5. Distances associées à l'espace des fréquences

Soient  $\{a_n\}$  et  $\{b_n\}$ , les deux ensembles des descripteurs de Fourier (déterminés par l'équation II-3) respectivement des deux éléments A et B éléments de l'espace  $\mathcal{F}^1$ , et si nous admettons que ces deux éléments ne peuvent être représentés que par  $N_c$  descripteurs de Fourier. La distance entre les descripteurs peut alors être donnée par :

$$d(A, B) = \sqrt{\sum_{n=-N_c}^{N_c} |a_n - b_n|^2} \quad [II-61]$$

Cette distance reste très sensible aux effets des transformations affines, à savoir une homothétie d'échelle (d'un facteur  $\alpha$ ), la rotation (d'un angle  $\phi$ ) et le choix du point de départ pour les mesures (qu'on note p). En effet, pour réduire ces effets, il est indispensable de minimiser les paramètres  $(\alpha, \phi, p)$  lors du calcul de la distance. Cela peut s'écrire de la manière suivante :

$$d^*(A, B) = \sqrt{\text{Min}_{\alpha, \phi, p} \sum_{n=-N_c, n \neq 0}^{N_c} |a_n - \alpha e^{j(np+\phi)} b_n|^2} \quad [II-62]$$

Cette distance est trop coûteuse en temps de calcul.

Par ailleurs, en se proposant d'utiliser la représentation dans l'espace complexe en échantillonnant le contour avec un pas régulier. Soit  $z'(n)$  une séquence de points

obtenue en faisant subir à la séquence d'origine  $z(n)$  une translation d'une quantité égale à  $z_t$ , une rotation d'angle  $\phi$ , une homothétie de facteur  $\alpha$  et avec un déplacement du point "origine des mesures" d'une quantité  $l$ . D'une manière explicite, la séquence  $z'(n)$  est liée à la séquence  $z(n)$  de la manière suivante:

$$z'(n) = \alpha z(n-l)e^{j\phi} \quad [\text{II-63}]$$

Les descripteurs de Fourier correspondant à la séquence  $z'(n)$  seront donc donnés par :

$$Z'(k) = \sum_{n=0}^{Nb-1} z'(n)e^{-j\frac{2\pi nk}{NB}} = \alpha e^{j\phi} \sum_{n=0}^{Nb-1} z(n-l)e^{-j\frac{2\pi nk}{NB}} \quad [\text{II-64}]$$

En posant  $m = n - l$ , on obtient :

$$Z'(k) = \alpha e^{j\phi} \sum_{m=-l}^{Nb-1-l} z(m)e^{-j\frac{2\pi mk}{NB}} e^{-j\frac{2\pi lk}{NB}} = \alpha e^{-j(\phi + \frac{2\pi lk}{N})} Z(k) = M'(k)e^{j\theta'(k)} \quad [\text{II-65}]$$

Avec

$$M'(k) = \alpha M(k) \text{ et } \theta'(k) = -\phi + \theta(k) + \frac{2\pi lk}{Nb} \quad [\text{II-66}]$$

D'après les tests opérés sur des données géographiques, nous remarquons que les amplitudes des descripteurs de Fourier présentent une large dynamique, ce qui rend très difficile l'établissement d'une distance de type euclidienne entre leurs valeurs. [Rui & al. 1998] proposent de générer deux séquences à partir des amplitudes des descripteurs de Fourier respectifs aux deux entités à mesurer. La première séquence est relative aux amplitudes des harmoniques en calculant le rapport des amplitudes à chaque fréquence  $k$ , et la deuxième est relative à la différence de phases entre les harmoniques en mesurant la quantité  $-2\pi l / Nb$ . Donc, pour mesurer la différence entre les descripteurs de Fourier, ils proposent de calculer la variance des deux séquences décrites ci-dessus et d'établir une distance sous la forme d'une combinaison linéaire<sup>20</sup> des variances des deux séquences calculées. Si les deux entités sont égales, on aura donc une distance nulle entre leurs descripteurs, et plus elles sont dissemblables, plus la distance croît. Grâce à un premier jeu d'essai, nous avons voulu tester cette distance sur des données géographiques réels, mais nous nous sommes rendu compte de sa défaillance en terme de robustesse face à un bruit progressif (cf. §II.5.2.3.). En effet, les valeurs de cet indicateur ne suivent pas les évolutions de l'amplitude du bruit d'une manière proportionnelle. Ceci nous amène à exclure son utilisation et à définir un nouvel indicateur qui s'exprime en terme d'une moyenne de pourcentage d'erreur à cause de la dynamique des valeurs des amplitudes. Cet indicateur est donné par :

$$I_{DF}(A, B) = \frac{1}{n} \sum_{i=0}^n \frac{|\text{amp}^A(i) - \text{amp}^B(i)|}{\text{amp}^A(i)} \quad [\text{II-67}]$$

$n$  est le nombre de fréquences utilisées et  $\text{amp}^A(i)$  exprime la valeur de l'amplitude du spectre de  $A$  à la fréquence du rang  $i$ .

<sup>20</sup> [Rui & al. 1998] proposent la combinaison linéaire suivante : distance = 0.9\*variance.(rapport des amplitudes) + 0.1\*variance(différence des phases)

Notons que la représentation par des descripteurs de Fourier peut être également appliquée pour la modélisation du polygone par la signature polygonale, en adoptant la même distance.

#### II.4.6. Distances associées à l'espace des moments

Nous avons montré que les entités surfaciques peuvent être modélisées par trois types de moments mathématiques : les moments géométriques, les moments de Legendre et les moments de Zernike.

Les moments géométriques ne constituent pas une base orthogonale de représentation. En conséquence, l'information portée par ces moments se trouve redondante. Par ailleurs, les moments géométriques, tels qu'ils sont définis, ne sont pas invariants par les transformations affines, ce qui oblige à les combiner entre eux pour définir un ensemble de moments invariants. Ces moments invariants constituent des descripteurs de la géométrie de l'entité surfacique. En effet, comparer la forme de deux entités revient à comparer leurs moments invariants respectifs. Dans un premier temps, il est naturel de considérer le cas de la distance euclidienne entre les moments invariants, les entités surfaciques sont représentées dans l'espace des invariants par un vecteur de 22 composantes rangées par ordre croissant. La distance entre les invariants est définie de la manière suivante :

Soient  $P_1$  et  $P_2$  deux entités surfaciques, soient  $\varphi^1$  et  $\varphi^2$  leurs vecteurs d'invariants respectifs, la différence de forme entre ces deux entités sera alors donnée par :

$$d(P_1, P_2) = \left( \sum_{i=1}^{22} (\varphi_i^1 - \varphi_i^2)^2 \right)^{1/2} \quad [\text{II-68}]$$

D'autres indices peuvent être définis entre les deux vecteurs d'invariants. [Mukundan 1998] propose un indice de corrélation entre les vecteurs d'invariants, donné par :

$$r(P_1, P_2) = \frac{\sum_{i=1}^{22} \varphi_i^1 \varphi_i^2}{\left| \sum_{i=1}^{22} (\varphi_i^1)^2 \right|^{1/2} \left| \sum_{i=1}^{22} (\varphi_i^2)^2 \right|^{1/2}} \quad [\text{II-69}]$$

L'indice de corrélation (équation II-69) est défini par [Mukundan 1998] pour des fins d'utilisation de reconnaissance des formes. Disposant d'une base de modèles, l'auteur compare les moments d'un objet donné à ceux stockés dans une base de données et considère comme modèle, l'objet dont les moments ont l'indice de corrélation le plus proche de l'unité. Nous notons que notre problématique n'est pas tout à fait la même que celle de la reconnaissance des formes, puisque nous ne possédons pas une base de modèles, mais plutôt un couple d'entités appariées dont nous fixons pour but de qualifier leur disparité de forme.

Les moments de Legendre d'ordre élevés sont très sensibles aux différentes formes de bruit. Cette sensibilité permet de redéfinir une distance entre les moments de Legendre s'appuyant sur la distance euclidienne en la pondérant avec un coefficient allant de 0 à 1 [Delattre 2000]. La définition du coefficient pondérateur dépend fortement de l'application envisagée. Si l'on souhaite, par exemple, rechercher une égalité parfaite de forme, il va falloir pondérer les moments d'ordre élevé. Par contre, si le but visé est la mise en correspondance des bases de données à des échelles différentes, on a intérêt à pondérer les moments de bas ordre au détriment de ceux d'ordre élevé. Par ailleurs, cette approche présente un inconvénient majeur qui réside dans sa complexité algorithmique, puisque la fonction définissant le coefficient pondérateur n'est pas universelle et qu'il faut la déterminer pour chaque couple d'entités à mesurer. Pour déterminer la fonction du coefficient pondérateur, nous nous basons sur la fonction de contribution (cf. figure II-14) en affectant la valeur 1 à la plus haute contribution et 0 à la contribution positive la plus faible (les contributions négatives sont automatiquement mises à zéro). Cet indicateur a été testé avec succès en l'utilisant sur quelques couples d'entités appariés [Delattre 2000]. Cependant, son utilisation pour un jeu de données complet est très coûteuse en terme de temps de calcul<sup>21</sup>.

Les moments de Legendre et les moments de Zernike sont calculés après avoir normalisé l'entité surfacique, soit dans un disque unitaire, soit dans un carré unitaire. D'une manière générale, la normalisation (notamment dans le domaine de reconnaissance des formes) se fait en normalisant séparément les deux entités à comparer. La normalisation des entités de la sorte ne permet que de détecter les différences de formes. Cependant, nous pouvons utiliser d'autres normalisations, en tenant compte de la disposition des entités surfaciques dans l'espace géographique [Bel Hadj Ali 2001a]. Ces normalisations alternatives sont données par la figure II-25.

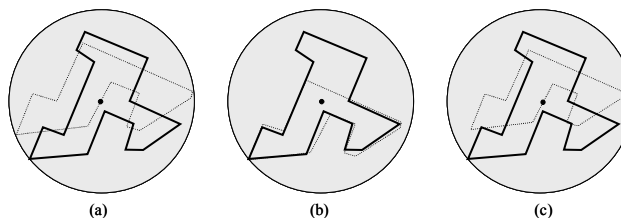


Figure II-25 : Différents types de normalisation

Le premier type de normalisation (figure II-25(a)) est le type le plus utilisé qui ne permet de détecter que les différences de forme entre les deux entités mesurées. Le deuxième type de normalisation (figure II-25(b)) consiste à préserver la position relative des entités, ce qui permet de traiter les entités telles qu'elles sont présentées dans les bases de données, mais en contre partie les mesures de forme seront biaisées. Le troisième type de normalisation (figure II-25(c)) correspond en quelque sorte au premier type de normalisation sans correction de l'effet de l'homothétie, ce qui permet de traiter les entités en préservant leurs tailles initiales.

<sup>21</sup> Le temps alloué pour le calcul de la distance pondérée pour un couple d'entités appariées englobe le calcul de leurs moments respectifs, l'établissement de la fonction de contribution des moments pour chaque entité, et enfin le calcul de la distance elle-même. Pour donner un ordre d'idée, cette opération peut prendre une trentaine de minutes pour un seul couple d'entités.

## **II.5. TESTS – CALIBRAGE DES MESURES**

Après avoir proposé des solutions alternatives pour la représentation et la modélisation de la géométrie des entités surfaciques, nous présentons dans cette section les travaux entrepris afin de tester la robustesse des représentations, ainsi que les mesures qui leurs sont associées face à des différents types de perturbations. Ceci permet d'effectuer un calibrage des mesures. Nous commençons par présenter la stratégie adoptée pour l'élaboration des tests, ainsi que les techniques de bruitage utilisées.

### **II.5.1. Stratégie utilisée (bruitage et simulations)**

En analysant de près la géométrie des données géographiques numériques, il se trouve que ces données peuvent subir des perturbations de deux types : perturbations liées à la façon de saisir les données et perturbations liées à la généralisation intentionnelle des données.

La majeure composante des perturbations est souvent commise lors de la saisie des données. Un point saisi a été considéré comme se trouvant d'une manière équiprobable "*n'importe où*" dans un disque centré sur les coordonnées nominales du point et de rayon  $\epsilon$ . Rappelons que cette supposition a été critiquée par [Vauglin 1997, pp. 40-41], en concluant que la bande  $\epsilon$  agit tout simplement comme un outil géométrique sans lien avec une mesure effective d'erreur sur les données géographiques. Dans ces travaux de thèse, Vauglin a également montré que les erreurs de la géométrie suivent une loi "mélange" d'une Gaussienne et d'une exponentielle symétrique. La composante Gaussienne correspond désormais à l'erreur de pointé, tandis que la composante exponentielle correspond aux effets de généralisation des primitives géométriques.

Des tests sur les représentations précédemment décrites ont été menés en tenant compte de toutes ces considérations sans oublier l'aspect de la structure de corrélation des écarts de la géométrie. Pour ce faire, nous avons utilisé l'algorithme développé par [Fouqué 1999]. Ces tests ont pour but d'évaluer la robustesse des indicateurs face aux bruits de différente intensité.

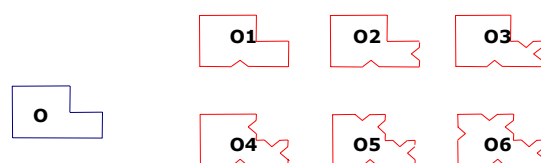
Nous envisageons deux types de tests. Le premier consiste à étudier le comportement d'un indicateur face à un bruit contrôlé affectant le contour de l'entité que nous appelons "robustesse au bruit". Le second consiste à bruitez les sommets par un bruit qui traduit d'une manière plus réaliste le comportement de l'erreur dans les bases de données géographiques. On appelle le second test "robustesse à la déformation".

Les tests ont été réalisés en deux étapes, d'abord sur des entités artificielles (générées à la main), puis sur des jeux de données réelles. Ces tests concernent toutes les modélisations et les indicateurs précédemment présentés.

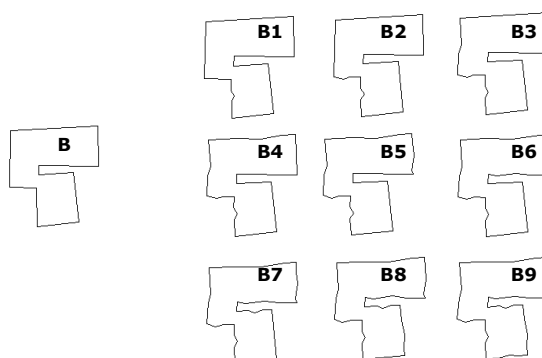
## II.5.2. Robustesses des indicateurs face aux perturbations (artefacts)

### II.5.2.1. Comportement de la fonction angulaire face au bruit

Pour mener à bien ce test, nous avons utilisé deux entités représentant du bâti. Sur l'une, nous avons ajouté un bruit progressif de même amplitude sur chacun des ses segments. Sur l'autre un bruitage aléatoire également sur chacun des segments. Ces deux entités sont représentées en figure II-26.



(a) entités avec bruitage régulier progressif



(b) entités avec bruitage aléatoire progressif

Figure II-26 : Entités tests

La distance entre la fonction angulaire de chacune des entités de la figure II-26 et ses "clones" bruités a été calculée. Etant donné, que le bruit qui affecte la première entité est progressif et présente la même amplitude, nous nous sommes attendus à voir une croissance dans les valeurs de la distance entre les fonctions angulaires, ce qui n'était pas tout à fait le cas. Les valeurs de la distance sont données comme suit :

$$0.329 - 0.444 - 0.513 - 0.568 - 0.591 - 0.571$$

On remarque, d'après ces mesures, que les valeurs de la distance suivent l'évolution du bruit jusqu'à la cinquième entité et décroît légèrement pour la sixième entité. Ceci est explicable par le fait que la fonction angulaire est établie en se servant de l'abscisse curviligne qui fait que le contour de l'entité est normalisé, et, par la suite, rend la distance moins sensible à un bruit uniforme. Cela rejoint la critique que fait [Veltkamp & Hagedoorn 1999] en la présentant comme la majeure défaillance de cette métrique. Cela pourrait expliquer en quelque sorte pourquoi la valeur de la distance

entre l'entité originale et la sixième entité (cf. figure II-26(a)) est inférieure à celle entre l'entité originale et l'entité 5, du fait que le bruit est plus uniformément réparti sur l'entité 6 que sur l'entité 5. La figure II-27 illustre les fonctions angulaires des cf. figure II-26(a), et sur laquelle on voit très bien le décalage qu'induit un bruit non uniforme sur les paliers de la fonction.

Le même test a été effectué pour l'entité (cf. figure II-26(b)) ce qui a permis de vérifier cette tendance (cf. figure II-28), surtout entre les distances au niveau de l'entité 4 et 5 par rapport à l'originale. Cependant, s'il existe une quelconque décroissance face à un bruit croissant, elle reste toujours minimale sans toutefois affecter les performances de cette distance à comparer les formes.

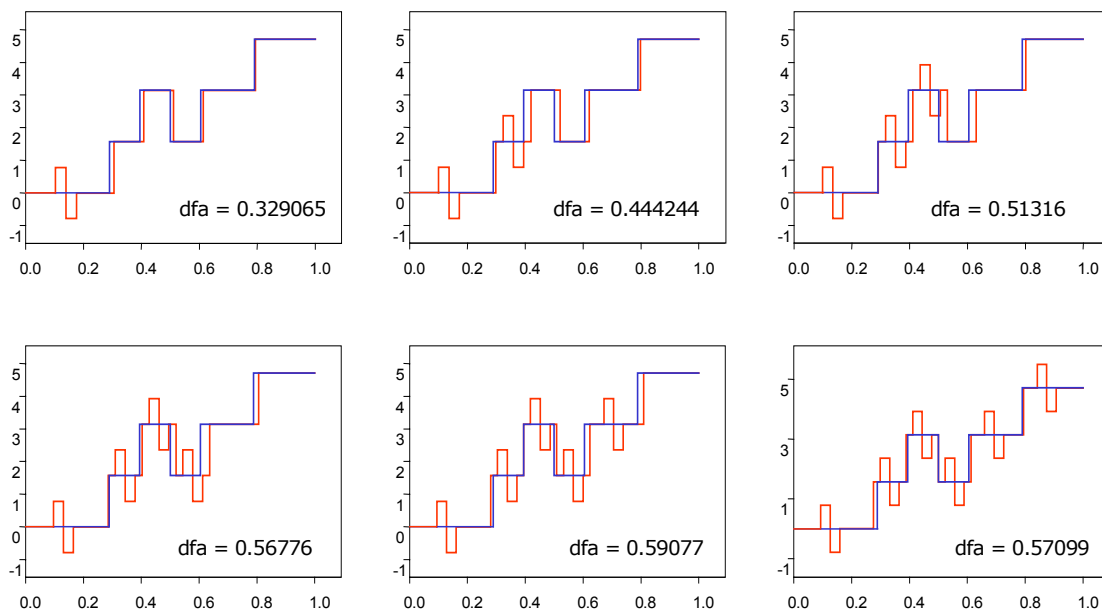


Figure II-27 : Fonctions angulaires relatives aux entités (figure II-26(a))

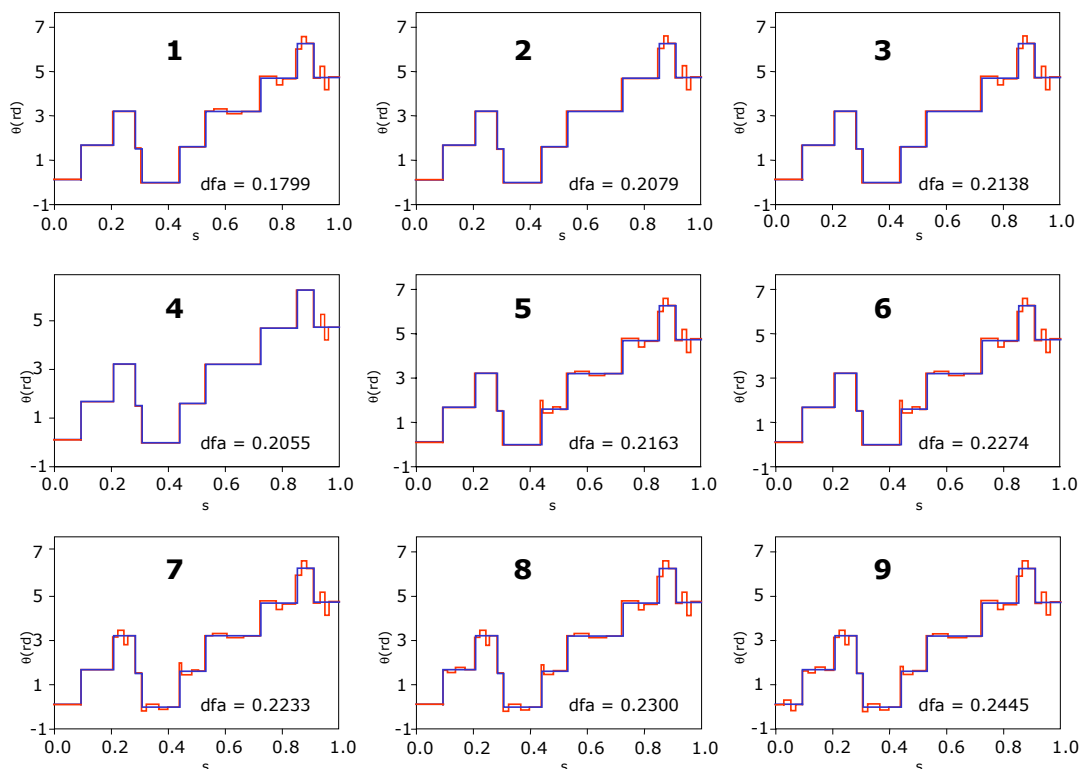


Figure II-28 : Fonctions angulaires relatives aux entités (figure II-26(b))

On note que le bruit que nous avons introduit sur les contours de l'entité cf. figure II-26(b)) est de faible amplitude en prenant le soin de ne pas créer de forte déviation angulaire. Ceci se traduit par des faibles variations sur les paliers de la fonction angulaire et par une faible variation au niveau des valeurs de la distance entre les fonctions angulaires (qui gravitent autour de la valeur 0.2).

### II.5.2.2. Comportement de la signature polygonale face aux bruits

Pour tester la robustesse de la signature polygonale face à un bruit progressif affectant le contour, nous avons utilisé les mêmes exemples que ceux utilisés dans le §II.5.2.1.

Les résultats de ce test démontrent également une légère sensibilité de la distance entre les signatures polygonales liée à l'utilisation de l'abscisse curviligne normalisée par le périmètre du polygone. Cette sensibilité peut s'expliquer par le fait que cette distance est moins sensible aux bruits uniformes. Les figures suivantes illustrent les signatures polygonales des polygones de la figure II-26.



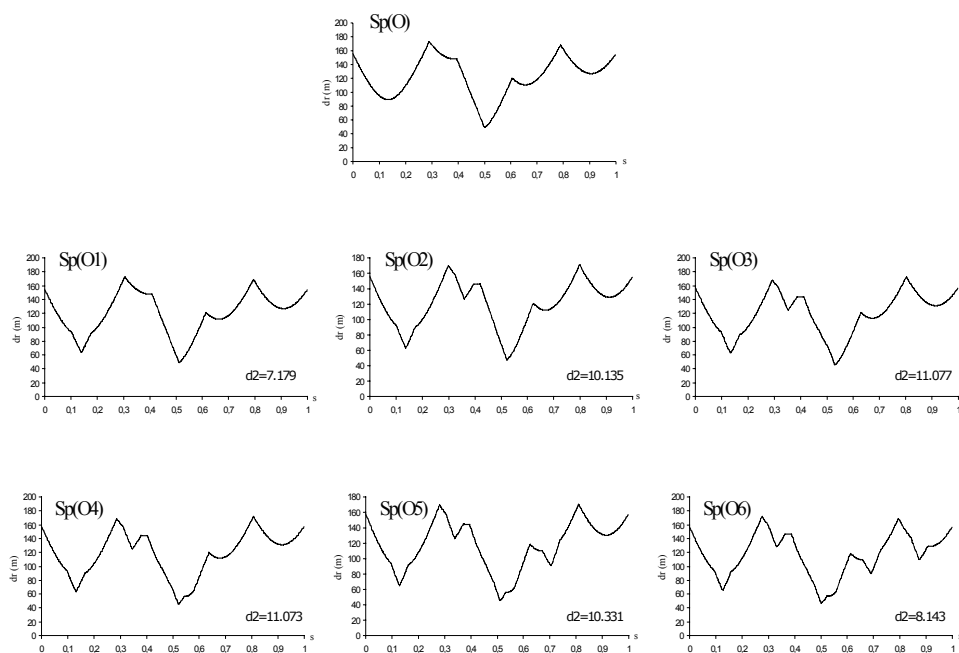


Figure II-29 : Signatures polygonales relatives aux polygones de la figure II-26(a)

Les chiffres en bas de chaque signature indiquent les distances entre chacune d'entre elles et la signature de l'objet O.

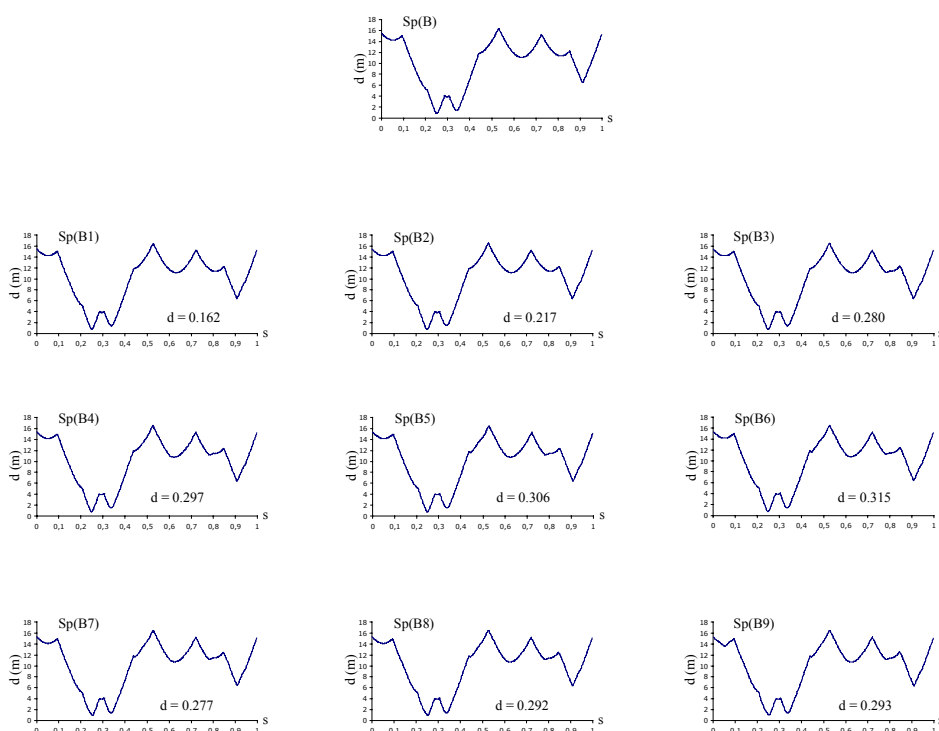


Figure II-30 : Signatures polygonales relatives aux polygones de la figure II-26(b)

Comme pour la distance entre les fonctions angulaires, la distance entre les signatures polygonales se montre trop sensible aux bruits uniformes qui affectent le

contour des polygones. L'uniformité du bruit atténue l'effet de normalisation et permet d'avoir un maximum de calage entre les courbes.

### II.5.2.3. Comportement des descripteurs de Fourier face aux bruits

Les polygones de la figure II-25 sont également utilisés pour tester la robustesse des descripteurs de Fourier, ainsi que l'indicateur qui leur est associé. Le but du test reste toujours de voir le comportement de l'indicateur face à un bruit progressif affectant le contour du polygone. Bien qu'un polygone donné puisse être représenté par quelques-uns de ses descripteurs de Fourier (cf. figure II-6), nous avons utilisé pour ce test l'ensemble complet des descripteurs de Fourier. Cette décision est motivée, essentiellement, par le fait que le calcul de ces descripteurs n'est pas d'une complexité algorithmique énorme du moment que nous utilisons l'algorithme de la transformée rapide de Fourier.

Dans un premier test, nous avons voulu tester l'indicateur défini par [Rui & al. 1998], et dont le résultat est consigné dans les tableaux suivants :

	O	O1	O2	O3	O4	O5	O6
O	0	2.696	6.906	12.175	8.024	8.033	12.741
O1			0.898	2.456	1.776	3.078	5.158
O2				1.713	2.197	6.269	3.924
O3					3.180	3.139	3.791
O4						0.758	3.266
O5							0.917
O6							0

Tableau II-1 : Mesures de similarité des objets de la figure II-26(a) par l'indicateur de [Rui & al. 1998]

	B	B1	B2	B3	B4	B5	B6	B7	B8	B9
B	0	0.946	1.550	1.553	1.511	2.794	2.633	3.766	3.353	4.228
B1			0.129	0.190	0.245	0.204	0.292	0.231	0.314	0.384
B2				0.017	0.027	0.140	0.161	0.182	0.174	0.284
B3					0.012	0.125	0.154	0.244	0.210	0.408
B4						0.087	0.122	0.170	0.146	0.284
B5							0.024	0.125	0.180	0.354
B6								0.158	0.162	0.283
B7									0.017	0.049
B8										0.048
B9										0

Tableau II-2 : Mesures de similarité des objets de la figure II-26(b) par notre indicateur

D'après les résultats obtenus de tableau II-1 et tableau II-2, on remarque que l'indicateur de similarité ne suit pas l'évolution du bruit. D'autre part, on voit sur la diagonale des deux tableaux que pour une même quantité de bruit, on n'a pas la même valeur de l'indicateur, ce qui reste problématique au niveau de la détermination d'une échelle de mesures.

Cependant, nous remplaçons l'indicateur de [Rui & al. 1998] par l'indicateur que nous avons défini dans le §II.4.5. (équation II-67). Cet indicateur est dissymétrique

puisque les mesures dans le sens de la référence vers le jeu de données ne sont pas égales à celles effectuées dans le sens contraire. Sa définition suppose donc qu'on mesure la similarité d'un jeu de données par rapport à une référence.

Les résultats obtenus par cet indicateur (cf. tableau II-3 et tableau II-4) montrent clairement qu'il ne présente pas les défaillances de celles de l'indicateur de [Rui & al. 1998] Les valeurs de l'indicateur évoluent d'une manière proportionnelle au bruit affectant le contour du polygone.

	O	O1	O2	O3	O4	O5	O6
O	0.000	0.707	0.917	1.146	1.150	1.318	1.499
O1			0.673	0.985	1.000	1.226	1.381
O2				0.744	0.907	1.161	1.274
O3					0.967	1.073	1.118
O4						0.618	0.943
O5							0.638
O6							0.000

Tableau II-3 : Mesures de similarité des objets de la figure II-26(a)

	B	B1	B2	B3	B4	B5	B6	B7	B8	B9
B	0.000	0.491	0.652	0.680	0.692	0.757	0.776	0.884	0.894	0.911
B1			0.207	0.250	0.265	0.288	0.346	0.353	0.390	0.407
B2				0.102	0.129	0.215	0.257	0.290	0.310	0.352
B3					0.069	0.188	0.234	0.291	0.297	0.364
B4						0.159	0.213	0.257	0.266	0.334
B5							0.113	0.224	0.258	0.318
B6								0.210	0.232	0.291
B7									0.093	0.154
B8										0.149
B9										0.000

Tableau II-4 : Mesures de similarité des objets de la figure II-26(b)

#### II.5.2.4. Comportement des moments face aux bruits

Nous reprenons, dans ce paragraphe, la même démarche utilisée pour les tests précédents en utilisant les mêmes exemples de la figure II-26. Nous cherchons à voir par ce test le comportement des différents moments utilisés, ainsi que les métriques qui leur sont associées face à un bruit progressif. Les résultats des mesures des objets  $O_i$  par rapport à l'objet O et des objets  $B_i$  par rapport à l'objet B sont présentés dans les tableaux suivants :

	d(O,O1)	d(O,O2)	d(O,O3)	d(O,O4)	d(O,O5)	d(O,O6)
Invariants	0.0231	0.0240	0.0543	0.0737	0.0860	0.0888
Zernike	0.0481	0.0769	0.1092	0.1132	0.1217	0.1403
Legendre	0.6645	1.0300	1.1296	1.3849	1.4444	1.6215

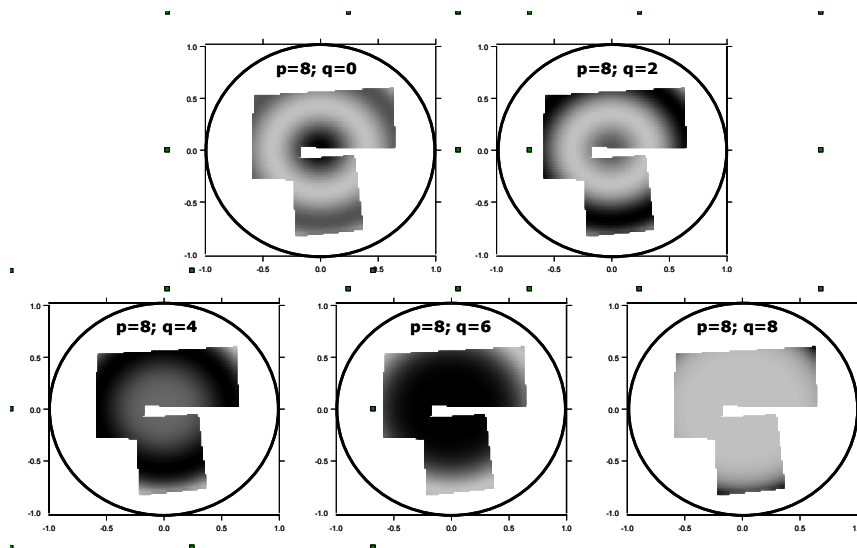
Tableau II-5 : Distances entre les moments relatifs aux objets de la figure II-26(a)

	d(B,B1)	d(B,B2)	d(B,B3)	d(B,B4)	d(B,B5)	d(B,B6)	d(B,B7)	d(B,B8)	d(B,B9)
Invariants	0.00346	0.00366	0.01010	0.00925	0.00997	0.01111	0.01116	0.12790	0.01582
Zernike	0.02262	0.02556	0.04495	0.05073	0.05119	0.05612	0.05946	0.05898	0.06851
Legendre	0.26380	0.34380	0.55070	0.66120	0.69420	0.75780	0.77270	0.79260	0.84790

Tableau II-6 : Distances entre les moments relatifs aux objets de la figure II-26(b)

Toutes les mesures obtenues montrent que les distances entre les moments invariants de Legendre et de Zernike évoluent dans le même sens que le bruit ajouté sur le contour des polygones tests. Les artefacts ajoutés sur l'objet O sont d'une amplitude conséquente, ce qui explique la dynamique des valeurs de la distance entre les invariants de Hu. Par ailleurs, les artefacts ajoutés sur l'objet B sont de faible amplitude, ce qui fait que les valeurs de la distance entre les invariants sont d'une variabilité faible et tournent à peu près autour de la valeur 0.01. Ceci confirme encore une fois que les invariants ne sont pas très sensibles aux faibles bruits et donnent une indication sur la forme globale du polygone. Bien que le bruit ajouté sur le polygone B soit faible, on trouve son impact sur les valeurs des distances entre les moments de Legendre et de Zernike, plus sensibles à ce genre de bruit.

Dans ce test, les moments ont été utilisés pour représenter des polygones simples. Cependant, toute l'information géométrique du polygone est utilisée, à savoir, son contour et son intérieur. La figure II-31 illustre la projection de la fonction  $f$  caractérisant l'objet O sur les polynômes de Zernike (Equation II-37) d'ordre 8 et de répétitions 0, 2, 4, 6 et 8.

Figure II-31 : Projection de l'objet O sur les polynômes  $R_{pq}$ ;  $p = 8$ 

*Le dégradé de noir illustre la variation des valeurs du polynôme  $R_{pq}$  des basses valeurs (couleur grise) aux hautes valeurs (couleur noire)*

La figure II-31 montre un exemple de projection de la fonction  $f$  caractérisant le polygone  $O$  sur les polynômes de Zernike d'ordre 8. On voit clairement sur cette figure que chaque répétition à cet ordre met l'accent sur une partie différente du polygone. Bien que ce test soit réalisé sur des polygones simples, sa crédibilité reste intacte pour le traitement des entités surfaciques complexes.

### II.5.3. Comportement des indicateurs face aux déformations

La saisie des données géographiques est souvent entachée d'erreurs. Plusieurs travaux ont essayé de modéliser le comportement statistique de ces erreurs [Goodchild 1991]. Ces travaux ont pu démontrer que les erreurs affectant les données géographiques sont essentiellement de deux types : des erreurs de pointé imputées à l'erreur humaine lors de la saisie des données et des "erreurs" dues aux procédés de généralisation des primitives géométriques. Les récents travaux de [Vauglin 1997], ainsi que ceux de [Hottier 1997] montrent que le premier type d'erreur suit une loi gaussienne, tandis que le deuxième type suit une loi exponentielle symétrique. Ces deux lois sont combinées pour donner naissance à un modèle hybride baptisé GES<sup>22</sup>. Cependant la composante de la gaussienne est plus majoritaire dans ce mélange. Ce modèle est utilisé pour bruiteur un échantillon composé de 439 bâtiments sur une zone pavillonnaire dans la région lyonnaise extrait de la BDTopo. A cet effet, nous avons utilisé l'algorithme de bruitage décrit et développé par [Fouqué 1999; Bonin 2000]. Dans le cadre de ce test, nous avons omis la composante exponentielle symétrique pour ne retenir que la composante gaussienne du bruit à introduire. Le bruit introduit suit alors une loi gaussienne de moyenne nulle dont l'écart type est donné par la formule suivante [Abbas 1994] :

$$\sigma = \frac{\pi}{2\sqrt{6}} \frac{emq}{\sqrt{c \cdot \ln(n)}} \quad [\text{II-70}]$$

où  $c$  est la constante d'Euler ( $c=0.577$ ) et  $n$  est le nombre de sommets du polygone

Le but de ce test est de voir l'évolution des valeurs des indicateurs précédemment définis face à différents niveaux de bruit. Ce test nous permettra de définir une plage ou un domaine d'évolution de la métrique en fonction du niveau du bruit introduit. A cet effet, nous avons simulé un bruit à différentes valeurs d'erreur moyenne quadratique que nous avons ajoutées aux données initiales. Les mesures sont alors effectuées entre les données initiales et les données bruitées. On note également que le bruit est introduit uniquement sur les sommets des polygones.

Les résultats des mesures d'écarts de forme entre les données initiales et les données bruitées sont illustrés dans les tableaux suivants :

<sup>22</sup> GES : pour Gaussienne Exponentielle Symétrique

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.0775	0.4189	0.1951	0.0615
0.7	0.0813	0.4550	0.2265	0.0649
1.0	0.0918	0.5681	0.2934	0.0741
1.3	0.1233	0.7507	0.3284	0.0866
1.5	0.1376	0.7157	0.3375	0.0970
1.8	0.1135	0.6923	0.3837	0.0807
2.0	0.1398	0.6923	0.3994	0.0800

Tableau II-7 : Mesures relatives à la distance entre les fonctions angulaires

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.0566	0.4347	0.1696	0.0493
0.7	0.0444	0.5429	0.2083	0.0899
1.0	0.0582	0.8265	0.4091	0.1605
1.3	0.1991	1.1797	0.4641	0.1311
1.5	0.0726	1.1709	0.5080	0.1993
1.8	0.1563	1.4736	0.6291	0.1760
2.0	0.1035	1.2854	0.7086	0.2170

Tableau II-8 : Mesures relatives à la distance entre les signatures polygonales

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.1250	4.7870	0.7598	0.5128
0.7	0.1920	4.5750	0.8670	0.5130
1.0	0.3840	5.7450	1.0294	0.6081
1.3	0.1410	6.0750	1.1088	0.6723
1.5	0.3930	9.2830	1.1916	0.8508
1.8	0.4690	7.4400	1.1791	0.7700
2.0	0.2970	8.3280	1.2195	0.7902

Tableau II-9 : Mesures relatives à l'indicateur entre les descripteurs de Fourier

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.0297	0.4945	0.1313	0.0644
0.7	0.0388	0.4875	0.1398	0.0606
1.0	0.0431	0.7789	0.2096	0.1178
1.3	0.0612	0.7785	0.2645	0.1116
1.5	0.0477	0.7863	0.2951	0.1132
1.8	0.0704	0.8000	0.3273	0.1353
2.0	0.0843	0.7958	0.3431	0.1349

Tableau II-10 : Mesures relatives à la distance entre les moments de Zernike

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.2609	2.9159	1.0927	0.4280
0.7	0.2906	2.4535	1.1868	0.3719
1.0	0.3721	3.6938	1.6262	0.6012
1.3	0.5548	4.2641	2.1668	0.6604
1.5	0.6396	3.5995	2.3328	0.4800
1.8	0.7157	3.9260	2.4765	0.5513
2.0	0.8261	3.9522	2.5723	0.5590

Tableau II-11 : Mesures relatives à la distance entre les moments de Legendre

EMQ	MIN	MAX	Moyenne	Ecart-Type
0.5	0.0026	0.4128	0.0460	0.0437
0.7	0.0012	0.3622	0.0470	0.0527
1.0	0.0036	0.5370	0.0773	0.0788
1.3	0.0110	1.8349	0.1061	0.1334
1.5	0.0117	0.6973	0.1250	0.1097
1.8	0.0121	1.3912	0.1320	0.1221
2.0	0.0083	5.9263	0.1326	0.0294

Tableau II-12 : Mesures relatives à la distance entre les invariants de Hu

En analysant l'évolution de la moyenne pour chacune des mesures effectuées, on remarque qu'elle croît d'une manière monotone avec l'erreur moyenne quadratique des données bruitées (cf. figure II-32). Cela montre que les indicateurs et les métriques que nous avons définis dans le cadre de ce travail suivent la même tendance du bruit.

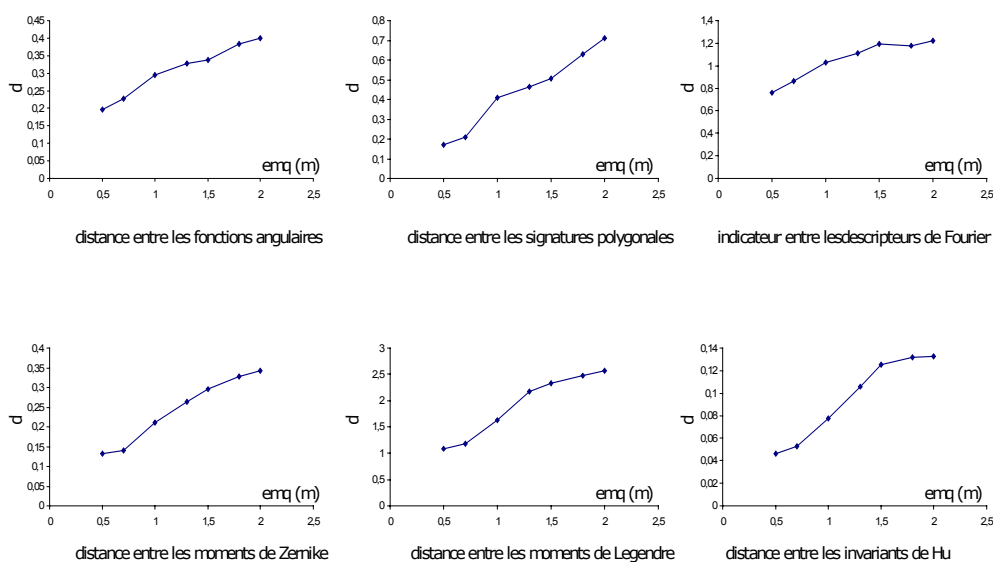


Figure II-32 : Evolution des moyennes des distances vs. l'erreur moyenne quadratique

Etant donné que l'emq de la BDTopo® [IGN 1997] est fixé à 2 mètres, ce test nous permet également de fixer les seuils pour les distances et les métriques que nous avons testées. Ces seuils peuvent avoir comme valeurs les valeurs des moyennes pour une erreur moyenne quadratique égale à 2m (cf. tableaux II-7 à II-12).

Ces tests ont permis de calibrer les mesures pour un éventuel contrôle de qualité en estimant l'erreur entre les bases de données surfaciques. Par ailleurs, pour la plupart des opérations de contrôle de qualité, on estime plutôt l'écart entre les entités des deux bases de données contrôlées. En effet, ces valeurs ne peuvent être utilisées que dans le cas où l'entité surfacique subit une déformation globale. Or, ce cas ne fait pas la règle pour les données géographiques puisqu'il existe des différences locales de forme telle une suppression d'un détail lors d'une opération de généralisation par exemple. Ce type de problème sera traité plus loin au chapitre III, en essayant de combiner l'utilisation des

indicateurs pour en dégager des règles pouvant renseigner sur la qualité de la forme et de la position.

## **II.6. SYNTHÈSE SUR LES REPRÉSENTATIONS ET LES INDICATEURS**

Le présent chapitre a deux buts principaux : le premier consiste à définir de nouveaux espaces de représentation qui permettent de traduire au mieux les caractéristiques de forme des objets géographiques et de s'affranchir de la représentation sous la forme d'une liste de points; le deuxième consiste à munir les espaces de représentation de métriques et d'indicateurs pouvant traduire la différence de forme et de position entre les entités géographiques. L'effort est porté principalement sur les entités surfaciques.

La qualité de la géométrie des données géographiques a été souvent abordée sous l'angle de la qualité de la position. Le contrôle de la forme de l'entité est souvent laissé à l'appréciation du contrôleur humain. Dans le cadre de cette thèse, nous avons mis l'accent sur le contrôle de la forme des entités surfaciques dans le but de pouvoir le réaliser d'une manière "automatique" pour s'affranchir des décisions humaines qui restent toutefois subjectives.

Un constat a été établi démontrant que la représentation cartésienne n'est pas la plus adaptée pour donner une idée sur la forme d'une entité quelconque, du fait que cette représentation donne une information localisée puisqu'elle s'appuie à l'origine sur la notion du point. A cet effet, de nouveaux espaces de représentation ont été définis qui permettent de modéliser la forme des entités surfaciques mieux que la représentation dans le plan cartésien. Ces représentations concernent deux types d'entités : les entités simples et les entités complexes.

### **II.6.1. Pour les entités simples**

Nous avons pris dans un premier temps le parti de définir de nouvelles représentations pour les entités simples. Pour la définition de ces nouvelles représentations, nous utilisons à la base la représentation cartésienne du contour de l'entité surfacique.

Dans ce chapitre, nous avons défini un nouvel espace de représentation dans lequel un polygone est représenté par une fonction "en escalier". Ces fonctions sont calculées à partir des angles que forment les segments du contour de l'entité entre eux. Elles sont définies dans l'intervalle  $[0,1]$  et prennent leurs valeurs dans  $\mathbb{R}$ . Cette représentation permet de donner une idée sur la géométrie du polygone, notamment au niveau de sa concavité ou de sa convexité. En se basant sur cette représentation, il est possible de détecter, outre la différence de forme, les deux points les plus homologues sur les contours des deux entités à comparer, ainsi que l'angle de rotation relatif entre elles.



Ces fonctions ne sont valables que pour les entités simples et ne peuvent pas être utilisées pour les entités complexes. La définition de ces fonctions dans l'intervalle  $[0,1]$  provient du fait qu'on normalise le point courant du contour par la longueur totale du périmètre. L'utilisation de cette normalisation rend cette représentation invariante par homothétie et facilite l'établissement d'une métrique dans cet espace. La métrique associée à cet espace est la norme  $L_2$ . Par ailleurs, les tests réalisés ont montré que cette normalisation peut avoir des effets indésirables puisque la métrique dans cet espace de représentation montre une légère insensibilité face aux bruits uniformes (nous entendons ici par "bruit uniforme" une même quantité de bruit ou déformation introduite sur tous les segments du contour de l'entité et non pas un bruit uniforme au sens statistique). Mis à part cette lacune, la métrique entre les fonctions angulaires a montré qu'elle suit la tendance de bruit, c'est à dire, elle croît si la quantité du bruit introduit croît.

Une représentation utilisant les distances séparant le centre de masse de l'entité aux points du contour a été également définie et testée. Cette représentation a été définie en normalisant le contour par sa longueur totale afin de pouvoir définir aisément une métrique dans cet espace de représentation. A l'instar de la métrique définie dans l'espace des fonctions angulaires, la métrique entre les signatures polygonales est définie à la base de la norme  $L_2$  et présente une même sensibilité face aux bruits uniformes. Cependant, cette métrique suit l'évolution du bruit. Cette représentation n'est valable que pour les entités simples. Nous notons aussi que l'utilisation de cette représentation permet de donner un indice de dilatation moyen entre les deux entités mesurées.

La dernière représentation que nous avons définie et testée pour les entités simples est une représentation utilisant les fréquences. Pour cette représentation, le contour de l'entité surfacique est échantillonné avec un pas régulier, et il sera transformé dans l'espace des fréquences. L'entité est donc représentée par deux séquences : l'une représentant les amplitudes de fréquences, l'autre les angles de phase des fréquences. Cette technique de représentation a été le plus souvent utilisée pour des besoins de lissage ou de filtrage. Peu de travaux l'ont utilisée pour des fins de qualification ou de mesures. A cet effet, nous avons repris quelques indicateurs définis dans le domaine de la reconnaissance des formes qui, suite aux tests réalisés sur des jeux de données géographiques, n'ont pas montré une grande robustesse face à différents types de bruit. Ces indicateurs ont été abandonnés au profit d'un nouvel indicateur que nous avons défini et qui a montré plus de robustesse face aux bruits. Nous notons qu'il est très difficile d'explicitier une interprétation physique de ces fréquences pour voir à quoi elles correspondent réellement en terme d'apport d'information sur la forme de l'entité. Par ailleurs, les tests ont montré que l'utilisation d'un nombre réduit de fréquences est suffisant pour décrire la forme globale d'une entité donnée.

Par opposition aux représentations citées précédemment qui ne sont valables que pour les entités simples, puisqu'elles n'utilisent que le contour des entités pour les représenter. Nous avons défini dans ce chapitre de nouveaux espaces de modélisation

qui utilisent toute l'information géométrique que contient l'entité surfacique. Elles tiennent compte de leurs contours et de leurs intérieurs, ce qui permet la représentation des entités complexes.

### **II.6.2. Pour les entités complexes**

Les entités complexes ont été définies comme étant des agrégats d'entités simples. Les représentations définies pour les entités simples ne sont donc pas obligatoirement valables pour représenter les entités complexes.

Pour pallier cette difficulté, nous avons développé dans ce chapitre une représentation des entités surfaciques par l'utilisation des techniques des moments mathématiques, à savoir, les moments géométriques et les moments orthogonaux de Legendre et de Zernike. Ces techniques ont prouvé leur utilité dans le domaine de la reconnaissance des formes. Cependant, leur utilisation dans le domaine de l'information géographique a soulevé quelques problèmes d'ordre pratique. Le premier problème rencontré consiste à définir le pas d'échantillonnage qu'il faut utiliser pour passer du mode vecteur au mode raster. Bien que ce problème ne diminue en rien l'intérêt de la méthode, sa résolution est utile pour le calcul pratique des moments. Nous avons montré que le choix de ce pas influe considérablement sur le calcul des moments. A cet effet, nous avons présenté une méthodologie pour choisir le pas optimal d'échantillonnage. Cette méthode a été intensivement testée sur des données issues de la BDTopo® et le pas retenu est de 0.5 mètres. Cette valeur n'est toutefois pas universelle. Pour l'utilisation d'autres types de données, nous recommandons la reprise de la méthode pour la détermination du pas d'échantillonnage correspondant au type de données.

Chaque entité surfacique peut être représentée par un ensemble infini de valeurs réelles ou complexes de ses moments. Nous avons montré également dans ce chapitre que l'utilisation d'un ensemble réduit de valeurs des moments est largement suffisante pour représenter et reconstruire une entité surfacique. La démonstration est faite pour les moments orthogonaux (moments de Zernike et de Legendre) puisqu'ils ont été définis sur la base de polynômes orthogonaux. Par conséquent, la démonstration a été établie par la reconstruction inverse des entités à partir de leurs moments. Il a été démontré qu'il suffit d'utiliser des moments de Legendre jusqu'à l'ordre 40 et les moments de Zernike jusqu'à l'ordre 20.

La représentation d'une entité surfacique par les moments rend compte de toutes ses caractéristiques géométriques puisque, pour leur détermination, nous utilisons le contour et l'intérieur de l'entité. Par la suite, chaque moment sera porteur d'une information concernant une partie particulière de l'entité. Cette représentation est adaptée pour mesurer les différences de forme entre les entités surfaciques. Pour accomplir cette tâche, nous avons défini une métrique entre les valeurs des moments. Les tests réalisés démontrent la robustesse de la métrique définie face à différents types de bruits.

## II.7. BILAN ET CRITIQUES

La représentation des entités surfaciques par des listes ordonnées de points dans l'espace cartésien a été critiquée pour son incapacité au niveau de la traduction des caractéristiques de la forme des entités et son inadaptation pour la représentation des entités complexes. Partant de ce constat, de nouveaux espaces de représentation ont été définis permettant ainsi de pallier ces problèmes.

Pour la définition des nouvelles représentations, nous avons utilisé la représentation des entités dans le plan cartésien par les listes ordonnées de points. Les nouvelles représentations ont permis la résolution du problème, notamment en ce qui concerne la qualification de la forme et de la position des entités surfaciques. Par ailleurs, elles présentent à leur tour des lacunes au niveau d'autres types d'utilisation de l'information géographique. Bien qu'elles présentent une souplesse et une robustesse dans le traitement de la forme des entités, les nouvelles représentations présentent une difficulté au niveau de leur utilisation pour le stockage des entités dans les bases de données ou bien pour leur visualisation. Cette difficulté est expliquée par la masse de l'information générée par ces représentations. Par exemple, un carré, qui au départ est définie par quatre coordonnées dans l'espace cartésien, sera représenté par un ensemble de 666 valeurs réelles dans l'espace des moments de Legendre.

Les représentations sont définies par rapport à un but bien déterminé et nous tenons à signaler que même pour ce but, il est très difficile qu'une seule représentation puisse aider à la résolution du problème. C'est pour cela que la définition de plusieurs représentations s'impose afin de cerner toutes les caractéristiques de la forme et de la position des entités surfaciques traitées. Il est à noter également qu'à chaque application correspond une ou plusieurs représentations dédiées. Donc, aucune représentation ne peut prétendre à l'universalité, ni à la généralité.

D'un point de vue pratique, toutes les représentations définies dans ce chapitre sont calculées "à la volée". Disposant initialement de la représentation cartésienne, les entités sont transformées dans les nouveaux espaces et les métriques sont calculées sans faire subir à la géométrie des données un quelconque changement, ni encombrer la base par de nouvelles informations. Seuls les résultats des mesures sont stockés dans la base.

La robustesse des indicateurs a été validée en introduisant du bruit qui approche au mieux un bruit réel. Ce bruit suit un modèle statistique de type exponentiel gaussien en utilisant un algorithme récemment développé par [Fouqué 1999; Bonin 2000]. Bien que l'algorithme de bruitage utilisé présente le mérite d'approcher la réalité, il est nécessaire de poursuivre les recherches concernant les méthodes de bruitage et notamment en tenant compte des corrélations entre les erreurs des points composants le contour du polygone, d'une part, et de la corrélation entre les polygones eux même, d'autre part. On a eu aussi recours à un bruit introduit à la main, et dont la quantité et l'amplitude sont connues afin de tester la sensibilité et l'évolution des métriques en sa présence.

Pour définir une représentation quelconque, nous nous sommes imposés au début de ce chapitre des critères de choix. Un des critères sélectionnés concerne la complexité algorithmique, ainsi que le temps dépensé pour le calcul des représentations. Bien que nous ne cherchions pas un calcul en temps réel (comme dans le cas de la reconnaissance automatique des formes), les algorithmes développés dans le cadre de ce travail ont été optimisés pour répondre à ce critère. Par exemple, parmi toutes les représentations définies, la plus coûteuse en temps de calcul est la représentation par les moments, puisqu'on passe par une étape de rasterisation des entités avant de les représenter dans l'espace des moments. A titre d'exemple, pour 500 couples d'entités à comparer, l'algorithme met 15 minutes<sup>23</sup> pour les représenter par les moments et calculer la disparité de forme entre elles.

---

<sup>23</sup> Ce temps inclus aussi le passage du mode vecteur au mode raster. Les utilitaires de calcul sont développés en C et tourne sur un PC PIII.

**CHAPITRE III :**  
**APPARIEMENT, ANALYSE DES MESURES**  
**ET CONTROLE QUALITE**

### III.1. INTRODUCTION ET APPROCHE

Nous avons défini dans le deuxième chapitre un ensemble de représentations et de mesures permettant d'évaluer la qualité de la géométrie des entités surfaciques dans les bases de données géographiques. Dans la première partie de ce chapitre, nous abordons les techniques d'appariement entre les bases de données surfaciques en proposant une méthode pour la mise en correspondance entre les objets surfaciques. Dans la deuxième partie de ce chapitre, nous proposons une méthode pour évaluer la qualité de la forme et de la position des entités appariées par une utilisation combinée des mesures définies dans le chapitre II.

Le problème majeur dans un processus d'appariement est la qualification des liens. On peut le formuler ainsi : quel est le degré de confiance qu'on peut accorder à un lien donné, afin de pouvoir conclure que les entités mises en correspondance représentent le même phénomène du monde réel? En effet, nous présentons une méthode dynamique s'appuyant sur la mesure de la probabilité d'association qui procure à l'algorithme une autonomie décisionnelle au niveau de l'établissement des liens d'association. Une fois que les liens d'association seront établis et validés, une étape de mesure sera engagée pour les qualifier et contrôler la qualité des entités surfacique. Cette approche est illustrée par la figure III-1.

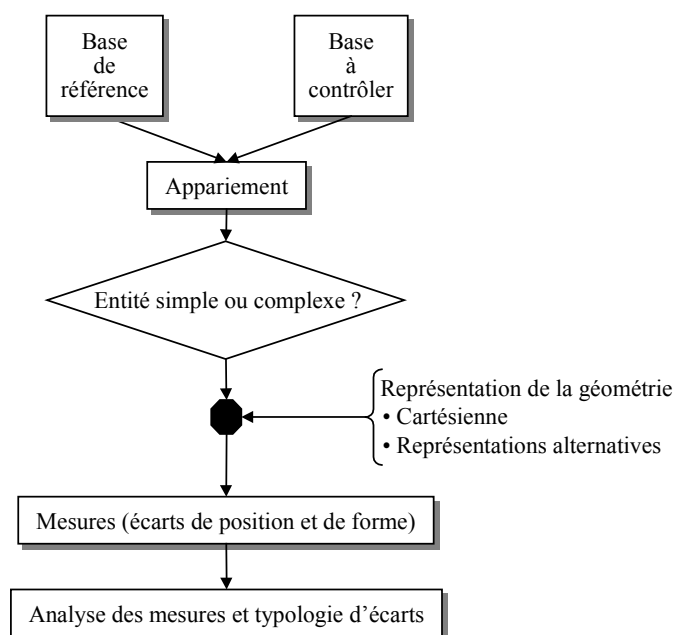


Figure III-1 : Approche générale

D'une manière globale, cette approche consiste, de prime abord, à appairer les entités provenant de deux bases différentes. Les liens obtenus à l'issue de l'opération d'appariement permettent de définir la nature des entités surfaciques. La classification des entités en deux classes (entités simples et entités complexes) se fait par l'analyse des cardinalités des liens. Selon la classification des entités, on sera amené à utiliser la représentation en coordonnées cartésiennes, si les entités sont considérées comme simples ou à utiliser de nouvelles représentations, dans le cas où les entités surfaciques

sont classées comme complexes. Les entités appariées subissent des mesures pour la qualification de la qualité de leurs formes et de leurs positions. Ces mesures sont utilisées d'une manière combinée pour déterminer les différentes classes d'écart. Cette approche est détaillée dans le présent chapitre.

## **III.2. APPARIEMENT DE DONNEES GEOGRAPHIQUES**

Les processus d'appariement sont souvent considérés comme complexes. Dans ce sens [Devegele 1997] établit le constat suivant :

*Les processus d'appariement sont complexes, ils confrontent un grand nombre d'informations et enchaînent plusieurs outils. La conception d'un processus d'appariement est donc une tâche fastidieuse. Néanmoins, la définition d'un processus générique facilite largement cette tâche. Cependant, la conception de ces processus peut encore être simplifiée en assistant l'utilisateur dans : le choix des outils d'appariement, le paramétrage de ces outils, le choix des filtres et enfin l'enchaînement des outils.*

Nous présenterons dans ce chapitre le développement d'une méthode d'appariement de données surfaciques fondé sur les techniques de croisement de données, ainsi que la démarche adoptée pour le contrôle de la qualité des liens d'associations entre les entités mises en correspondances. Nous présentons également le contrôle de la qualité géométrique des entités, qu'on trouve dans un cas d'appariement d'un jeu de données avec une référence.

Bien que l'appariement soit bien souvent cité dans la littérature, les méthodes permettant sa mise en œuvre sont bien souvent passées sous silence [Lemarié 1996]. Il existe beaucoup d'algorithmes d'appariement géométrique traitant le problème de façons diverses. Parmi toutes les applications utilisant l'appariement, nous nous concentrons sur l'estimation de la qualité des primitives géométriques d'une base de données par rapport à une autre plus précise, considérée comme référence. A cet effet, notre attention sera particulièrement portée sur l'appariement géométrique, qui lui-même s'appuie à l'origine sur les opérations de croisement de données tout en essayant de résoudre le problème d'appariement entre des bases de données n'ayant pas le même niveau d'abstraction. Les opérations de croisement de données ont été considérées comme un outil majeur pour l'analyse des données surfaciques, dont nous exposerons les développements qui leur sont rattachés, ainsi que les problèmes rencontrés.

### **III.2.1. Appariement géométrique des données surfaciques**

Par opposition à des algorithmes et méthodes déjà développés pour appairer les données géographiques et se basant sur l'utilisation de l'information linéaire, nous présentons dans cette section une méthode d'appariement fondée sur l'utilisation de la géométrie des surfaces. Les entités surfaciques ne seront donc pas appariées en ne s'appuyant que sur l'utilisation de leur contour, mais en utilisant leur intérieur également.

La méthode développée dans le cadre de cette thèse s'articule autour de quatre étapes principales. La première étape consiste à établir la liste des liens d'association entre les entités surfaciques des deux bases à appairer. Cette liste présente un nombre conséquent de couples pré-appariés, parmi lesquels il existe un certain nombre de liens indésirables qui seront éliminés lors d'une opération de filtrage (deuxième étape). Ces liens représentent des associations de type 1-à-1, or ce type de lien est basique et ne traduit pas pleinement la réelle liaison qui existe entre les objets des deux bases. A cet effet, il est nécessaire de chercher les liens multiples de type n-à-m qui peuvent éventuellement exister entre les entités (troisième étape). D'une manière générale, l'établissement des liens multiples est une opération fastidieuse et dont le résultat est loin d'être exempt d'erreur. A cet effet, nous présentons également une méthode itérative pour l'accomplissement de cette tâche, ainsi que l'adjonction à la méthode d'un outil de mesure afin de lui conférer une certaine autonomie pour décider de la vraisemblance du lien n-à-m établi. La dernière étape de la méthode consiste en une étape de filtrage final et de suppression des liens qu'on juge inutile.

Nous détaillons dans les paragraphes suivants toutes les étapes de la méthode, ainsi que les tests réalisés pour valider les résultats.

### III.2.1.1. Liens d'association

La première étape de la méthode consiste à mettre en correspondance, *a priori*, les entités des deux jeux de données. Cette étape se fonde sur une méthode probabiliste, initialement développée pour appairer des segments d'images avec un jeu de données vecteur [Servigne 1993][Salmeron & Milgram 1986; Phalakarn 1991; Le Men & Jamet 1990]. En pratique, cette étape se fonde sur les techniques de superposition des jeux de données (terme anglais : *map overlay*) [Andrews & al. 1994; Chrisman & al. 1992; Wagner 1988].

Les liens d'associations sont établis en utilisant les relations topologiques générées lors de l'étape de superposition de données. Un lien d'association entre deux entités est défini de la manière suivante :

Soient  $JD_1 = \{A_i\}$  et  $JD_2 = \{B_j\}$  deux jeux de données à appairer, avec  $A_i$  et  $B_j$  représentant respectivement les entités surfaciques simples composant les jeux de données  $JD_1$  et  $JD_2$ , On dit qu'il existe un lien d'association entre les deux objets  $A_i$  et  $B_j$  si et seulement si  $Surface(A_i \cap B_j) \neq 0$ .

Nous notons que les jeux de données doivent avoir le même système de projection pour pouvoir réaliser cette opération. On signale également que les jeux de données ne subissent aucune action manuelle de correction ou de calage géométrique, comme c'est le cas pour la méthode établie par [Walter & Fritsch 1999], où ils procèdent par un recalage géométrique des jeux de données avant d'opérer l'appariement proprement dit.

A ce niveau de l'appariement, les liens détectés sont stockés dans un graphe d'association (figure III-2). Le graphe généré est un graphe bipartite incomplet dont les



nœuds sont les entités des deux jeux de données à appairer et les arêtes sont les liens d'association.

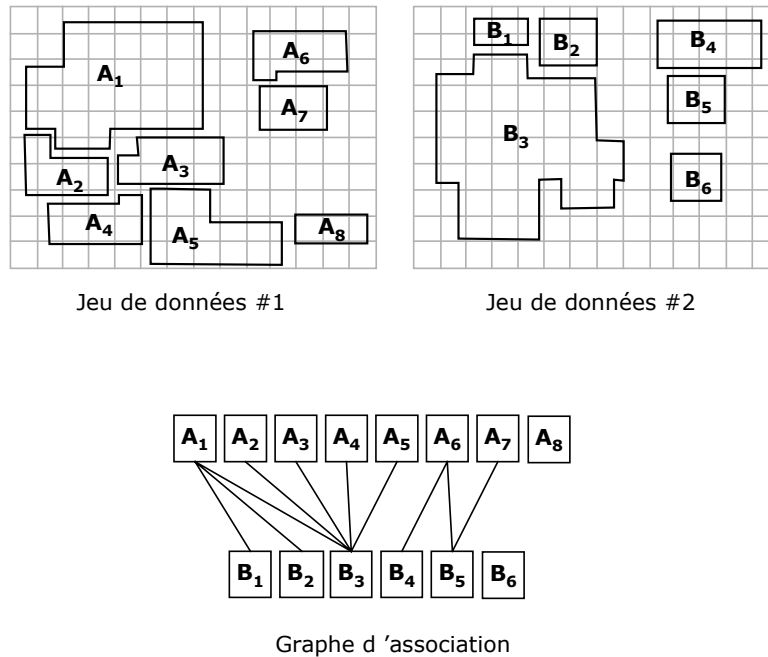


Figure III-2 : Etablissement du graphe d'association

Les arêtes du graphe reçoivent comme poids la mesure de l'aire de l'intersection entre les deux objets liés. Il se trouve qu'il existe des liens d'association générés à cause des intersections parasites. Ces liens ne traduisent aucune réalité physique et doivent être supprimés du graphe. Pour ce faire, nous définissons le filtre suivant :

Soient  $JD_1 = \{A_i\}$  et  $JD_2 = \{B_j\}$  deux jeux de données à appairer, avec  $A_i$  et  $B_j$  représentant respectivement les entités surfaciques simples qui composent les jeux de données  $JD_1$  et  $JD_2$ , On conserve un lien d'association entre les deux objets  $A_i$  et  $B_j$  si et seulement si

$$\text{Surface}(A_i \cap B_j) \leq \frac{S_m}{2}$$

$$\text{Avec } S_m = \left( \inf_{i=1}^n \text{Surface}(A_i); \inf_{j=1}^n \text{Surface}(B_j) \right). \quad [\text{III-1}]$$

Pour définir la valeur de  $S_m$  de manière alternative, on peut se référer aux documents de spécifications en prenant la valeur de la plus petite entité autorisée à être représentée dans la base de données (résolution).

Lorsque le choix est possible, il est préférable de prendre la valeur du seuil à partir des spécifications, car elle reflète une réalité physique de saisie et parce qu'elle est constante pour tous les jeux de données ayant les mêmes spécifications.

A défaut, la valeur de seuil sera calculée à partir de la surface de la plus petite entité des deux jeux à appairer. Le choix de la valeur de seuil par cette méthode dépend de la méthode de découpage mise en œuvre pour extraire les jeux de données. Un découpage selon un rectangle, par exemple, peut induire des entités artificielles dont la

surface est plus petite que celle autorisée. Dans ce cas, une détection automatique de la valeur de seuil peut être biaisée. Cette situation est illustrée par la figure III-3.



Figure III-3 : influence de l'effet de bord sur le choix du seuil de filtrage

*Les entités en couleur jaune sont générées par le découpage et leurs surfaces peuvent être inférieures à la résolution de la base de données.*

Cependant, si l'effet de bord ne se présente pas, la fixation du seuil de filtre à la valeur de la moitié de la plus petite surface est justifiée par le fait que si la plus petite entité présente dans l'un ou l'autre des deux jeux de données participe à un lien d'association, elle doit avoir au minimum 50% d'inclusion pour que son lien soit accepté.

Dans le cas de l'exemple de la figure III-2, le lien entre A6 et B5 sera supprimé puisqu'il ne satisfait pas la condition fixée par le filtre. Ce genre de liens indésirables est généré par la création des polygones parasites (terme anglais : *spurious polygons*) lors du croisement des deux jeux de données [McAlpine & Cook 1971]. Cette situation est souvent rencontrée lors de l'appariement des bases de données représentant l'occupation du sol.

Le filtre défini ici permet de supprimer les liens générés par les intersections parasites. Cependant, le graphe contient à ce stade des liens qui ne reflètent pas une réalité physique au niveau de l'appariement. Ces liens correspondent à des intersections insuffisantes : une intersection est jugée insuffisante si le recouvrement entre les deux entités est inférieur à 20% de la plus petite surface des deux entités. Ce filtre utilise la fonction d'inclusion que nous avons définie dans le paragraphe II.4.2.3. (chapitre II), comme suit :

*Soient  $JD_1 = \{A_i\}$  et  $JD_2 = \{B_j\}$  deux jeux de données à appairer, avec  $A_i$  et  $B_j$  représentant respectivement des entités surfaciques simples qui composent les jeux de données  $JD_1$  et  $JD_2$ , On dit qu'un lien d'association entre les deux entités  $A_i$  et  $B_j$  est inutile avec  $Surface(A_i \cap B_j) \neq 0$  si  $Fi(A_i, B_j) < 0,2$*

A ce stade de l'appariement, les entités des deux jeux de données sont associées les unes aux autres par des liens de type 0-à-1, 1-à-0 ou 1-à-1. Cependant, les liens de type 1-à-1 ne sont pas suffisants pour refléter les situations des appariements multiples de type n-à-m (plusieurs entités liées à plusieurs entités).

### III.2.1.2. Appariement complet

L'établissement des liens de type n-à-m est souvent considéré comme un problème majeur dans les algorithmes d'appariement de données. Pour le résoudre, on s'appuie sur une description statistique du croisement des données surfaciques.

#### III.2.1.2.1. Description statistique de la méthode de croisement de données

Le croisement des données, et leur superposition peuvent être décrits par des règles en termes de mesures d'association et de modèles [Zaslavsky 1995]. Ces modèles sont initialement fondés sur une observation probabiliste d'une entité sous la condition d'indépendance. Dans le cas du croisement de deux jeux de données, le modèle de relation entre les objets des deux bases peut être représenté sous la forme simple du diagramme de Venn [Venn 1881]. Le diagramme de Venn est décrit graphiquement par cinq dispositions (cf. figure II-23, chapitre II).

En d'autres termes, nous pouvons considérer une relation d'association entre deux objets comme étant une probabilité conditionnelle, en l'occurrence la probabilité de présence d'un objet B dans une base sachant la présence d'un objet A dans l'autre base. Zaslavsky a essayé de formaliser ce modèle en faisant le rapprochement avec une analyse statistique utilisée dans le dépouillement des questionnaires avec des variables nominales [Chesnokov 1982]. [Zaslavsky 1995] a utilisé cette analyse pour proposer une méthode alternative à la matrice de contingence classique pour mesurer les écarts entre deux jeux de données à pavage complet. Cette analyse est dite "*Determinacy analysis*"<sup>24</sup>.

Soient a et b deux éléments de l'ensemble  $\mathcal{F}$  appartenant respectivement à deux jeux de données différents. L'intersection de ces deux éléments est différente de l'ensemble vide et donne naissance à un élément c ( $c = a \cap b$ ) de  $\mathcal{F}$ . L'analyse de la "*Détermination*" s'appuie sur le postulat suivant : "**Si a, alors b**" explicitement défini dans le contexte particulier. L'utilisation de la "*Détermination*" tente donc d'expliquer la présence de l'élément a dans un des jeux de données connaissant la présence de l'élément b dans un autre jeu de données. Cette explication se fait en accompagnant le postulat de deux occurrences non nulles appelées l'exactitude et la complétude<sup>25</sup>, qui permettent de décrire le postulat d'une manière analytique. La "*Détermination*" peut être représentée par une relation statistique entre les catégories "a" et "b" qui sont considérées respectivement comme l'argument et la fonction de la "*Détermination*".

L'exactitude de la "*Détermination*", notée  $I(a \rightarrow b)$  est définie comme une probabilité conditionnelle de "b" sachant "a" et est donnée par :

<sup>24</sup> "*Determinacy*" n'a aucun synonyme dans la langue française. Nous employons le terme "*Détermination*" comme traduction de "*Determinacy*".

<sup>25</sup> On indique par exactitude et complétude les deux caractéristiques du postulat. Ils n'ont rien à voir avec les deux termes abondamment utilisés dans le jargon de la qualité des données géographiques.

$$P(b|a) = \frac{S(a \cap b)}{S(a)} \quad \text{[III-2]}$$

Avec  $S(a \cap b)$  désignant la mesure de la surface commune entre "a" et "b" et  $S(a)$  la mesure de la surface de "a".

La complétude de la "Détermination", notée  $C(a \rightarrow b)$  est donnée par:

$$P(a|b) = \frac{S(a \cap b)}{S(b)} \quad \text{[III-3]}$$

Ces deux notions représentent également les probabilités d'association dans les deux sens  $a \rightarrow b$  et  $b \rightarrow a$  que nous avons définies en §II.4.2.3.

La "Détermination" est utilisée pour évaluer tous les liens d'appariement établis par l'analyse de l'exactitude et de la complétude de la "Détermination" de chacun d'eux. De part leurs définitions, l'exactitude et la complétude prennent leurs valeurs dans l'intervalle  $[0,1]$ . Si  $I(a \rightarrow b)$  tend vers 1, alors a est inclus dans b. De plus si  $C(a \rightarrow b)$  tend vers 1, alors l'objet "référence" contient l'objet du "jeu de données".

Une analyse combinée des deux paramètres permet de contrôler la qualité des liens établis. Par exemple, les fortes valeurs de  $I(a \rightarrow b)$  donnent une indication sur l'inclusion totale de l'ensemble a dans l'ensemble b. L'analyse de  $C(a \rightarrow b)$  pour les fortes valeurs de  $I(a \rightarrow b)$  montre que l'inclusion de a dans b évolue proportionnellement (ligne supérieure de la figure III-4). Cette interprétation est illustrée par la figure III-4.

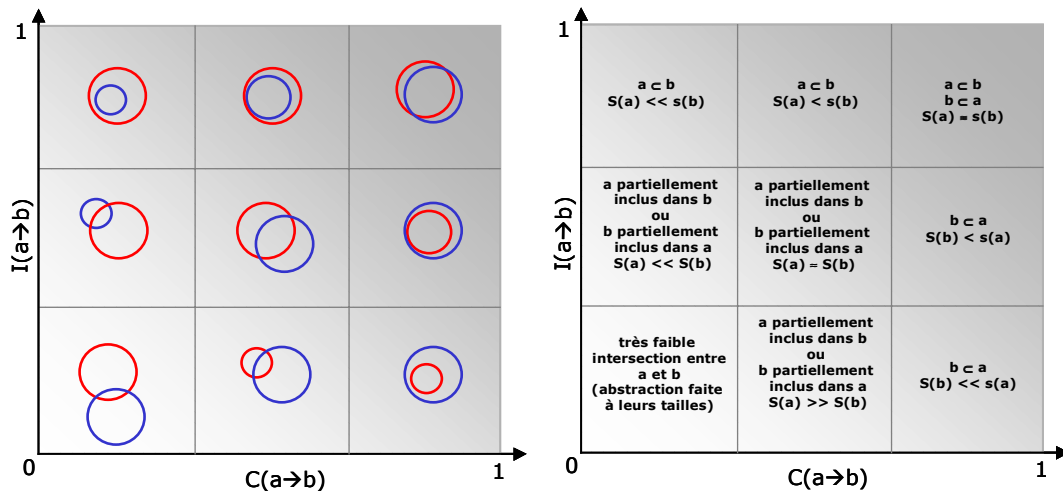


Figure III-4 : Evaluation des liens d'appariement par l'analyse de leur "Détermination"

### III.2.1.2.2. Etablissement des liens multiples

A l'issue de l'étape de l'établissement des liens d'associations et après le filtrage des liens parasites, nous obtenons un graphe bipartite incomplet. Ce graphe est re-traité afin de détecter les appariements de type n-à-m. Pour réaliser cette tâche, le graphe d'association est transcrit dans une matrice de M lignes et N colonnes (M et N désignent respectivement les nombres d'entités de la première base et de la deuxième base qui

participent au moins à un lien d'association). Cette transcription est illustrée par la figure III-5.

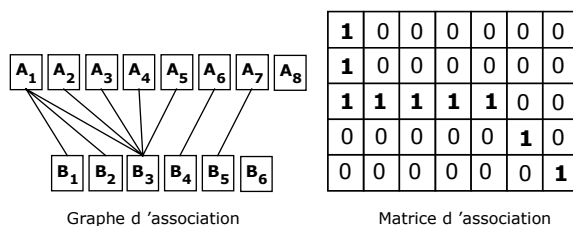


Figure III-5 : matrice d'association

La mise en forme des liens d'association sous la forme d'une matrice permet d'aider la recherche des liens multiples. Pour cela, nous avons développé l'algorithme qui permet le parcours de la matrice et l'établissement des liens multiples. La méthode développée pour détecter un lien de type n-à-m est donnée par le pseudo-code suivant :

```

List1tot = {ID1} // ensembles de tous les nœuds du graphe représentant les entités de la BDG#1
List2tot = {ID2} // ensembles de tous les nœuds du graphe représentant les entités de la BDG#2
NumApp = 1 // initialiser les identifiants des liens multiples à 1
Tant que (list1tot ≠ ∅)
  Créer un ensemble vide L1
  Créer un ensemble vide L2
  Sélectionner tous les liens dont le premier élément de la liste y participe
  Ajouter le premier élément de l'ensemble List1tot dans L1
  Pour tous les liens sélectionnés
    Ajouter la valeur ID2 du lien dans L2
  FIN
  Initialiser un compteur à 0
  Définir une valeur test à 1
  Tant que (test ≠ 0)
    Indicateur = compteur
    Pour tout élément i de L2
      Sélectionner tous les liens contenant le nœud i
      Pour tous les liens sélectionnés
        Val = la valeur ID1
        si (val ∉ L1) alors
          ajouter val à l'ensemble L1
          compteur = compteur+1
        Fin
      Fin
    Fin
    Supprimer les doublons de l'ensemble L1
    Pour tout élément i de L1
      Sélectionner tous les liens contenant le nœud i
      Pour tous les liens sélectionnés
        Val = la valeur ID2
        si (val ∉ L2) alors
          ajouter val à l'ensemble L2
          compteur = compteur+1
        Fin
      Fin
    Fin
    Supprimer les doublons de l'ensemble L2
    Test = indic - compteur
  Fin
  Pour tout élément i de l'ensemble L1
  Pour tout élément j de l'ensemble L2
  Sélectionner toutes arêtes ayant pour nœuds i et j
  si (l'ensemble des arêtes ≠ ∅) alors
    créer un lien multiple et lui attribuer le Numéro NumApp
  Fin
Fin
Fin
Retirer tous les éléments appartenant à l'ensemble L1 de l'ensemble list1tot
Retirer tous les éléments appartenant à l'ensemble L2 de l'ensemble list2tot
numapp = numapp + 1
Fin

```

L'application de cet algorithme sur les données de l'exemple de la figure III-5 permet la détection de deux liens simples et de un lien multiple de type 5-à-3. A ce stade, nous aboutissons à l'établissement des liens multiples sans toutefois pouvoir vérifier leur validité. Pour tester la validité des liens d'appariement, nous proposons une

analyse utilisant la notion de la "*Détermination*". Pour valider un lien de type n-à-m ou à défaut le remplacer par un lien plus satisfaisant n'-à-m' (avec  $n' \leq n$  et  $m' \leq m$ ), nous proposons la méthode suivante :

Calculer l'exactitude et la complétude pour toutes les combinaisons possibles entre les n et m objets appariés d'un lien donné en respectant le graphe d'association. En effet, on isole les liens multiples et on traite chacun d'eux à la fois par l'élimination d'une ou plusieurs arêtes de son graphe, et par la recherche de la configuration qui maximise à la fois l'exactitude et la complétude. Maximiser l'exactitude et la complétude du lien revient à maximiser la probabilité d'existence des entités d'un des jeux de données, participantes au lien sachant l'existence des entités qui leurs sont homologues dans l'autre jeu de données, et inversement. La maximisation des deux paramètres de la "*Détermination*" permet ainsi d'obtenir un lien reflétant le plus possible la réalité physique. Nous convenons d'appeler le lien obtenu par "lien optimal".

En reprenant l'exemple de la figure III-2, nous calculons l'exactitude et la complétude pour le lien multiple de type 5-à-3. La recherche du lien optimal se fait en maximisant l'exactitude et la complétude du lien multiple parmi toutes les valeurs calculées. La maximisation se fait en recherchant la valeur maximale de la somme des valeurs de l'exactitude et de la complétude. Dans le cas de l'exemple de la figure III-2, cette condition est atteinte en retirant l'arête A5-B3 du graphe d'association (Exactitude = 0.797; complétude = 0.835). L'analyse visuelle de la configuration des entités renforce le résultat obtenu du fait que l'entité A5 bien qu'elle ait un lien d'association avec B3 et puisse participer à la formation du lien multiple, son retrait de la configuration initiale rend le lien optimal. Donc, la cardinalité du lien optimal sera réduit à 4-à-3. L'implémentation de cet outil présente une grande complexité algorithmique. Pour cela, nous proposons de limiter le champ de choix des arêtes à supprimer en se basant sur la fonction d'inclusion entre les objets (définie dans le paragraphe II.4.2.3., chapitre II). Toutes les arêtes ayant une valeur de la fonction d'inclusion inférieure à 0.5 seront des candidats à la suppression lors de la recherche du lien optimal.

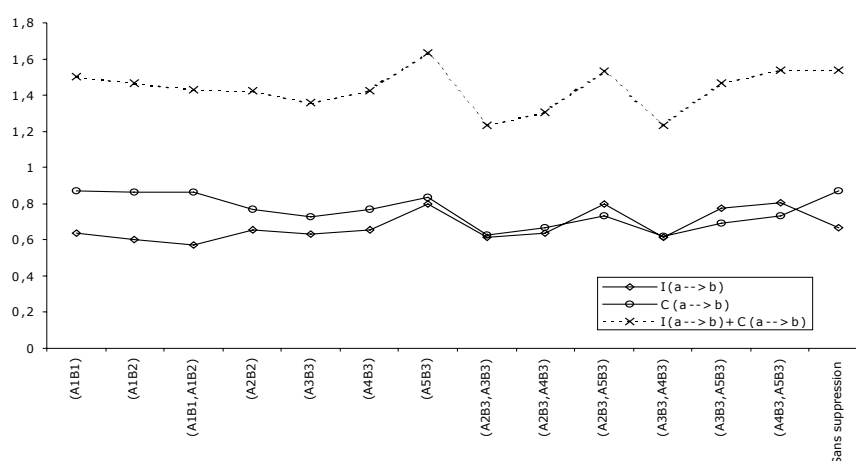


Figure III-6 : Exactitude et complétude du lien multiple 5-à-3 de l'exemple de la Figure III-2

L'évaluation de liens d'appariement nécessite la fixation d'un seuil pour les valeurs de l'exactitude et de la complétude. [Zaslavsky 1995] adoptait la valeur 0.5 pour l'exactitude en se plaçant dans un contexte de contrôle d'un jeu de données par rapport à un autre. Cependant, si on se place dans un autre contexte tel la construction d'un serveur multi-échelles, la notion de référence n'aura plus lieu d'exister, et la complétude aura la même importance que l'exactitude. Ainsi le seuil adopté pour l'exactitude peut être utilisé pour la complétude. Ces valeurs de seuils dépendent des données appariées et ne peuvent jamais être définies d'une manière universelle. Elles seront fixées d'une manière empirique lors des tests d'appariement sur des jeux de données réels.

Cette méthode dote l'algorithme d'appariement d'un outil lui permettant de décider de la façon optimale pour construire un lien multiple. Cette méthode permet une première évaluation des liens de correspondances que nous allons analyser plus loin dans ce chapitre en analysant les mesures faites entre les entités mises en correspondance.

Les deux composantes de la "Détermination" sont dissymétriques. Cependant, la complétude d'un jeu de données vers un autre indique l'exactitude dans le sens inverse. L'utilisation de ces deux paramètres est recommandée dans le cas où les deux jeux de données à appairer ne sont pas saisis à la même échelle. Par ailleurs, si les deux jeux de données sont à la même échelle (cas d'un contrôle de qualité, par exemple), l'utilisation de la distance surfacique (cf. §II.4.2.4., chapitre II) peut largement suffire. La distance surfacique peut être exprimée en fonction des paramètres de la "Détermination" de la façon suivante [Bel Hadj Ali 2001b] :

Soient deux entités  $A$  et  $B$  appartenant respectivement aux deux jeux de données  $JD1$  et  $JD2$  et ayant un lien d'appariement entre eux. Soit  $(x, y) \in [0, 1]^2$ , avec  $x = I(A \rightarrow B)$  et  $y = C(A \rightarrow B)$ . La distance surfacique  $D_s(A, B)$  est donnée par :

$$D_s(A, B) = \frac{2xy - x - y}{xy - x - y} \quad [\text{III-4}]$$

L'équation III-4 est illustrée graphiquement par la figure suivante :

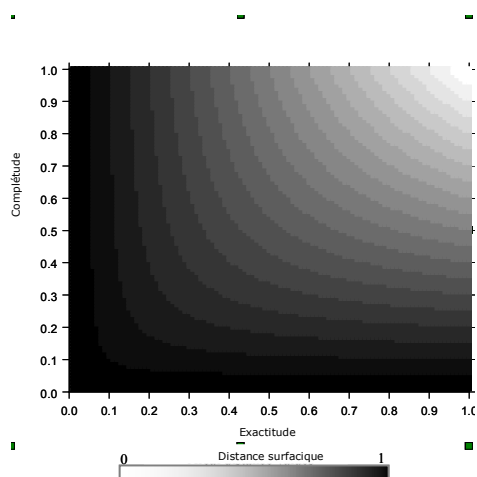


Figure III-7 : Distance surfacique en fonction de l'exactitude et de la complétude

### III.2.1.2.3. Liens particuliers

Lors de l'établissement des liens multiples, il se trouve que l'on rencontre des parties du graphe d'association qui se présentent comme des graphes bipartites complets. Ces parties du graphe donnent naissance à des liens particuliers qu'on appelle liens circulaires. Il est très important de détecter ces liens et de signaler leur présence, notamment si on se place dans le cas d'une application visant l'unification des bases de données ou la construction des serveurs multi-échelles pour la mise à jour, du fait que l'intervention sur une des entités participant au lien influe d'une manière conséquente sur toute la configuration. La figure III-8 illustre un exemple.

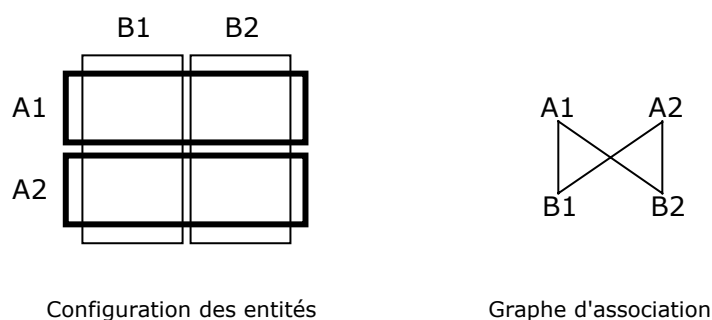


Figure III-8 : Exemple d'un lien circulaire

Hormis l'importance de signaler ce type de liens particuliers détectés à l'appariement, ils ne font pas l'objet d'un traitement spécifique dans la suite de ce chapitre.

Pour tester la robustesse de la méthode que nous avons développée, nous avons essayé d'appliquer la méthode pour appairer une multitude de jeux de données ayant des caractéristiques variées. Ces tests sont détaillés dans le paragraphe suivant.



### III.2.2. Tests de fiabilité de l'algorithme

Etant donné la spécificité des bases de données géographiques, le processus d'appariement de deux bases se confronte souvent à des conflits d'intégration. [Devoegele 1997] a établi une taxonomie relative aux problèmes d'intégration des BDG, répartie en six classes :

- ✓ Les conflits de sources de données qui apparaissent quand les types de sources de données employées ou les caractéristiques de celles-ci sont différentes;
- ✓ Les conflits d'hétérogénéité qui portent sur les critères globaux à définir pour chaque BDG (modèles de données, positionnement des éléments, modélisation de l'altitude, mode de représentation, méta-données liées à la géométrie et aux relations topologiques);
- ✓ Les conflits de définition des classes qui regroupent les problèmes liés à la classification des éléments des BDG, à leur spécification et au découpage des phénomènes du monde physique;
- ✓ Les conflits de structure qui surviennent lorsque les éléments en correspondance sont décrits par des concepts différents (classe, attribut, relation) ou lorsqu'une information gérée par la BDG correspond à une information qui doit être déduite;
- ✓ Les conflits de description sémantiques et géométriques qui résultent des différences entre les propriétés (attributs, méthodes) des classes en correspondance;
- ✓ Les conflits de données qui surviennent lorsque les objets en correspondance ont des valeurs différentes pour les attributs en correspondance.

Compte tenu de cette taxonomie et dans le but de tester la fiabilité de la méthode développée, plusieurs jeux de données issues de bases de données différentes et variées ont été utilisés. Ces tests sont réalisés sur :

❶ Un jeu de données composé de deux couches décrivant le thème du "Bâti" sur la ville d'Angers, acquises à deux dates différentes (1994 et 1996) en utilisant les mêmes procédés d'acquisition et les mêmes spécifications de saisies.

❷ Un jeu de données composé de deux couches extraites respectivement de la BDTopo® et du Cadastre représentant le thème "bâti" d'une zone pavillonnaire située dans la ville de Lyon.

❸ Un jeu de données composé de deux couches extraites respectivement de la BDTopo® et du Cadastre représentant le thème "bâti" sur la région de Cesson.

❹ Un jeu de données représentant l'occupation du sol de la BDTopo® dans la région de Bédarieux. Ce jeu de données est composé de deux couches obtenues par stéréo-restitution à partir de photographies aériennes en utilisant les mêmes spécifications de contenu et de saisie, mais il est réalisé par deux photo-interprètes différents.

⑤ Un jeu de données composé de deux couches extraites de la base de données "Forêt" de la Tunisie. La première couche est réalisée en 1993 par la photo-interprétation des images SPOT multi-spectrale, la deuxième couche est réalisée en 1999 par la photo-interprétation des ortho-photographies aériennes. On fait remarquer que les spécifications de contenu de ces deux couches ne sont pas les mêmes.

Le tableau ci-après résume les caractéristiques des jeux de données utilisés pour les tests.

	Couches	Nombre d'objets simples	Emprise géographique	Thème.	Loc. Géog.
Jeu de données #1	BDTopo 1994	986	5 Km <sup>2</sup>	Bâtiment quelconque	Ville d'Angers
	BDTopo 1996	1156			
Jeu de données #2	Cadastré	3083	5 Km <sup>2</sup>	Bâtiment	Lyon
	BDTopo	1183			
Jeu de données #3	Cadastré	3818	10.5 Km <sup>2</sup>	Bâtiment	Cesson
	BDTopo	2129			
Jeu de données #4	BDTopo	726	14 Km <sup>2</sup>	Occupation du sol	Bédarieux
	BDtopo	269			
Jeu de données #5	Forêt 1993 – Images Satellitales	148	60 Km <sup>2</sup>	Forêt	Metline Tunisie
	Forêt 1999 – Photographies aérienne	346			

Tableau III-1 : Données utilisées

Les tests effectués sur ces jeux de données sont classés selon quatre catégories :

- ✓ Le premier test consiste à appairer des jeux de données ayant les mêmes spécifications de contenu et de saisie mais à deux actualités différentes. Le jeu de données #1 sera utilisé pour ce test.
- ✓ Le deuxième test consiste à appairer des jeux de données ayant la même échelle de saisie mais n'ayant pas les mêmes spécifications. Pour ce test, les jeux de données #2 et #3 représentant le Cadastre et la BDTopo® seront utilisés.
- ✓ Le troisième test consiste en l'utilisation d'un jeu de données saisi avec les mêmes spécifications de contenu et de saisie, mais réalisé par deux photo-interprètes différents (jeu de données #4). Le but recherché par ce test est d'étudier l'erreur induite par l'interprétation de l'être humain.
- ✓ Le quatrième test consiste à appairer deux jeux de données représentant le même thème mais étant acquis à deux échelles différentes avec des spécifications différentes, et à deux actualités différentes (jeu de données #5). Ce test est réalisé dans le but de voir la capacité de la méthode de détecter les mises à jour, d'une part, et de voir la possibilité d'une éventuelle intégration dans un serveur multi-échelles, d'autre part.

### III.2.2.1. Test #1 : appariement entre deux actualités différentes d'une même base de données

Ce jeu de données représente un extrait du thème "bâti quelconque" de la BDTopo® et est composé de deux couches représentant chacune le même thème à deux dates d'acquisition différentes (1994 et 1996). Pour ce genre d'application, le processus d'appariement est utile pour détecter les mises à jour possibles entre les deux jeux de données. L'appariement de ces deux jeux de données génère 967 liens d'appariement dont 949 liens simples et 18 liens multiples. Le nombre de liens simples est plus élevé que celui des liens multiples, ce qui explique que les deux jeux de données appariés ont en quelque sorte la même échelle.

Les fonctions d'exactitude et de complétude sont calculées pour chacun de ces liens. On signale que le jeu de données d'Angers1994 a été considéré comme jeu de référence et celui de 1996 est considéré comme le jeu de données "à contrôler".

Parmi les 949 liens simples, il existe 756 liens dont l'exactitude et la complétude valent l'unité, ce qui indique une égalité géométrique entre les entités participantes à ces liens. Ces liens seront exclus de l'analyse. Les valeurs des mesures des liens restants sont illustrées par le graphique en figure III-9 :

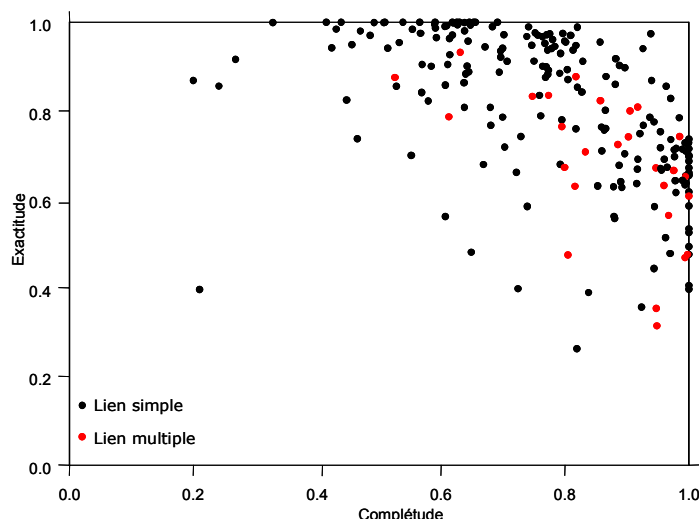


Figure III-9 : Exactitude vs. complétude des liens d'appariement relatifs au jeu de données #1

En adoptant une valeur de seuil égale à 0.5 pour l'exactitude et pour la complétude [Zaslavsky 1995], nous acceptons les liens qui respectent la condition suivante :

$$I(a \rightarrow b) > 0.5 \text{ ou } C(a \rightarrow b) > 0.5 \quad \text{[III-5]}$$

En se basant sur l'équation III-3, cette condition peut s'écrire en fonction de la distance surfacique comme suit :

$$D_s(a,b) < 0.66 \quad \text{[III-6]}$$

On détecte, pour ce jeu de données, un seul lien dont les valeurs ne respectent pas cette condition ( $I(a \rightarrow b) = 0.396$ ,  $C(a \rightarrow b) = 0.21$  et  $D_s(a,b) = 0.841$ ). Cependant, ce lien (cf. figure III-10(a)) sera signalé sans être supprimé du processus. On retrouve également à l'issue du processus d'appariement les objets des deux couches qui n'ont pas de correspondants. 158 objets de la couche Angers1996 n'ont pas d'homologues et un seul objet de la couche d'Angers1994 n'a aucun correspondant. L'objet de la couche de

1994 (cf. figure III-10(b)) a participé initialement à la construction des liens d'associations, mais il a été supprimé par le filtre de la surface minimale d'intersection.

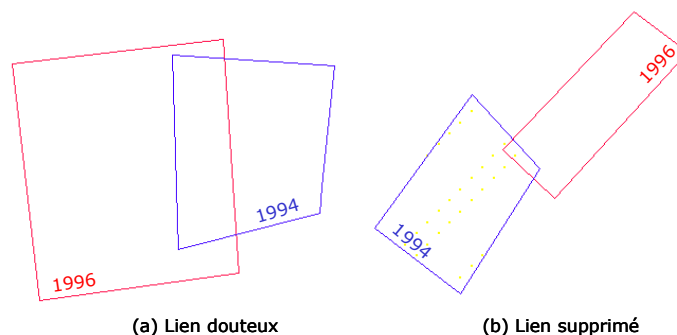


Figure III-10 : exemples de liens d'appariement douteux

La validation des liens se fait d'une manière visuelle en vérifiant que les entités appariées occupent une partie commune de l'espace géographique et ont des formes presque semblables. A ce niveau la validation n'est que quantitative, les liens ne sont pas validés (ou invalidés) de manière définitive. La validation définitive est faite après l'établissement des mesures de forme et de position. En effet, le lien que nous qualifions de "douteux"<sup>26</sup> sera analysé en se basant sur des mesures de forme et de position (dans la suite de l'étude) afin de le valider ou de le rejeter définitivement. Donc, pour cet exemple, nous pouvons déduire que le processus d'appariement a réussi à appairer la totalité (à l'exception de l'exemple de la figure III-10(b)) des entités du jeu de données. Soit 99.9% des entités des deux couches sont appariées d'une manière complètement automatique.

### III.2.2.2. test #2 : appariement des données n'ayant pas les mêmes spécifications

Le présent test est réalisé sur un jeu de données composé de deux couches extraites de la BDTopo® et du Cadastre décrivant le thème du "bâti". Il est à noter que les couches ne présentent pas les mêmes spécifications de contenu, ni les mêmes spécifications de saisie (cf. figure III-11).

<sup>26</sup> Un lien est qualifié de douteux si les entités qui participent à sa formation présentent un fort biais ou bien l'une des entités est incluse dans l'autre avec une différence conséquente de leur surfaces respectives.

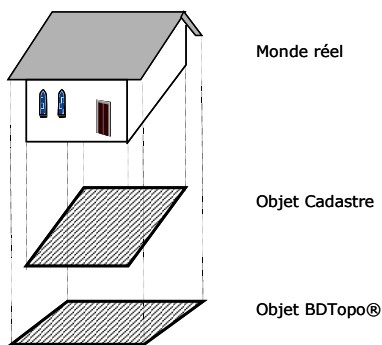


Figure III-11 : Spécification de saisie d'un bâtiment dans le Cadastre et dans la BDTopo

Pour cet exemple, le Cadastre est considéré comme le jeu de données de référence et le jeu de données BDTopo® est considéré comme celui à contrôler. A l'issue du processus d'appariement, on a obtenu 1067 liens d'appariement répartis comme suit : 387 liens simples de type 1-à-1 et 680 liens multiples de type n-à-m. Le nombre élevé des liens multiples indique que les deux bases de données ne présentent pas la même granularité. Les histogrammes des mesures de l'exactitude et de la complétude sont illustrés dans la figure III-12 :

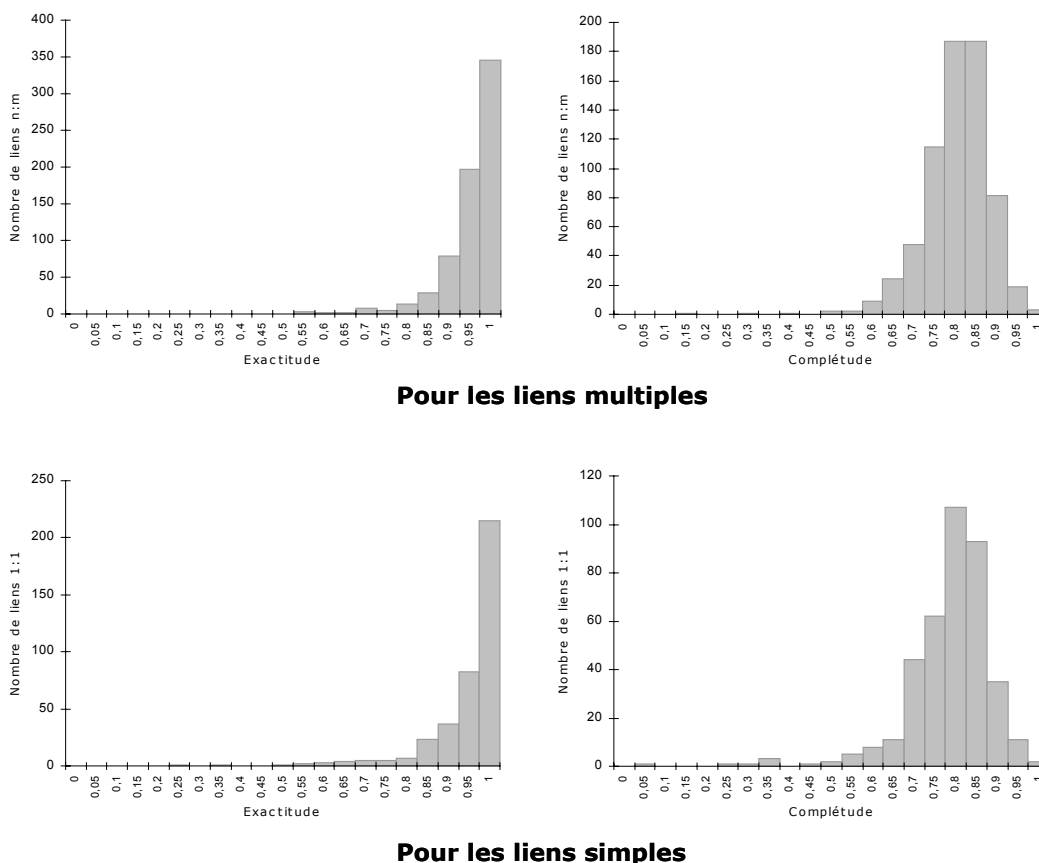


Figure III-12 : Histogrammes de l'exactitude et de la complétude -jeu de données Cadastre-BDTopo-

Les histogrammes de l'exactitude montrent l'existence d'un grand nombre de liens (560 liens soit 52.5% du total des liens d'appariement) ayant une exactitude supérieure à

0.95. Cette constatation confirme le fait que les deux jeux de données n'aient pas les mêmes spécifications de saisie, et notamment la règle que nous avons illustrée par la figure III-11. On remarque aussi que la croissance des histogrammes commence à être sans interruption à partir de la valeur 0.5 pour les deux paramètres de la "Détermination".

Ces tendances ont été remarquées en effectuant les tests d'appariement sur le jeu de données #3 représentant le thème du "bâti" du Cadastre et de la BDTopo®. Les histogrammes de l'exactitude et de la complétude sont illustrés dans la figure suivante :

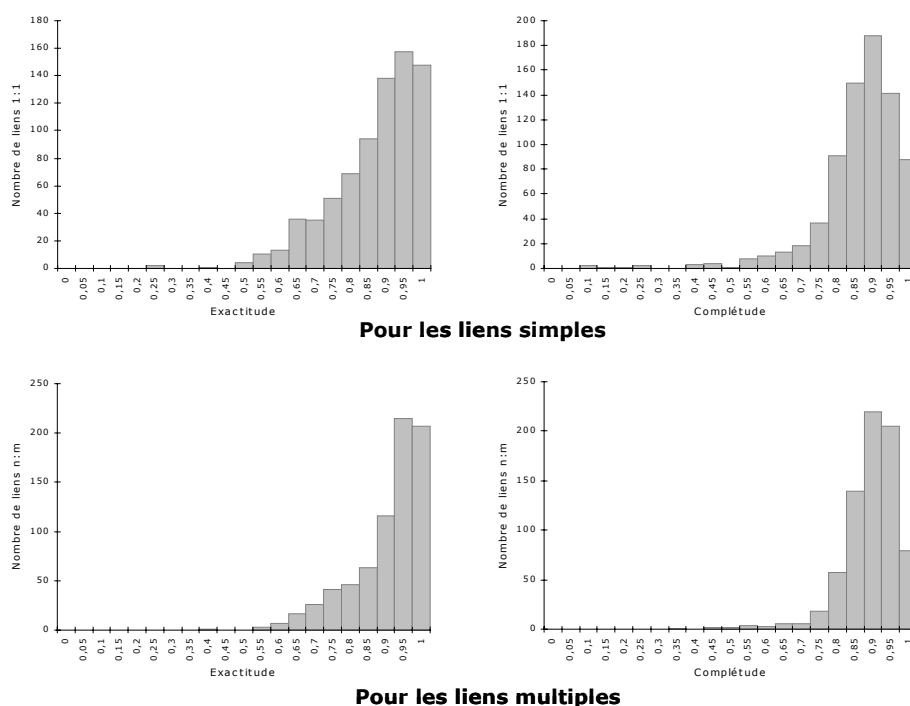


Figure III-13 : Histogrammes de l'exactitude et de la complétude relatifs au jeu de données #3

Sur les histogrammes, on remarque la présence d'un nombre conséquent de liens dont la valeur de l'exactitude pour les liens simples et multiples est supérieure à 0.95. Cela exprime une inclusion totale des bâtiments du Cadastre dans les bâtiments de la BDTopo® et permet de vérifier la différence de spécifications évoquée précédemment (cf. figure III-11).

Pour la fixation d'un seuil de filtrage sur les valeurs de l'exactitude et de la complétude, nous avons augmenté la valeur de seuil de coupure d'une manière progressive afin de dénombrer les liens supprimés à chaque valeur de seuil. Les résultats obtenus pour les jeux de données 1, 2 et 3 sont illustrés dans la figure III-14.

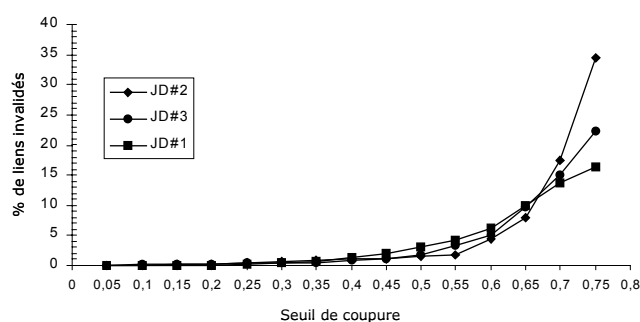


Figure III-14 : Nombre de liens invalidés vs. seuil de coupure

A chaque valeur de filtre, les liens invalidés sont visualisés pour s'assurer que les entités mises en correspondance par ces liens ne représentent pas le même objet géographique. La figure III-14 montre que les trois courbes obtenues présentent une tendance comparable. Le nombre des liens invalidés croît d'une manière exponentielle avec la croissance du seuil de coupure. Pour ce type de données, le décollage des courbes est constaté pour les trois jeux de données à des valeurs du seuil allant de 0.2 à 0.35. Pour les données représentant le thème "bâti", et suite à une validation visuelle, nous fixons le seuil de coupure sur les valeurs de l'exactitude et de la complétude à 0.3. Par conséquent, tous les liens d'appariement qui répondent à la condition suivante seront invalidés :

$$(I(a \rightarrow b) < 0.3) \text{ ou }^{27} (C(a \rightarrow b)) < 0.3 \quad [III-7]$$

L'application de cette condition sur les liens d'appariement établis entre les couches du jeu de données #2 invalide 5 liens, soit 0.47% du nombre total des liens. Parmi les 5 liens invalidés, 4 sont des liens simples. Les entités participantes à ces liens sont illustrées par la figure III-15. Ce type de liens concerne une configuration bien particulière. Les entités, faisant partie de ce type de lien, présentent une inclusion presque totale de l'une dans l'autre avec une différence significative au niveau de leurs surfaces respectives. Parmi ces liens, on peut distinguer deux catégories : l'inclusion des entités du premier jeu dans le deuxième (exemples 1-2 et 4 de la figure III-15) ou bien l'inverse (exemple 3 de la figure III-15). D'une manière générale, la distinction de ces catégories se fait en se basant sur l'interprétation des valeurs de l'exactitude et de la complétude d'une manière séparée et en se basant sur la classification de la figure III-4.

<sup>27</sup> L'utilisation du "ou" dans cette condition est pour s'abstraire de la notion d'un jeu de données de référence. Par conséquent la complétude  $C(a \rightarrow b)$  sera considérée comme l'exactitude dans le sens inverse, c'est à dire,  $C(a \rightarrow b) = I(b \rightarrow a)$

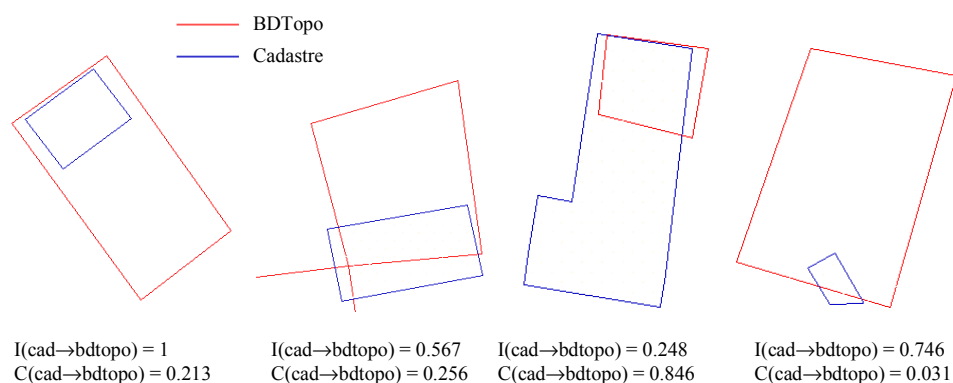


Figure III-15 : Exemples de liens invalidés

Ces liens ne seront pas supprimés définitivement, mais ils seront stockés dans une structure à part et ils seront signalés à l'utilisateur. La structure de stockage des liens d'appariement est détaillée dans le §III.2.3.

En utilisant le jeu de données #2 (Cadastre/BDTopo®), un test a été engagé consistant à voir le comportement de l'exactitude et de la complétude en la présence d'un biais généralisé progressif selon l'un des axes. Les entités des deux jeux ont été appariées en formant 128 liens (56 simples et 72 multiples). Le test consiste à déplacer le jeu de données du Cadastre selon l'axe des abscisses d'une manière progressive et de l'apparier avec le jeu de données BDTopo à chaque déplacement. A chaque appariement, les valeurs moyennes de l'exactitude et de la complétude sont calculées. Ces valeurs sont représentées sur la figure suivante :

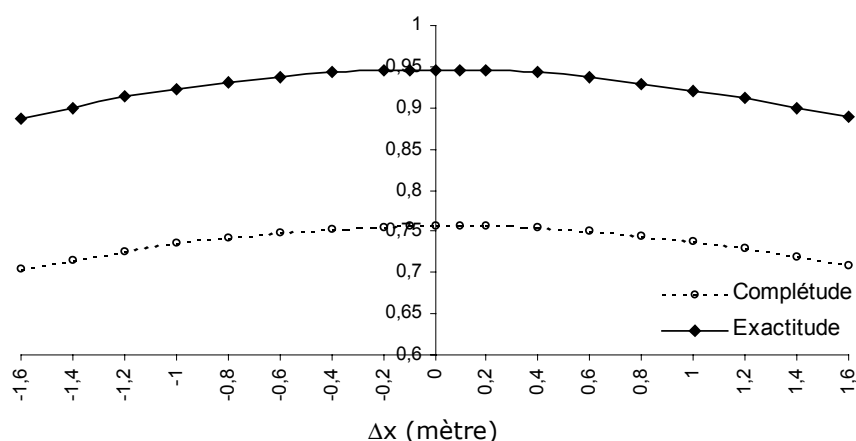


Figure III-16 : Evolution de l'exactitude et de la complétude en la présence d'un biais généralisé

On signale que, même en la présence du biais, les mêmes liens d'appariement entre les entités ont été conservés. Les deux paramètres montrent une décroissance quand le biais croît (en valeur absolue). On constate aussi que, même en la présence d'un biais conséquent (1.6 mètres<sup>28</sup>), la variation des deux paramètres de la "Détermination" est minimale (aux alentours de 0.05). Ce comportement indique une

<sup>28</sup> Cette valeur est choisie parce qu'elle approche l'erreur moyenne quadratique de la BDTopo.

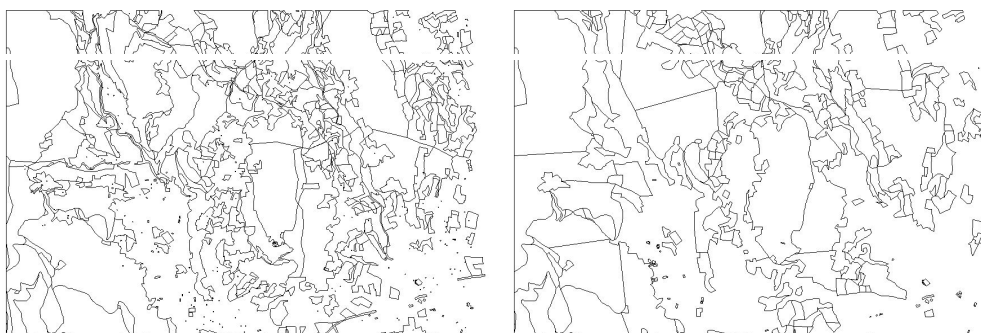


robustesse de ces deux indicateurs en la présence d'un biais généralisé, puisque, même à forte valeur de biais, ils présentent une variation faible. Par conséquent leur aptitude à détecter les liens d'appariement valides reste intacte.

### III.2.2.3. test #3 : appariement des données à pavage "presque" complet et ayant les mêmes spécifications

Pour effectuer ce test, nous utilisons un autre type de jeux de données surfaciques représentant une couverture presque complète de l'espace géographique. Ces jeux de données représentent essentiellement des données de l'occupation du sol. L'espace est donc représenté sous la forme d'entités surfaciques partageant les mêmes frontières, mais ayant des thématiques différentes. Les entités sont collées les unes aux autres comme des pavés, d'où l'appellation de pavage complet (si tout l'espace est couvert) et presque complet (s'il existe une partie de l'espace qui n'est pas porteuse de données).

Les données utilisées représentent deux couches de l'occupation de sol, saisies par deux photo-interprètes en utilisant les mêmes spécifications et les mêmes photographies aériennes à la même actualité. Chaque jeu de données est composé de 12 thèmes (cf. figure III-17) dont le pavage n'est pas complet. Une première analyse visuelle montre que le premier photo-interprète s'est appliqué à saisir plus de détails que le second d'une part (le premier jeu de données est composé de 726 objets et le second de 264 objets), et que, d'autre part, les limites définies des quelques entités, représentant notamment le thème bois, ne sont pas les mêmes.



(a) Réalisé par le photo-interprète #1 (b) Réalisé par le photo-interprète #2

Figure III-17 : Deux saisies de l'occupation de sol - BDTopo - Bedarieux

La méthode de l'appariement a été appliquée sur ce jeu de données avec l'enchaînement décrit dans le §III.2. Il en résulte 165 liens d'appariement dont 103 liens simples de type 1-à-1 et 62 liens multiples de type n-à-m. Les liens multiples sont composés en majeure partie par des liens de type 1:n ou n:1 (55 liens sur 62). Ce type de configuration de liens souligne une différence au niveau de l'appréciation faite par chacun des photointerprètes, ce qui soulève des problèmes liés aux niveaux de l'agrégation sémantique. La figure III-18 illustre un exemple.

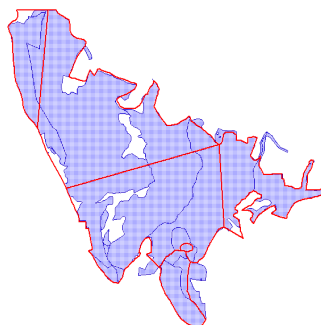


Figure III-18 : Exemple d'un appariement multiple de type 17:5

Etant donnée la spécificité de ce type de données, il est judicieux de chercher un seuil de coupure approprié et de ne pas généraliser l'utilisation de la valeur trouvée lors des tests précédents. Nous avons adopté la même méthode que précédemment pour l'établissement d'un seuil de coupure sur les valeurs de l'exactitude et de la complétude. Etant donnée la symétrie qui existe entre les deux jeux de données, l'analyse du seuil sera réalisée sur les valeurs de la distance surfacique. Nous en déduirons par la suite le seuil sur les valeurs de l'exactitude et de la complétude en utilisant la relation donnée par l'équation III-4. A cet effet, nous faisons varier le seuil de 0.05 à 0.95 et nous vérifions visuellement la validité des liens répondant à la condition. La figure III-19 représente l'histogramme de la distance surfacique, ainsi que le pourcentage cumulé du nombre de liens. On observe sur l'histogramme un pic matérialisant les liens ayant une distance surfacique supérieure à 0.95. Une analyse visuelle montre que ces liens mettent en correspondance des entités représentant des constructions légères avec des parcelles de bois ou de broussailles. Ces liens peuvent être imputés à un oubli commis par le photointerprète #2. Celui ci ayant omis de saisir les constructions légères dans la base de données.

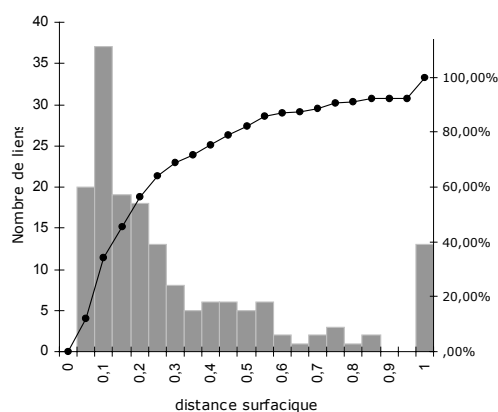


Figure III-19: Histogramme de la distance surfacique (barres grises) et pourcentage cumulé (points noirs)

On observe également sur la figure III-19 que la décroissance de l'histogramme n'est pas monotone, puisqu'il existe deux paliers, l'un situé au niveau des valeurs entre 0.4 et 0.6 et l'autre pour les liens ayant une distance surfacique supérieure à 0.6. Le premier palier des liens représente les correspondances entre des entités de même

thème occupant localement une partie commune de l'espace. Ces liens peuvent être imputés à une mauvaise interprétation de la part de l'opérateur de saisie (cf figure III-20(a)). Le deuxième palier est interprété de la même façon, mais la différence entre les surfaces des entités appariées est plus conséquente (cf. figure III-20(b)). Nous ne pouvons pas généraliser cette remarque du fait que l'appariement n'a été fait que sur un seul jeu de données de ce type. Par conséquent, ce comportement ne peut être validé que par d'autres tests sur des données de même type.

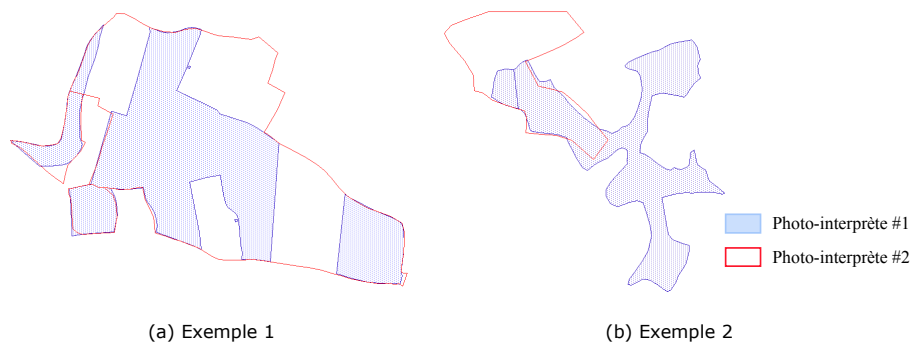


Figure III-20 : Exemples de mauvaise interprétation

La divergence au niveau de l'interprétation des limites des entités est également présente, même pour des liens dont la valeur de la distance surfacique est inférieure à 0.4. Mais, la différence entre les surfaces des entités appariées est de moindre amplitude. Ainsi, nous avons choisi de fixer le seuil sur les valeurs de la distance surfacique à 0.6. Cela signifie que tous les liens de valeurs de distance surfacique inférieure à 0.6 seront considérés comme des liens valides. Cette condition peut être également exprimée en fonction de l'exactitude et de la complétude de la manière suivante :

Lien valide si  $I(a \rightarrow b) > 0.57$  ou  $C(a \rightarrow b) > 0.57$

L'application de cette condition valide 143 liens soit 86.7% des 165 liens initialement établis. Ces 143 liens subiront d'autres mesures (cf. §II.4) afin d'affiner le résultat de l'appariement et de contrôler la qualité des entités participant à leur formation. Par ailleurs, les liens invalidés ne seront pas supprimés d'une manière définitive du processus, puisque les entités participant à leur formation subiront une étape de mesure pour aboutir à une décision finale sur la validation ou l'invalidation du lien.

#### **III.2.2.4. test #4 : appariement des données à pavage complet avec des spécifications différentes, sources de saisie différentes et actualité différente**

Le présent test consiste à mettre en correspondance deux jeux de données qui n'ont pas les mêmes spécifications, ni les mêmes sources de saisie, ni la même actualité. Mais, ils représentent la même thématique sur une même zone géographique. Les jeux

de données utilisés sont extraits d'une base de données représentant le thème forêt<sup>29</sup>. Le premier jeu de données a été réalisé par une photo-interprétation des images SPOT<sup>30</sup> multi-spectrales datées de 1993 et corrigées au niveau 2B. Ce type de correction géométrique d'images n'inclut pas l'effet du relief et la rectification de l'image ne se fait que d'une manière planimétrique par l'utilisation d'un modèle polynomial. Le second jeu de données a été réalisé en 1999 par une photo-interprétation des ortho-photographies aériennes en utilisant de nouvelles spécifications de la base de données. Des extraits de ces deux jeux de données sont illustrés par la figure III-21. Le but visé par ce test est de voir la limite de la méthode lors de l'appariement de ce genre de données.

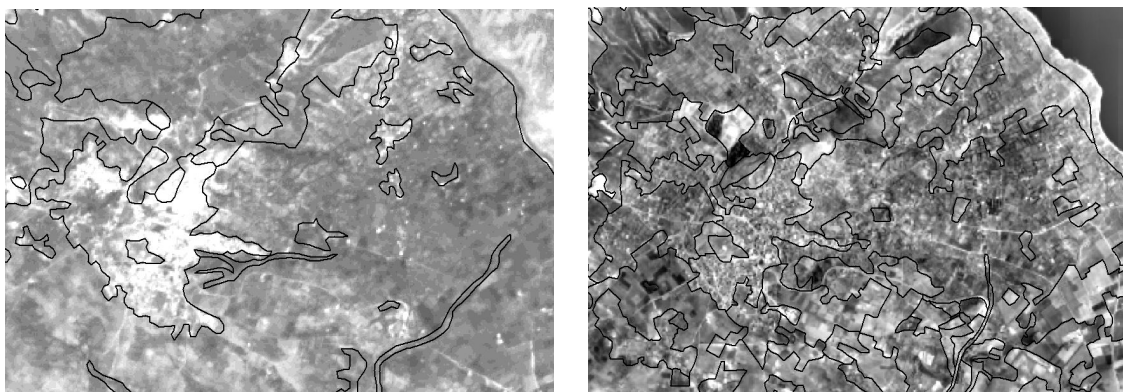


Photo-interprétation réalisée à partir d'une image SPOT-XS corrigée au niveau 2B -1993-

Photo-interprétation réalisée à partir d'une ortho-photo -1999-

**Figure III-21 : Extraits des jeux de données utilisés pour le test #4**

Une première analyse visuelle des jeux de données, montre que le jeu de données de 1999 est plus détaillé que celui de 1993. Ceci est logique, du fait que les échelles des sources de saisie ne sont pas comparables et que les spécifications de la base de données 1999 prévoient un éclatement au niveau des postes de légende. Les deux sources de données n'ont pas la même résolution planimétrique (20 mètres pour les images SPOT et 2 mètres pour les ortho-photos), ce qui induit une erreur géométrique au niveau des contours des entités. D'autre part, elles n'ont pas la même résolution radiométrique. La base de données 1999 est donc plus détaillée que celle de 1993. Ces deux différences des deux sources de saisies sont à l'origine de la génération des intersections parasites lors du croisement des deux jeux de données, et, par conséquent, à l'origine de la génération des liens d'appariement multiples. Ce type de problème est également rencontré dans le domaine de fusion de données d'occupation de sol [Jones & al. 2000]. La méthode de résolution du problème est proche de celle de [Chrisman & Lester 1991] qui utilise des indicateurs simples de type compacité, élongation, mesure de la surface des polygones, etc.

La figure III-21 (Image Spot -au centre et à gauche-) montre un effet de généralisation au niveau de l'interprétation. Le polygone entourant la parcelle blanche représente (dans la base de données) un espace urbain, or il existe à l'intérieur de ce

<sup>29</sup> Bases de données construites par le centre national de télédétection tunisien et détenues par le ministère de l'agriculture tunisien

<sup>30</sup> SPOT: Satellite français d'observation de la terre

polygone des parcelles grises qui représentent des exploitations agricoles que le photo-interprète n'a pas jugé utile de différencier en les laissant partie intégrante de la zone urbaine. La distinction entre ces deux thèmes est faite pour la base de données 1999 (cf. figure III-21 - ortho-photo - ). Les deux jeux de données utilisés pour ce test sont illustrés par la figure suivante :

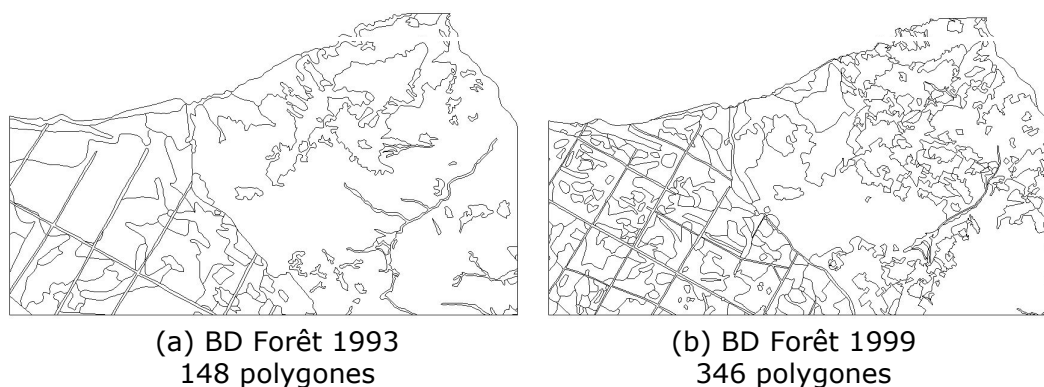


Figure III-22 : BD forestière utilisées pour le test #4

Les deux jeux de données n'ont pas été saisis selon les mêmes spécifications et ils n'ont pas le même niveau de détails. Ils présentent un décalage géométrique entre les contours des entités. Etant données ces deux contraintes et que les deux jeux assurent un pavage complet de l'espace, il se trouve que l'appariement géométrique entre ces deux jeux s'avère problématique, puisqu'il y a une génération d'un nombre important de liens indésirables. Par conséquent, il est nécessaire de redéfinir les valeurs de seuil pour supprimer les liens de correspondances de type 1-à-1 inutiles. L'utilisation de la valeur 0.2 (cf. §III.2.1.1) comme filtre pour supprimer les liens parasites a donné lieu à un appariement entre les deux jeux comprenant 8 liens seulement, dont 4 simples et 4 multiples. Parmi les liens multiples, il existe un lien qui met en correspondance 307 polygones de la base de données 1999 à 127 polygones de la base de données 1993, ce qui s'apparente à une situation où tous les polygones d'un jeu de données sont appariés avec tous les polygones de l'autre jeu de données.

Ce résultat montre bien l'insuffisance de la valeur définie pour le seuil de suppression des liens parasites pour ce type de données. Par conséquent, la recherche d'une nouvelle valeur s'impose. Pour ce faire, nous avons augmenté la valeur de ce seuil progressivement et analysé le résultat obtenu d'une manière visuelle. L'augmentation de la valeur de seuil aura pour effet l'éclatement des gros liens multiples en un ensemble de liens simples et d'autres liens multiples de moindre taille. La figure III-23 donne le nombre de liens d'appariement générés selon la valeur de seuil utilisée :

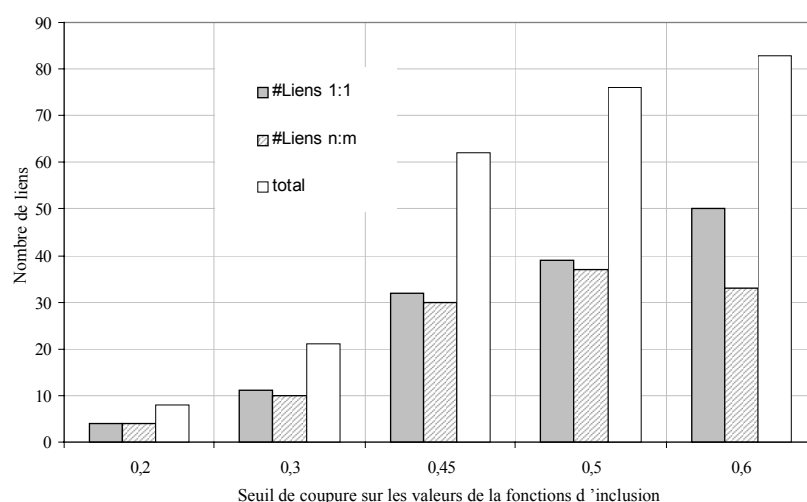


Figure III-23 : Répartition des liens selon le seuil de coupure utilisé

Nous rappelons que dans le cadre de ce travail, l'appariement entre les jeux de données repose uniquement sur l'information géométrique. Par conséquent, aucune information sémantique n'a été utilisée pour guider le processus de mise en correspondance.

Par ailleurs, l'information sémantique a été utilisée pour chaque résultat d'appariement afin de générer des matrices de confusion dans le but de voir l'impact du choix du seuil sur ces matrices. Etant donné la différence de spécification et l'éclatement des postes de légende, il paraît très difficile d'identifier entre les deux jeux de données les thèmes désignant le même phénomène. A cet effet, aucun travail d'unification des légendes n'a été entrepris en amont de l'appariement. En effet, pour affiner l'analyse, nous préconisons la génération des matrices de confusion et l'analyse des fortes valeurs d'accord entre les thèmes, afin de pouvoir établir une correspondance sémantique entre les thèmes appariés. L'analyse des matrices de confusion permettra également d'assister le choix d'un seuil de coupure en maximisant les taux d'accord.

Dans le cadre de ce test, le choix de la valeur 0.45 est déterminé d'une manière visuelle en analysant les liens après l'application des différents seuils de coupure. Cette conclusion rejoint celle faite par [Bel Hadj Ali 1997] pour appairer des jeux de données issues de la BDCarto® de l'IGN et de la base de données européenne Corine-LandCover.

Les liens multiples ont été établis en maximisant l'exactitude et la complétude de chacun d'entre eux. Cependant, le résultat de l'appariement révèle l'existence d'un ensemble de liens (39 liens parmi un ensemble de 62 liens) dont l'exactitude ou la complétude ont des basses valeurs (inférieures à 0.5). Les entités participant à la formation de ces liens présentent une différence conséquente au niveau de leur géométrie. La figure III-24 illustre quelques-uns des liens invalidés.

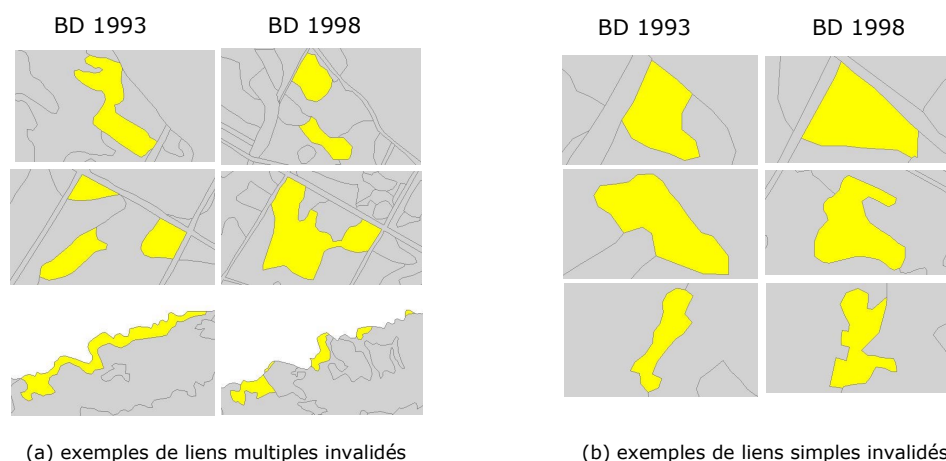


Figure III-24 : Exemples de liens invalidés

Il est à noter que les liens invalidés représentent en partie une réalité physique. Leur invalidation est essentiellement due à une forte différence, soit dans l'interprétation des limites des thèmes, soit dans l'interprétation de la nature du thème elle-même.

### III.2.2.5. Critique des tests

La méthode d'appariement que nous avons développée dans le cadre de ce travail s'appuie essentiellement sur l'utilisation de l'information géométrique de la mise en correspondance entre deux bases de données. L'appariement géométrique est considéré parmi les plus difficiles à mettre en œuvre [Lemarié 1996; Devogele 1997; Badard 2000]. Les résultats obtenus en appariant différents jeux de données par l'utilisation de cette méthode, et notamment sur des jeux de données "comparables", montrent la robustesse de la méthode. Pour des jeux de données ayant la même granularité, en l'occurrence les jeux de données représentant le thème du "bâtiment", le résultat de l'appariement automatique avoisine 90% des entités mises en correspondance de part et d'autre.

Les tests réalisés sur les jeux de données de l'occupation du sol ont montré également la robustesse de la méthode, dans le cas où les bases de données auraient la même granularité, voire les mêmes spécifications. Cela est démontré par le test effectué sur les jeux de données "occupation du sol" extraits de la BDTopo®. Le test effectué sur la base de données "forestière" a montré les limites de l'appariement géométrique, dans le cas où les deux bases de données n'auraient ni la même échelle, ni les mêmes sources de saisie, ni les mêmes spécifications. Pour aboutir à un appariement reflétant la réalité, il est nécessaire de procéder *a priori* par une étape d'unification des schémas des deux bases, puis de supprimer les liens 1-à-1 indésirables avant de procéder à l'établissement des liens multiples par l'utilisation de l'information sémantique. L'utilisation d'une telle démarche peut paraître bénéfique dans le cas d'une mise en œuvre d'un serveur multi-échelles, mais, en contre partie, elle peut biaiser les résultats d'un contrôle de la qualité sémantique.

La méthode de l'appariement précédemment décrite a été mise en œuvre sur la plate-forme *ArcView*. Les résultats de l'appariement sont stockés dans des tables *DBase*. Le prototype est présenté dans la section suivante.

### III.2.3. Présentation du prototype réalisé

Pour la mise en œuvre de la méthode d'appariement, nous avons opté pour le stockage de toutes les étapes de la méthode, notamment lors des phases de filtrage. Ce choix est pris dans l'objectif de conserver tous les liens d'appariement, même ceux qui sont considérés comme indésirables pour une exploitation et une analyse ultérieure, si le processus d'appariement devait être repris. Chaque jeu de données est identifié par une table sémantique (JD1 et JD2) dans laquelle chaque entité surfacique est reconnue par un identifiant unique. Les liens supprimés lors des différents stades de filtrage sont stockés dans la table "LIENSUPP" en indiquant le filtre qui les a écartés. Le reste des liens valides figure dans une table nommée "TABAPP". Le résultat de la détection des liens multiples est stocké dans l'attribut "IDAPPMLT". Les mesures de disparité de forme et de position entre les entités appariées par les métriques présentées dans le chapitre II sont stockées dans la table "MESURES". Les entités d'un jeu de données qui n'ont pas de correspondants dans l'autre jeu de données, et réciproquement, sont stockées dans deux tables nommées "LIENS10" et "LIENS01". Toutes ces tables sont liées entre elles comme le montre le modèle de stockage donné par la figure III-25.

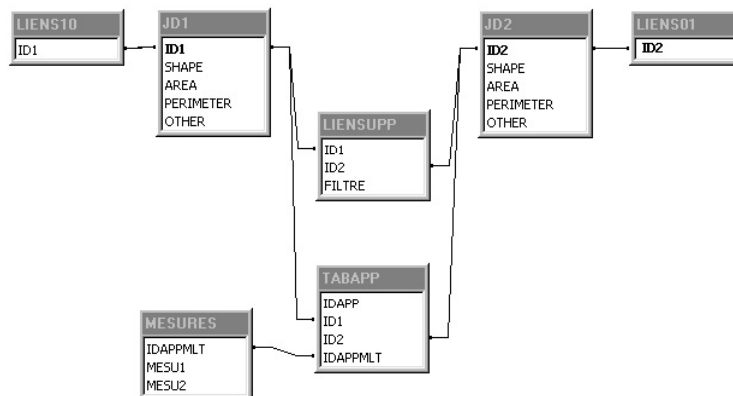


Figure III-25 : Modèle de stockage des résultats de l'appariement

Le prototype réalisé est illustré en annexes C et D.

A l'issue du processus d'appariement, les entités représentant le même phénomène physique sont mises en correspondance, sans pour autant vérifier leur exactitude géométrique en terme de forme et de position. A cet effet, toutes les entités participant à un lien d'appariement jugé valide subiront l'ensemble des mesures décrites dans le chapitre II pour qualifier leur différence de forme et de position.

Nous présentons, dans la suite de ce chapitre, une méthode permettant d'analyser l'ensemble des mesures effectuées sur les entités appariées. Cette analyse aura pour but de classer les liens selon la différence de la forme et de la position des entités qui



participent à leur formation. Des tests sur les jeux de données précédemment décrits (en §III.2.2) seront également présentés pour montrer la validité de la méthode.

### III.3. ANALYSE DES MESURES

A cette étape de l'étude, la méthode de l'appariement présentée permet la mise en correspondance entre les entités des deux bases de données. La mise en correspondance des entités et la vérification de la validité de leurs liens d'appariement sont réalisées en se basant sur les taux d'inclusion des entités les unes dans les autres.

Après leur mise en correspondance, les entités seront classées en entités simples ou complexes selon les cardinalités de leurs liens. Puis, elles subiront les mesures définies dans le chapitre II en les représentant dans les espaces appropriés. Les mesures utilisées sont résumées dans le tableau suivant :

Mesure	Désignation de la mesure	Entités	
		Simple	Complexes
Ds	Distance surfacique	✓	✓
Dh	Distance de Hausdorff	✓	✓
Dcm	Distance entre les centres de masse	✓	✓
Dfa	Distance entre fonctions angulaires	✓	
Rotr	Rotation relative entre les entités	✓	✓
Dsp	Distance entre les signatures polygonales	✓	
Imd	Indice moyen de dilatation –signature polygonale-	✓	
Dmgi	Distance entre les moments géométriques invariants	✓	✓
Dleg	Distance entre les moments de Legendre	✓	✓
Dzer	Distance entre les moments de Zernike	✓	✓

Tableau III-2: Distances et indicateurs utilisés

Les distances présentées dans le tableau III-2 ont été testées d'une manière individuelle au chapitre II sans tester l'existence d'une éventuelle corrélation entre elles. L'étude de la corrélation entre les distances et les indicateurs permet de réduire le nombre des métriques utilisées et d'alléger la tâche de l'utilisateur dans le cas de l'utilisation de ces métriques pour un contrôle de la qualité de forme et de position des entités surfaciques.

L'utilisation de l'ensemble de ces distances et indicateurs, pour qualifier l'exactitude de la forme et de la position des entités surfaciques, nécessite l'établissement des seuils et des limites sur les valeurs de ces métriques afin de classer les entités appariées selon les différences de leur forme et de leur position. La fixation des seuils d'une manière individuelle pour chacune des distances utilisées paraît une tâche fastidieuse et très complexe [Bel Hadj Ali & Vauglin 2000; Bel Hadj Ali 2000]. A cet effet, il est nécessaire d'utiliser les mesures d'une manière combinée, d'essayer de déterminer les seuils sur leurs valeurs d'une manière automatique. Cette méthode proposée est illustrée par l'organigramme suivant :

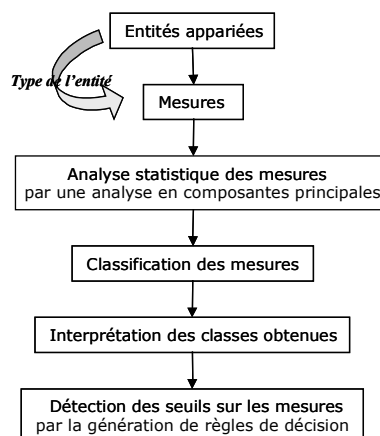


Figure III-26 : Méthode proposée pour l'analyse des mesures

Dans la suite de l'étude, toutes les étapes de la méthode seront expliquées par l'utilisation des mesures réalisées sur les entités simples appariées entre le Cadastre et la BDTopo® (cf. III.2.2.2.).

### III.3.1. Mesures et analyse statistique

L'exemple que nous traiterons dans ce chapitre est constitué d'un ensemble de 439 entités simples appariées représentant le thème du "bâti" entre le Cadastre et la BDTopo®. Six mesures ont été appliquées aux entités appariées, dont trois concernent des mesures de position et trois des mesures de forme. Les écarts de position sont calculés par l'utilisation de la distance surfacique, la distance entre les centres de masses et la distance de Hausdorff. Les différences de forme sont détectées par l'utilisation de la distance entre les fonctions angulaires, la distance entre les signatures polygonales et la distance entre les moments géométriques invariants. Comme nous l'avons indiqué précédemment, l'utilisation d'une seule mesure est insuffisante pour discriminer les écarts de position ou les différences de forme. Par exemple, pour une même distance surfacique, il est possible de trouver différentes configurations d'écart de position. Ainsi, l'utilisation d'une autre mesure, telle la distance de Hausdorff, par exemple, est nécessaire pour les différencier. Ce constat est illustré par quelques exemples (cf. figure III-27).

Le même constat est fait pour les mesures de forme en ce qui concerne la distance entre les moments géométriques invariants et la distance entre les signatures polygonales, par exemple. Il a été démontré en chapitre II que la distance entre les moments géométriques invariants est insensible aux détails affectant les contours des polygones. De plus, elle rend compte essentiellement de la forme globale de l'entité et la distance entre les signatures polygonales est beaucoup plus sensible à ce type de distorsions.

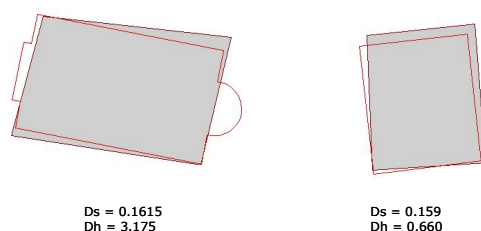


Figure III-27 : Exemple de l'insuffisance de l'utilisation d'une seule mesure pour quantifier les écarts de positions

Cependant, il est nécessaire d'analyser les corrélations entre les mesures utilisées pour éviter l'utilisation de deux mesures hautement corrélées, porteuses de la même information au niveau de la distorsion géométrique des entités. Les corrélations entre les 6 mesures sont données par le tableau suivant :

	Ds	Dh	Dcm	Dfa	Dsp	Dmgi
Ds	1,000	0,501	0,537	0,420	0,365	0,385
Dh		1,000	0,803	0,657	0,842	0,549
Dcm			1,000	0,459	0,603	0,396
Dfa				1,000	0,839	0,539
Dsp					1,000	0,595
Dmgi						1,000

Tableau III-3 : Corrélations entre les mesures

*Ds*: Distance surfacique - *Dh*: distance de Hausdorff - *Dcm*: distance entre les centres de masse - *Dfa*: distance entre les fonction angulaires - *Dsp*: distance entre les signatures polygonales - *Dmgi*: distance entre les moments géométriques invariants.

Le tableau III-3 montre l'existence d'une forte corrélation entre la distance de Hausdorff (*Dh*) et la distance entre les centres de masses des entités mesurées (*Dc*). Cette corrélation peut s'expliquer par la nature de chacune de ces deux mesures, puisqu'elles traduisent une mesure ponctuelle et localisée. Il existe également une forte corrélation entre la distance entre les fonctions angulaires (*Dfa*) et la distance entre les signatures polygonales (*Dsp*). Cette constatation permet de conclure que ces deux mesures de forme ont les mêmes performances pour la détermination des disparités de forme entre les entités surfaciques. A cet effet, l'utilisation de l'une ou l'autre de ces deux mesures doit être suffisante. Etant donnée que la distance de Hausdorff renvoie la distance maximale entre les deux points les plus éloignés d'une part, la distance entre les signatures polygonales s'appuie sur le calcul de la fonction des distances euclidiennes entre les points des deux contours. Il est naturel que les deux distances (*Dh* et *Dsp*) aient une forte corrélation.

Pour mieux comprendre le comportement de ces mesures, les unes par rapport aux autres, ainsi que leur contribution pour expliquer une configuration d'écart, nous allons employer la technique d'analyse dite "Analyse en Composantes Principales" (ACP) [Chamussy & al. 1994].

L'analyse en composantes principales consiste à transformer les mesures initialement obtenues, corrélées entre elles, en de nouveaux indices synthétiques qui sont globalement décorrélés et qui sont obtenus par une combinaison linéaire des mesures initiales. Parmi tous les indices possibles, l'ACP recherche d'abord celui qui

permet de voir au mieux les individus, c'est à dire, celui pour lequel la variance des individus est maximale. Cet indice est appelé première composante principale ou encore premier axe principal. Une certaine proportion de la variation totale des individus est expliquée par cette composante principale. Ensuite, une deuxième composante est calculée sous deux conditions :

- ✓ Avoir une corrélation nulle avec la première;
- ✓ Avoir à son tour la plus grande variance.

Le processus se déroule jusqu'à l'obtention de la  $p^{\text{ième}}$  et dernière composante principale. L'analyse en composantes principales permet également d'analyser les mesures initiales.

L'analyse en composantes principales permet la génération de 6 variables synthétiques qui sont exprimées en fonctions des mesures initiales comme suit :

$$\begin{aligned}
 \text{CP1} &= 0.32\text{Ds} + 0.47\text{Dh} + 0.4\text{Dcm} + 0.42\text{Dfa} + 0.46\text{Dsp} + 0.36\text{Dmgi} \\
 \text{CP2} &= -0.68\text{Ds} - 0.06\text{Dh} - 0.44\text{Dcm} + 0.38\text{Dfa} + 0.34\text{Dsp} + 0.3\text{Dmgi} \\
 \text{CP3} &= 0.5\text{Ds} - 0.32\text{Dh} - 0.42\text{Dcm} + 0.04\text{Dfa} - 0.21\text{Dsp} + 0.65\text{Dmgi} \\
 \text{CP4} &= 0.38\text{Ds} - 0.18\text{Dh} - 0.35\text{Dcm} + 0.59\text{Dfa} + 0.14\text{Dsp} - 0.58\text{Dmgi} \\
 \text{CP5} &= -0.18\text{Ds} - 0.54\text{Dh} + 0.57\text{Dcm} + 0.48\text{Dfa} - 0.34\text{Dsp} + 0.1\text{Dmgi} \\
 \text{CP6} &= -0.11\text{Ds} + 0.59\text{Dh} - 0.17\text{Dcm} + 0.34\text{Dfa} - 0.7\text{Dsp} + 0.02\text{Dmgi}
 \end{aligned}
 \tag{III-8}$$

Les équations montrent que les mesures initiales contribuent d'une manière équitable à la formation du premier plan principal qui est porteuse de 78% de l'information. En se référant à la représentation des mesures initiales sur le premier plan principal (cf. figure III-28), on retrouve que les trois distances initialement dédiées pour mesurer les différences de forme vont dans le même sens, ce qui montre encore une fois leurs vocations communes. Les distances affectées pour mesurer les écarts de position vont globalement dans la même direction avec de légères divergences, ce qui montre que chacune de ces mesures est porteuse d'une information que les autres sont incapables de discriminer.

Le deuxième plan principal (75% de l'information initiale) permet de montrer que la distance surfacique et la distance entre les moments géométriques invariants vont dans le même sens. Cette tendance paraît logique du fait que les mesures des aires des entités surfaciques contribuent d'une manière conséquente à la formation des moments géométriques invariants, d'une part, et que la distance surfacique est définie sur la base de mesures des aires, d'autre part.

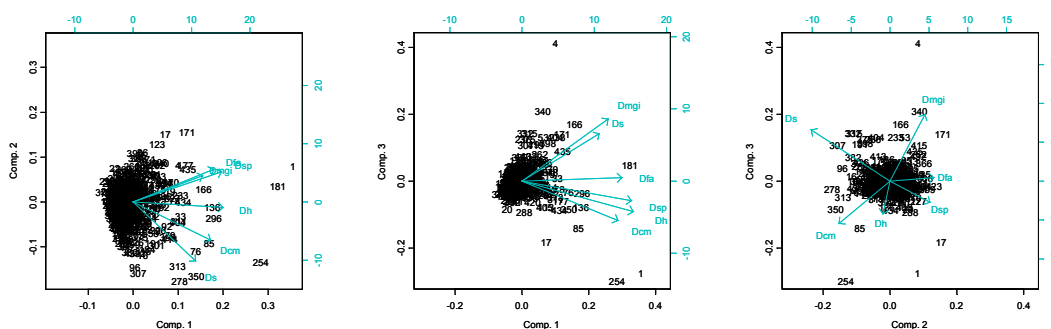


Figure III-28 : Représentation des mesures initiales sur les 3 plans principaux

Le graphique suivant illustre l'importance relative de chacune des composantes principales générées.

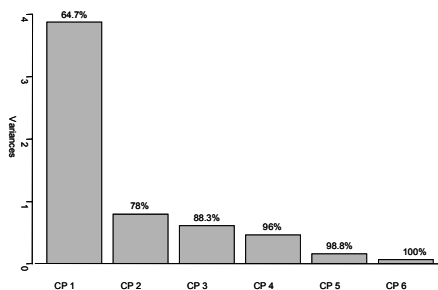


Figure III-29 : Importance des composantes principales

L'analyse de la variance des composantes principales montre que la première composante présente la plus grande variance (3.881) par rapport aux autres composantes. D'autre part, on remarque que la variance varie peu entre les 4 dernières composantes principales. L'analyse de la variance permet également d'écartier les composantes qui présentent une valeur faible de variance et dont on estime qu'elles ne véhiculent pas assez d'information. Dans le cas de l'exemple présent, nous pouvons nous limiter à l'utilisation des quatre premières composantes principales, d'autant plus qu'elles sont porteuses de 96 % de l'information initiale. L'analyse se réduira donc au traitement de 4 variables synthétiques, au lieu de l'analyse de 6 initialement prévues.

Les entités appariées sont donc représentées et identifiées par un quadruplet de mesures synthétiques que nous cherchons à ranger dans des classes ayant les mêmes caractéristiques en terme de différence de formes et de positions.

Pour classer les entités en fonction de leurs mesures respectives, plusieurs techniques de classifications sont envisageables. Ces techniques peuvent être classées selon deux catégories : classifications supervisées et classification non supervisées.

- ✓ Les classifications supervisées : ce type de classification nécessite une interprétation *a priori* des classes par l'analyse de quelques individus. Puis, le reste des individus sera classé après une étape de pré-classification. Ainsi, le nombre de classes est fixé au départ du processus de classification.

- ✓ Les classifications non supervisées : pour ce type de classification, ni le nombre de classes, ni leurs interprétations ne sont connus au départ du processus. La définition des classes est faite en se basant uniquement sur les mesures qualifiant un individu. Il existe plusieurs algorithmes pour classer les données d'une manière non supervisée, nous citons quelques-uns:
  - ✓ Classification par partitions : cette classification consiste à découper l'espace des mesures en définissant à priori un nombre de classes.
  - ✓ Classifications hiérarchiques : ce type de classification est composé d'un ensemble d'individus regroupés d'une manière itérative en fonction des mesures. Ce type de classification peut se faire selon deux façons différentes :
    - ✓ Classifications par agglomérations (classifications ascendantes) : il s'agit d'une classification hiérarchique en considérant au départ que chaque individu constitue une classe à part. De plus, il faut regrouper les classes d'une manière itérative jusqu'à la formation d'une seule classe représentant l'ensemble complet des individus.
    - ✓ Classifications par divisions (classifications descendantes) : à l'inverse de la classification par agglomérations, cette classification commence par considérer l'ensemble des individus comme une classe qu'il va falloir diviser d'une manière itérative jusqu'à l'aboutissement en  $n$  classes dont chacune contient un seul individu.

Nous avons opté pour le choix d'une classification non supervisée afin de détecter toutes les configurations possibles de différences de forme et de position. L'utilisation d'une classification supervisée suppose une connaissance de toutes les configurations possibles de disparités et donc une définition *a priori* du nombre de classes à générer.

Les classifications non supervisées peuvent se faire par plusieurs méthodes (cf. figure III-30). Nous en exposerons quelques-unes dans la section suivante. L'utilisation de plusieurs méthodes de classification permet de vérifier la capacité des indicateurs à classer les entités appariées dans une classe donnée d'une manière stable. Pour tester également la stabilité des classifications à classer les entités de la même manière, les différentes classifications seront réalisées sur les mesures initiales et sur les mesures synthétiques issues de l'analyse en composantes principales.

### **III.3.2. Classification des liens sur la base des mesures**

Comme nous l'avons signalé plus haut, chacune des entités géographiques appariées sera représentée par un ensemble de mesures. Elles sont alors représentées par un point dans un espace à  $n$  dimensions ( $n$  est le nombre de mesures utilisées). Dans ce paragraphe, nous présentons les différentes classifications utilisées avec l'application de chacune d'entre-elles sur un échantillon contenant 16 individus. La méthode retenue sera appliquée sur l'ensemble des mesures de toutes les entités mises en correspondance et présentée en figure III-30. Pour la construction de cet échantillon, nous avons choisi des

entités représentant visuellement la même classe d'écart. Nous estimons que les entités retenues pour les tests consistent quatre classes d'écart. Chaque classe est composée de quatre entités (les entités 1, 2, 3 et 4 constituent une même classe d'écart). Ce choix est fait délibérément afin de tester les différentes techniques de classification à pouvoir retrouver d'une manière automatique ce que l'opérateur a pu interpréter visuellement.

L'échantillon d'entités utilisées, ainsi que leurs mesures respectives, est illustré par la figure III-30.

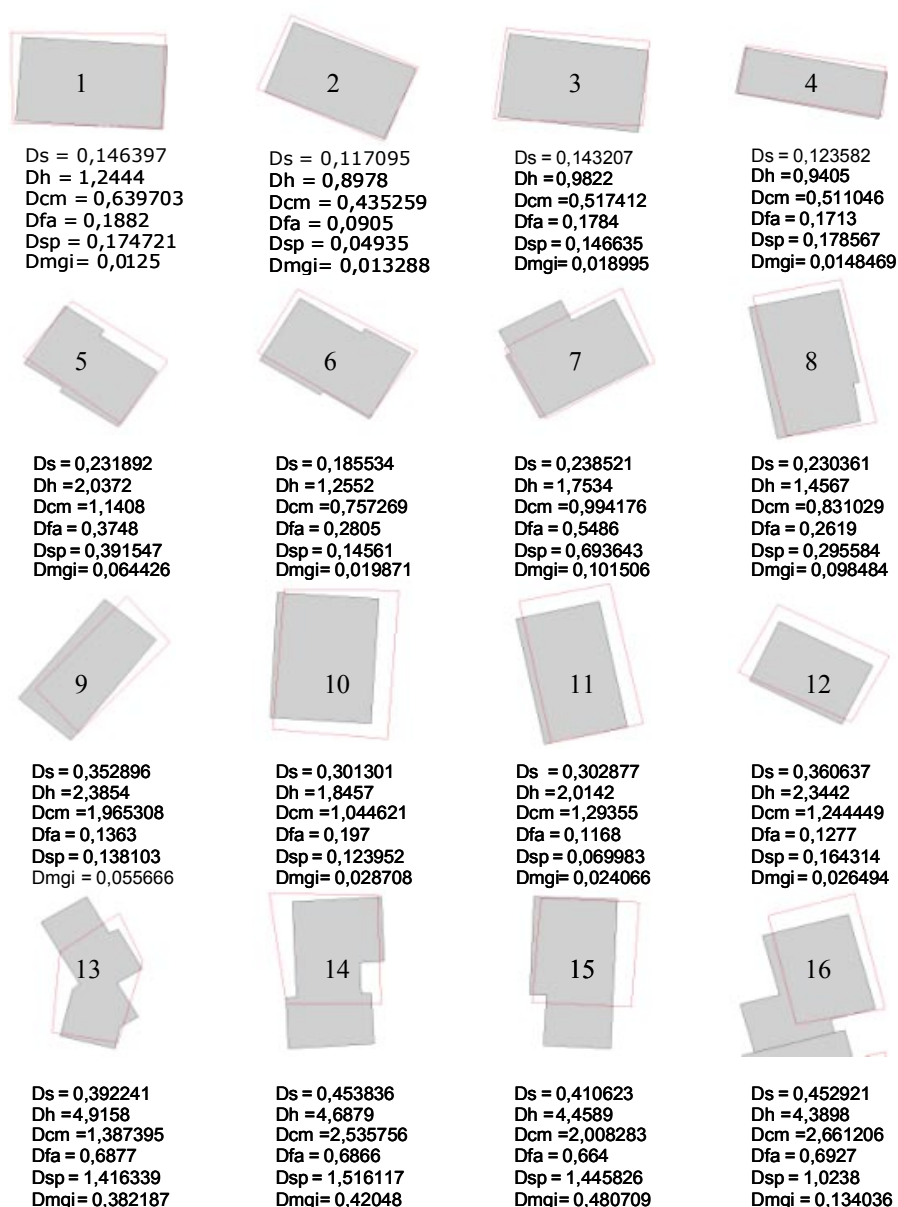


Figure III-30 : Echantillon utilisé pour les classifications

Avant de procéder à la classification des mesures, il est nécessaire de procéder par l'établissement d'une matrice des distances entre les différentes valeurs des mesures.

Cette matrice est souvent appelée matrice de "dissimilarité"<sup>31</sup> [Kaufman & Rousseeuw 1990]. Le calcul de cette matrice se fait de la manière suivante :

Soit un ensemble E composé de n liens d'appariement, dont chaque lien est caractérisé par un ensemble de m mesures entre les entités qui les mettent en correspondance. La matrice de dissimilarité est calculée en mesurant la distance euclidienne entre ces mesures. Elle a la forme suivante :

$$\left[ \begin{array}{cc} d_{euc}(\{mes_2\}\{mes_1\}) & 0 \\ d_{euc}(\{mes_n\}\{mes_1\}) & d_{euc}(\{mes_n\}\{mes_2\}) \dots 0 \end{array} \right] \quad \text{avec}$$

$$d_{euc}(\{mes_i\}\{mes_j\}) = \sqrt{(mes_{1i} - mes_{1j})^2 + (mes_{2i} - mes_{2j})^2 + \dots + (mes_{mi} - mes_{mj})^2} \quad [\text{III-9}]$$

La matrice de dissimilarité peut être calculée, soit sur les mesures initiales, soit sur leurs valeurs normalisées. Le tableau suivant illustre la matrice de dissimilarité pour l'échantillon de 16 individus<sup>32</sup> (cf. figure III-30).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0															
2	0.44	0														
3	0.29	0.18	0													
4	0.34	0.22	0.14	0												
5	0.98	1.42	1.27	1.31	0											
6	0.16	0.53	0.38	0.44	0.91	0										
7	0.89	1.30	1.13	1.15	0.47	0.83	0									
8	0.33	0.76	0.61	0.64	0.68	0.28	0.60	0								
9	1.76	2.15	2.03	2.07	0.97	1.67	1.36	1.49	0							
10	0.74	1.15	1.02	1.07	0.39	0.67	0.68	0.49	1.07	0						
11	1.03	1.42	1.30	1.35	0.45	0.95	0.86	0.78	0.77	0.32	0					
12	1.28	1.68	1.56	1.61	0.49	1.22	0.94	1.01	0.72	0.55	0.35	0				
13	4.00	4.41	4.28	4.30	3.10	3.97	3.29	3.72	2.96	3.40	3.27	2.94	0			
14	4.21	4.64	4.49	4.51	3.24	4.15	3.44	3.89	2.82	3.56	3.36	3.08	1.18	0		
15	3.78	4.21	4.06	4.08	2.83	3.73	3.02	3.47	2.54	3.15	2.98	2.68	0.78	0.58	0	
16	3.88	4.31	4.17	4.20	2.90	3.80	3.15	3.57	2.37	3.19	2.96	2.70	1.46	0.66	0.86	0

Tableau III-4 : Exemple d'une matrice de dissimilarité

La matrice de dissimilarité sert à initialiser le processus de classification en établissant une distance entre les mesures de chacun des individus de l'échantillon. A titre d'exemple, en utilisant les valeurs de cette matrice, on peut conclure *a priori* que l'individu numéro 2 est plus proche de l'individu numéro 4 avec une distance de 0.22 (ligne 5 de la matrice) que l'individu numéro 14, qu'on considère comme le plus éloigné avec une distance de 4.51 (colonne 5 de la matrice). Par ailleurs l'analyse de la ligne 5 de la matrice montre que l'individu 3 est le plus proche de l'individu 4 avec une distance de 0.14.

Dans la suite de cette section, nous allons essayer de classer l'ensemble des 16 individus de la figure III-30 en utilisant les différents types de classification et de montrer les avantages et les inconvénients de chacune d'entre-elles.

<sup>31</sup> Egalement appelée matrice de similarité

<sup>32</sup> Nous utilisons le terme "individu" pour désigner l'ensemble des mesures entre deux entités appariées.



### III.3.2.1. Classification par partitions

Ce type de classification se fait en définissant au départ le nombre de classes ( $k$  classes). La définition du nombre de classes conditionne le choix de  $k$  individus d'une manière aléatoire qu'on considère comme individus représentatifs des classes. Chaque ensemble de mesures d'un lien donné est affecté dans la classe qui lui est proche. En d'autres termes, soit l'ensemble de mesures  $i$ , soit la classe  $c_i$  dont l'individu représentatif est  $m_{c_i}$ , soit  $m_c$  les individus représentatifs des autres classes, on affecte  $i$  à la classe  $c_i$  si et seulement si :  $d(i, m_{c_i}) \leq d(i, m_c)$  pour tout  $c = 1, \dots, k$ .

Au départ, les individus représentatifs sont choisis au hasard, mais ils seront remplacés en cours de processus d'une manière itérative en les inter-changeant par d'autres individus représentatifs tout en minimisant la fonction suivante :

$$\sum_{i=1}^n d(i, m_{c_i}) \quad \text{[III-10]}$$

L'application de ce type de classification sur l'échantillon des 16 liens donne les résultats suivants en fixant le nombre de classes à générer au nombre de 2, 3 et 4 :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
<b>2 cl.</b>	①	①	①	①	①	①	①	①	①	①	①	①	②	②	②	②
<b>3 cl.</b>	①	①	①	①	②	①	②	②	②	②	②	②	③	③	③	③
<b>4 cl.</b>	①	①	①	①	②	①	②	②	③	③	③	③	④	④	④	④

Tableau III-5 : Classification autour des individus représentatifs

Les cases grisées du tableau III-5 représentent les individus représentatifs de chaque classe. La fixation de classes permet de différencier les entités appariées présentant, entre elles, un fort décalage en position (individus 13, 14, 15 et 16) et en forme par rapport aux autres entités. Cette division des individus en 2 classes permet de discriminer d'une manière grossière la nature des écarts entre les entités. L'augmentation du nombre de classes en entrée de processus permet d'affiner la classification des individus (de 1  $\rightarrow$  12).

La répartition des individus en 4 classes permet de couvrir les différents types d'écarts de forme et de position présents entre les entités de l'échantillon. Les quatre classes peuvent être interprétées de la manière suivante :

- ✓ Classe 1 : entités ayant la même forme et occupant la même position dans l'espace.
- ✓ Classe 2 : entités occupant la même position avec un écart en forme.
- ✓ Classe 3 : entités ayant la même forme avec une position biaisée.
- ✓ Classe 4 : entités ayant un fort biais en position et une différence conséquente de forme.

Ce type de classification reste toujours dépendant de la définition du nombre de classes à générer, ce qui présente un handicap majeur vu la méconnaissance *a priori* de ce paramètre. Pour l'analyse d'un échantillon réduit de liens, on peut se permettre la génération de plusieurs classes et la fixation du nombre *a posteriori* au moment où l'interprétation physique de la nature des écarts s'avère impossible. Une telle analyse est impossible en présence d'un nombre conséquent d'entités à analyser, comme des jeux de données réels.

### III.3.2.2. Classifications hiérarchiques

Les classifications hiérarchiques se divisent en deux grandes classes : les classifications hiérarchiques qui procèdent par division et les classifications hiérarchiques qui procèdent par agglomération.

#### III.3.2.2.1. Classifications hiérarchiques par division (Descendante)

L'ensemble total des individus à analyser est considéré comme une classe qu'on divise d'une manière itérative jusqu'à l'obtention de plusieurs classes contenant chacune un seul individu. Cette classification opère de la manière suivante :

1- Chercher l'objet le plus disparate (ayant la plus grande moyenne de dissimilarités par rapport au reste des autres objets). Cet objet initie ce qu'on appelle le "*splinter group*".

2- Pour tout objet en dehors du "*splinter group*", on calcule la quantité suivante :

$$V_i = \text{moyenne}_{j \in \text{splinter group}} d(i, j) - \text{moyenne}_{j \notin \text{splinter group}} d(i, j) \quad [\text{III-11}]$$

si  $V_h > 0$  alors l'objet  $h$  est moyennement proche du "*splinter group*" et donc il lui sera ajouté.

3- L'opération (étape 2) sera répétée jusqu'à ce que  $V_h < 0$ . De plus, on scinde l'ensemble des objets en deux classes.

Choisir la classe qui présente le plus grand diamètre et refaire les étapes 1, 2 et 3.

Répéter toutes les étapes précédentes jusqu'à l'aboutissement à des classes contenant chacune un seul objet.

Le résultat de cette classification est donné sous la forme d'un arbre. L'utilisation de cette méthode pour classer les 16 individus de l'échantillon est réalisée en s'appuyant sur les mesures initiales. La classification est également réalisée en utilisant les deux premières composantes principales (porteuses de 96% de l'information initiale). Le résultat est illustré par la figure III-31 :

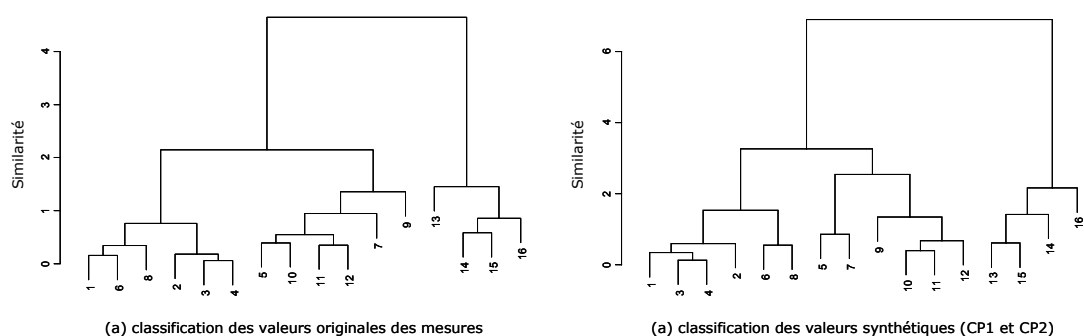


Figure III-31 : Exemple de classification hiérarchique par divisions

On note que selon que l'on utilise les mesures initiales ou les mesures synthétiques, le résultat de la classification n'est pas le même. L'identification des classes se fait d'une manière empirique par la coupure de l'arbre de classification à différents niveaux et par l'interprétation des classes à chaque seuil de coupure. L'exemple des arbres de classification figure III-31(a) et (b) montre que l'arbre peut être découpé au niveau d'une similarité égale à 3 pour générer deux classes (les mêmes classes retrouvées par une classification par partitions). Cependant, si on tente d'affiner la classification, on retrouve deux résultats différents selon qu'on utilise l'un ou l'autre des arbres de classifications. Le choix d'un seuil de coupure autour d'une similarité à 1.8 de l'arbre (a) génère 3 classes. Ceci permet de retrouver une classification comparable à celle effectuée autour des individus représentatifs, mis à part l'individu 8 qui ont basculé de la classe 2 à la classe 1. Au-delà de cette valeur de seuil, l'interprétation physique des écarts de forme et de position des entités appariées s'avère relativement confuse.

Par ailleurs, le choix d'un seuil de coupure autour de la valeur de 2.8 pour l'arbre (b) permet la génération de 3 classes (les mêmes générées en coupant l'arbre (a)). Mais, la coupure autour de la valeur 2.2 permet d'aboutir à la définition de quatre classes. Cette classification en quatre classes est également la même que celle réalisée par la méthode des partitions, à l'exception de l'individu 8 qui est passé de la classe 2 à la classe 1.

### III.3.2.2.2. Classifications hiérarchiques par agglomérations (Ascendante)

Par opposition aux techniques de classifications hiérarchiques par divisions, les classifications hiérarchiques par agglomérations commencent par considérer chaque individu comme une classe à part qu'on regroupe d'une manière itérative jusqu'à la formation d'une seule classe regroupant tous les individus analysés.

Ce type de classification s'opère de la manière suivante :

1- Créer  $n$  classes, contenant chacune un individu de l'ensemble de données; les distances (similitudes) entre les individus étant connues (matrice de dissimilarité).

2- Chercher dans la matrice de dissimilarité les deux classes  $i$  et  $j$  les plus similaires ayant une faible distance.

3- Fusionner les classes  $i$  et  $j$  en une classe  $(ij)$ . Les distances seront recalculées en fonction de la méthode utilisée. Le nombre des classes, ainsi que la taille de la matrice décroîtront de 1 à chaque étape.

Répéter l'étape 2 et 3, jusqu'à l'aboutissement d'une seule classe.

L'étape de regroupement (ou de fusion) des classes peut se faire de plusieurs manières [Johnson & Wichern 1998; Mirkin 1987; Ward 1963] par l'utilisation de différents modes de calcul pour minimiser les distances entre les classes regroupées et reconstruire une nouvelle matrice de dissimilarité à chaque étape de fusion. On énumère les modes les plus utilisés :

**Fusion avec le plus proche voisin:** (*single linkage*) le mode de calcul utilisé est donné par l'équation suivante :

$$d(C_{(ij)}, C_k) = \text{Min}\{d(C_i, C_k), d(C_j, C_k)\} \quad \text{[III-12]}$$

$d$  indique la distance entre les groupes et  $C_i$  représente l'ensemble des individus de la classe  $i$ . Le calcul de la distance se fait pour tous les individus  $k = 1, \dots, n$  avec  $k \neq i, j$ . La matrice de dissimilarité est calculée à chaque étape de fusion en recevant les valeurs issues de ce calcul.

**Fusion avec le voisin le plus loin:** (*complete linkage*), pour ce mode de calcul, l'équation III-7 sera remplacée par l'équation suivante :

$$d(C_{(ij)}, C_k) = \text{Max}\{d(C_i, C_k), d(C_j, C_k)\} \quad \text{[III-13]}$$

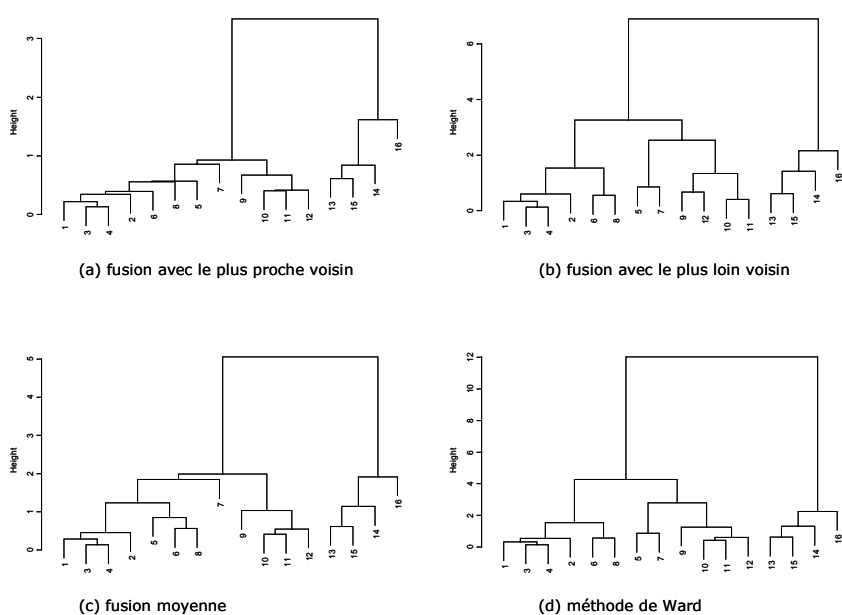
**Fusion moyenne:** (*Average linkage*), pour ce mode de calcul, l'équation III-7 sera remplacée par l'équation suivante :

$$d(C_{(ij)}, C_k) = \frac{\sum_{l=1}^{n_{(ij)}} \sum_{m=1}^{n_k} d(\text{individu}_l, \text{individu}_m)}{n_{ij} \cdot n_k} \quad \text{[III-14]}$$

avec  $\text{individu}_l \in C_{ij}$ ,  $\text{individu}_m \in C_k$ ,

**Fusion par la méthode de Ward** [Ward 1963]: la distance utilisée pour cette méthode est calculée par la sommation des carrés des dissimilarités.

L'application de ces méthodes est illustrée par la figure III-32.



**Figure III-32 : Classification par les méthodes hiérarchiques par agglomérations**  
**-les classifications sont réalisées sur les deux premières composantes principales-**

Parmi les quatre méthodes de classifications hiérarchiques par agglomérations, on remarque que seules les méthodes utilisant la fusion avec le voisin le plus loin et la méthode de Ward donnent le même résultat que les classifications précédemment testées.

Nous notons que le but visé par l'utilisation de plusieurs classifications (bien que la liste des méthodes ne soit pas exhaustive) n'est pas d'effectuer une étude comparative entre elles, mais plutôt de voir la stabilité d'affectation des individus dans les classes par l'emploi de différentes méthodes, ainsi que la suffisance des mesures à couvrir tous les écarts possibles de forme et de position entre les entités surfaciques appariées. Dans la suite de l'étude, pour classifier des mesures d'un jeu de données entier, nous décidons de n'employer que deux méthodes de classification, à savoir : une classification par partitions et une classification hiérarchique par agglomérations utilisant la méthode de Ward. Le choix de ces deux méthodes de classification est justifié par le fait que ces deux techniques de classification renvoient un résultat proche de celui réalisé en interprétant visuellement les configurations d'écart entre les entités testées.

L'application de quelques-unes de ces classifications sur 3 mesures (distance surfacique, distance de Hausdorff et distances entre les fonctions angulaires) des entités appariées d'un jeu de données réel [Bel Hadj Ali 2000] montre une confusion de classifications de quelques individus. Ces individus se trouvent généralement sur les limites de classes et ils restent susceptibles d'être classés dans une classe ou dans une autre selon la méthode de classification utilisée. Le même constat est fait par [Reiners 1998]. Ce déplacement d'une classe à une autre est du essentiellement à l'incapacité des mesures utilisées à décrire toutes les caractéristiques géométriques des entités analysées.

### III.3.3. Analyse des classifications et contrôle qualité

Dans ce paragraphe, nous détaillons l'application des deux méthodes de classifications précédemment présentées sur l'ensemble des entités simples appariées entre le Cadastre et la BDTopo®. Les mesures utilisées sont celles pour l'échantillon des 16 individus de la figure III-30. Nous rappelons que l'analyse de la classification se fait d'une manière visuelle, en faisant appel à une intervention humaine pour interpréter les différentes configurations des écarts de position et de forme. Par conséquent, la définition des classes à générer se fait d'une manière itérative et l'arrêt du processus se fait au moment où l'interprétation s'avère impossible.

Une fois que les classes seront définies d'une manière définitive, une étape d'apprentissage sera enclenchée dans le but de générer des règles de classification. La génération de ces règles, ainsi que leur utilité seront détaillées plus loin.

La classification est effectuée sur des mesures synthétiques, en l'occurrence les quatre premières composantes principales (cf. §III.3.1.). Le regroupement de ces mesures par la classification hiérarchique par agglomérations en utilisant la méthode de Ward est illustré par l'arbre de classification (cf. figure III-33). On distingue sur l'arbre de classification deux grandes classes d'entités ( $Deg^{33} > 30$ ). La première classe (à gauche de l'arbre) regroupe toutes les entités présentant, soit un fort décalage en position, soit une différence conséquente de forme. Cette classe contient 28 individus dont les statistiques sur les valeurs de leurs mesures sont données par le tableau III-6.

	<b>Ds</b>	<b>Dh</b>	<b>Dcm</b>	<b>Dfa</b>	<b>Dsp</b>	<b>Dmgi</b>
<b>Minimum</b>	0.322	2.689	0.460	0.390	0.735	0.068
<b>Maximum</b>	0.585	16.419	6.954	1.046	5.784	2.029
<b>Moyenne</b>	0.434	6.984	2.863	0.683	1.658	0.609
<b>Ecart type</b>	0.073	3.594	1.537	0.141	0.987	0.498

Tableau III-6 : Statistiques sur les valeurs des mesures des entités ayant des écarts conséquents de forme et de position

<sup>33</sup> Deg :Distance entre groupe

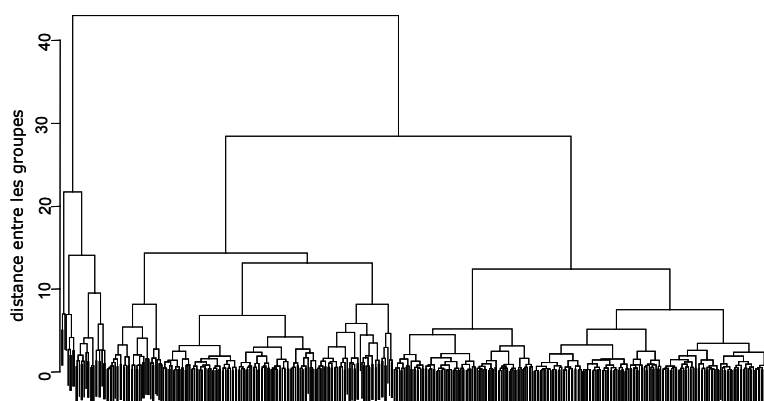


Figure III-33 : Arbre de classification des mesures entre les entités simples

La deuxième classe (située à droite de l'arbre et contenant 411 individus ) peut, à son tour, être découpée en sous-classes pour affiner la classification. En effet, le sous-arbre illustrant la classe 2 sera isolé et découpé de nouveau. Ce sous-arbre peut être découpé en deux classes au niveau d'une Deg autour de 20. Cependant, un seuil de coupure autour de la valeur 10 permet la génération de 5 classes. Le choix d'un seuil inférieur à 10 permet la génération de plus de classes, difficilement interprétables au niveau de la détection des différentes configurations des écarts. En effet, on se contente de stopper le découpage de l'arbre de classification à ce niveau. Ce découpage est illustré par la figure III-34.

On note que le découpage de l'arbre de classification se fait d'une manière empirique. A chaque étape de découpage, l'arbre est scindé en sous-arbres qui sont à leur tour analysés et découpés d'une manière isolée. Le découpage est arrêté au moment où on ne distingue plus visuellement de différence dans les configurations d'écarts entre deux classes.

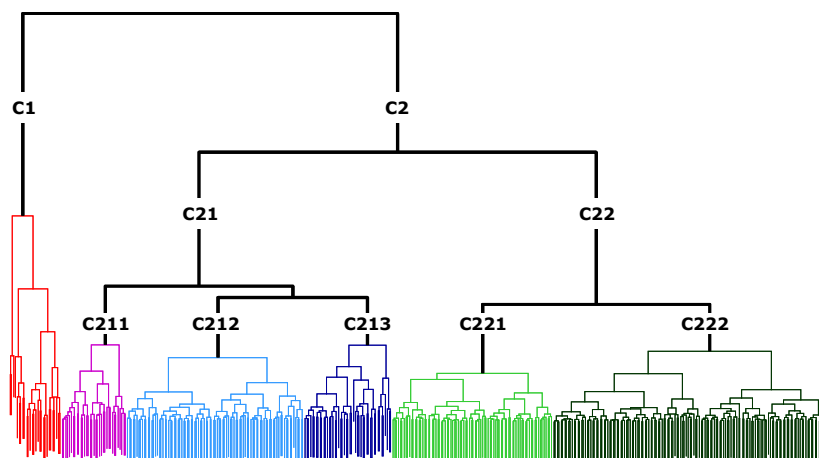


Figure III-34 : Répartition des individus analysés en 6 classes

### III.3.3.1. Interprétation des classes

L'interprétation des classes se fait d'une manière visuelle en essayant de qualifier chaque classe selon la nature des écarts existant entre les entités appariées. Cette interprétation est donnée dans le tableau suivant :

Classe	Interprétation	Exemples
C1	Forte différence de forme induisant un fort biais en position. Ces écarts peuvent être causés par une mauvaise interprétation des entités ou par l'omission d'une partie importante d'une entité.	
C211	Forte différence de forme avec un écart moyen de position.	
C212	Une légère différence de forme (généralisation, suppression d'un détail non significatif) avec une occupation de la même position dans l'espace.	
C213	Différence conséquente de forme (forte généralisation, suppression de gros détails) et occupation de la même position dans l'espace géographique.	
C221	Exactitude de forme, avec un léger biais en position.	
C222	Exactitude en forme et en position.	

Tableau III-7 : Interprétation de la configuration des écarts

Les exemples donnés dans le tableau III-7 sont représentatifs de la majeure partie des individus dans chacune des classes. Par ailleurs, comme nous l'avons signalé précédemment, il existe des individus dans une classe dont l'interprétation de la configuration des écarts paraît très difficile à élucider. Cela est dû au fait que l'individu se trouve à la frontière de la classe et que son interprétation se confond avec celle de la classe adjacente.

La méthode de classification par partitions a été également appliquée sur l'ensemble des mesures initiales de la classe 2. Le résultat de cette classification est



illustré par la partition en figure III-35. La représentation des partitions est faite sur le premier plan principal.

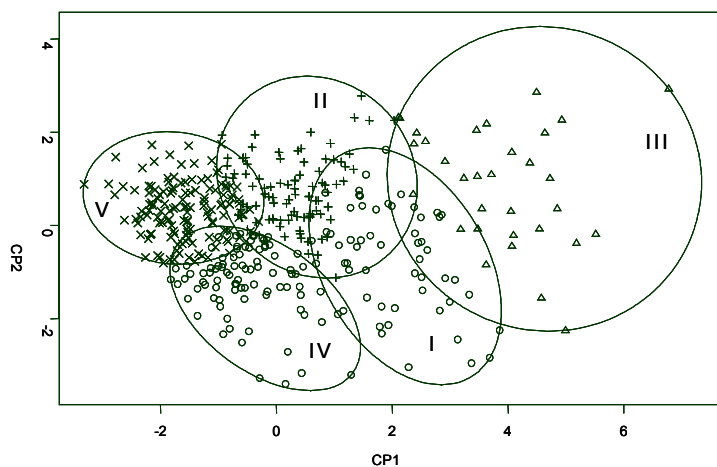


Figure III-35 : Classification par partitions

D'une manière générale, à l'issue du processus d'appariement, les entités mises en correspondance sont plus ou moins comparables, ce qui implique que leur représentation dans l'espace des mesures paraît trop compacte. L'identification des classes ne peut donc pas se faire d'une manière complètement séparée, puisqu'il existe des individus se retrouvant en frontières de classes. De plus, leur appartenance à une classe donnée est souvent faite d'une manière incertaine et confuse. En effet, les 5 classes obtenues lors de la première classification et les 5 obtenues par la deuxième classification ne sont pas les mêmes. Cependant, il existe des individus qui ont été rangés dans la même classe indépendamment de la méthode utilisée. La matrice de confusion suivante montre le nombre d'individus bien classés et ceux qui ont changé de classe en fonction de la méthode utilisée.

	I	II	III	IV	V	Total
C211	<b>23</b>	1	13	7	0	44
C212	10	<b>61</b>	0	2	0	73
C213	23	9	<b>19</b>	0	0	51
C221	0	1	0	<b>56</b>	12	69
C222	0	24	0	22	<b>128</b>	174
Total	56	96	32	87	140	

Tableau III-8 : Matrice de confusion -comparaison de deux classifications-

L'utilisation de ces deux méthodes de classification a permis de bien classer (la diagonale de la matrice tableau III-8) 70% des individus de la deuxième classe. Cependant, l'analyse visuelle des individus qui contribuent à la confusion entre les classes, permet de constater que la classification par la méthode hiérarchique par agglomérations respecte plus l'interprétation donnée en tableau III-7 que la classification par partitions autour des individus représentatifs. Ainsi, nous nous en tenons à la seule utilisation de la classification hiérarchique par agglomérations, étant donné qu'elle permet une meilleure différenciation entre les classes générées.

Nous avons utilisé, dans le cadre de cette étude, des techniques de classification non supervisée, puisque nous ne connaissons pas à priori le nombre de classes d'écartes

dans le jeu de données analysé. Cependant, si le nombre de classes est connu ou si nous disposons d'échantillons représentatifs de chacune d'entre elles, on pourrait<sup>34</sup> envisager l'emploi d'autres méthodes de classification telle que la méthode dite "*Learning Vector Quantization (LVQ)*" [Kohonen 1986] afin de minimiser la confusion entre les classes d'écart. Ce type de classification vise à définir des surfaces de décision entre les différentes classes. Ces surfaces de décision présentées comme des hyperplans linéaires par morceaux sont obtenues par un processus stochastique d'apprentissage supervisé sur les échantillons représentatifs et qui approchent la probabilité de l'erreur minimale bayésienne de classification.

On présente dans le tableau III-9 les statistiques des mesures au sein de chacune des classes générées. En agrégeant, par exemple, les classes C22x en une seule classe C22 et en les considérant comme des classes dont les entités sont légèrement biaisées en position avec une légère différence de forme, des premières conclusions peuvent être tirées en analysant les statistiques de cette classe. Les valeurs de la distance surfacique ont une moyenne égale à 0.24 (avec un écart type de 0.06), ce qui permet de dire, *a priori*, que les entités appariées ayant une distance surfacique inférieure ou égale à cette valeur peuvent être acceptées comme des entités correctement appariées. Les valeurs de la distance de Hausdorff ont une moyenne de 1.669 mètres. Cette valeur se rapproche de la valeur de l'erreur quadratique moyenne de la BDTopo®.

		Ds	Dh	Dcm	Dfa	Dsp	Dmgi
Classe C211 (44 individus)	Minimum	0,258	1,204	0,484	0,233	0,130	0,031
	Maximum	0,583	5,506	2,449	0,728	1,231	1,111
	Moyenne	0,399	3,021	1,414	0,468	0,713	0,284
	Ecart type	0,081	0,895	0,529	0,116	0,308	0,225
Classe C213 (51 individus)	Minimum	0,119	1,693	0,144	0,417	0,520	0,015
	Maximum	0,333	9,136	1,969	0,726	2,420	0,495
	Moyenne	0,254	3,063	1,002	0,564	0,947	0,192
	Ecart type	0,049	1,221	0,438	0,077	0,352	0,122
Classe C212 (73 individus)	Minimum	0,205	1,279	0,114	0,284	0,248	0,016
	Maximum	0,450	3,017	1,217	0,611	1,026	0,359
	Moyenne	0,287	2,123	0,731	0,410	0,565	0,155
	Ecart type	0,048	0,366	0,260	0,062	0,133	0,085
Classe C221 (69 individus)	Minimum	0,258	0,937	0,102	0,117	0,069	0,002
	Maximum	0,416	3,176	1,965	0,389	0,485	0,161
	Moyenne	0,311	1,993	0,989	0,238	0,235	0,054
	Ecart type	0,033	0,476	0,386	0,065	0,103	0,040
Classe C222 (174 individus)	Minimum	0,072	0,661	0,083	0,091	0,049	0,003
	Maximum	0,283	2,986	1,398	0,522	0,671	0,341
	Moyenne	0,208	1,582	0,723	0,259	0,275	0,072
	Ecart type	0,039	0,373	0,272	0,076	0,127	0,062

Tableau III-9 : Statistiques des mesures au sein de chaque classe

Nous notons que la définition du nombre de classes reste toujours dépendante des jeux de données à analyser du fait qu'un jeu de données ne peut pas contenir toutes les

<sup>34</sup> Cette technique n'a pas été étudiée dans le cadre de notre travail. Par ailleurs elle pourrait être une piste intéressante à étudier afin d'affiner les résultats de classification.

configurations des écarts. En admettant qu'un écart de position ou de forme peut être décrit quantitativement par quatre classes (pas d'écart, faible, moyen et fort), une combinaison de ces classes d'écarts de position et de forme des entités appariées permet l'obtention de 13 configurations, comme le montre le tableau suivant :

		Ecart de forme			
		Nul	Faible	Moyen	Fort
Ecart de position	Nul	✓			
	Faible	✓	✓	✓	✓
	Moyen	✓	✓	✓	✓
	Fort	✓	✓	✓	✓

Tableau III-10 : Différentes classes d'écart de forme et de position

Dans l'exemple du Cadastre et de la BDTopo, nous n'avons détecté que 6 classes parmi les 13 classes du Tableau III-, que nous pouvons ré-interpréter comme suit :

		Ecart de forme			
		Nul	Faible	Moyen	Fort
Ecart de position	Nul				
	Faible		C222	C212	C213
	Moyen		C221		C211
	Fort				C1

Tableau III-11 : Ré-interprétation des classes des écarts entre les entités du cadastre et de la BDTopo®

Le jeu de données Cadastre-BDTopo® utilisé pour détecter les classes d'écarts ne contient pas toutes les configurations possibles (6 configurations parmi 13). En effet, la méthode de mesures et de classification devrait être poursuivie sur d'autres jeux de données afin de détecter toutes les configurations d'écarts.

### III.3.4. Règles de classification

La classification des mesures entre les entités appariées et l'interprétation des configurations des écarts est une étape fastidieuse et très coûteuse. A cet effet, il est nécessaire de compresser toute la méthode pour la fourniture de règles de classification utilisables dans le cas où on traiterait des données avec les mêmes spécifications et la même nature que celles permettant la génération de ces règles.

Pour la génération des règles de décision, nous utilisons les techniques d'apprentissage supervisé [Mustière 2001], en utilisant un algorithme considéré parmi les plus stables connu sous le nom de l'algorithme C4.5 [Quinlan 1993].

La classification étant faite sur les indicateurs synthétiques (composantes principales), chaque individu se classe donc dans une des classes générées et la génération des règles de décision se fait en utilisant les mesures initiales. Cet "aller-retour" entre les données initiales et les données synthétiques permet de confronter les résultats et d'en déduire des interprétations correctes. [Chamussy & al. 1994] préconisent également de confronter l'interprétation des composantes principales et les mesures initiales pour une meilleure compréhension des données :

*"[...] Chaque résultat doit être interprété par confrontation avec le tableau initial qu'il concourt à analyser et l'ensemble de l'analyse a pour but de mieux le connaître, le comprendre*

*et le juger. La confrontation et le retour permanent aux données initiales est la règle d'or d'une bonne interprétation." [Chamussy & al. 1994, p. 148]*

Reprenons l'exemple des 439 entités appariées entre le cadastre et la BDTopo® en utilisant le découpage en 6 classes résultant de la classification hiérarchique par agglomérations, les 439 individus sont partagés au hasard en deux jeux : un jeu utilisé pour générer les règles et un jeu pour contrôler les règles obtenues. Le jeu d'exemple est donc divisé en deux jeux en prenant les 2/3 des individus pour générer les règles et le 1/3 restant pour les contrôler, comme cela est préconisé dans les techniques d'apprentissage [Mustière 2001, p. 87]. Le choix de l'échantillon pour effectuer le processus d'apprentissage est fait en respectant les mêmes proportions des individus à l'intérieur de chaque classe. Le reste des mesures servira pour contrôler le résultat des règles afin de les valider et déterminer les erreurs dues aux mauvaises classifications. Les règles générées sont données par la figure III-36. Les règles sont générées d'une manière automatique en employant les mesures qui paraissent les plus discriminantes. On constate que les règles obtenues n'utilisent pas la distance entre les moments géométriques invariants. Ceci pourrait être dû à deux facteurs : soit cette mesure n'est pas trop discriminante et n'apporte pas une information supplémentaire sur les écarts par rapport aux autres mesures, soit l'échantillon utilisé est insuffisant pour pouvoir générer des règles robustes pour la classification. La conclusion ne peut être définitivement tirée qu'après l'analyse des règles de classification obtenues en analysant leur robustesse à bien classer l'échantillon de contrôle.

Les pourcentages donnés sur la figure III-36 (à droite de chaque libellé de règle) indiquent les taux de confiance qu'on peut accorder à la règle pour aboutir à une classe donnée. Les taux de confiance qui sont inférieurs à 100% indiquent que la règle correspondante peut mal classer un individu en l'imputant à une classe à laquelle il n'appartient pas au départ. Dans le cadre de cet exemple, les règles obtenues ont mal classé 10 individus parmi les 293 utilisés pour leur génération, soit un taux d'erreur de 3.125%. Ce taux d'erreur est appelé erreur apparente et il est considéré comme un estimateur peu fiable et trop optimiste de l'erreur réelle [Mustière 2001]. Par conséquent, il faut procéder à une évaluation empirique par l'application des règles obtenues sur le jeu de contrôle et mesurer la confusion et l'accord entre la classification initiale et la classification par les règles.

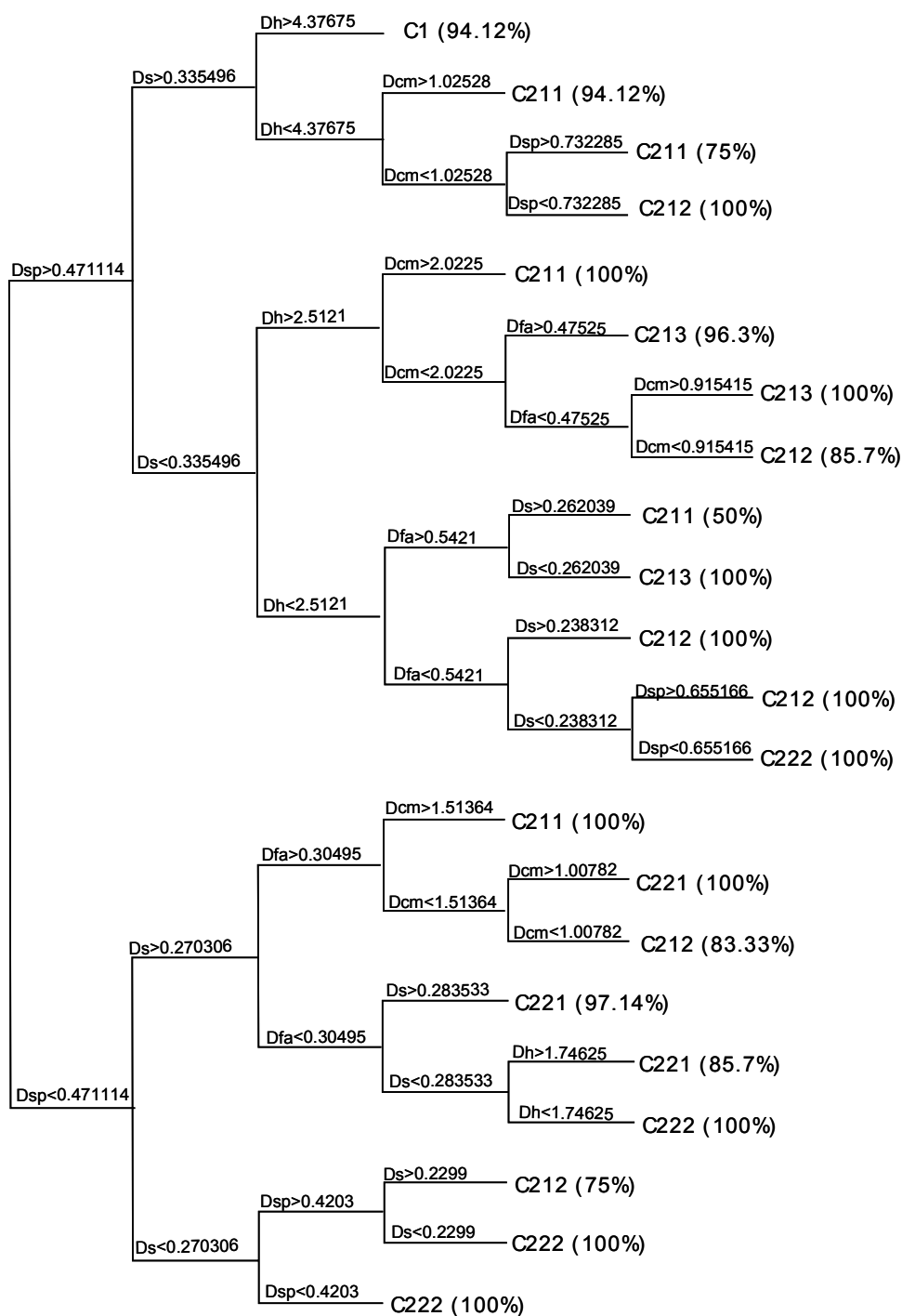


Figure III-36: Règles de classification générées par les méthodes d'apprentissage supervisé

Les règles obtenues sont appliquées sur les 146 individus (échantillon de contrôle) afin de les re-classifier et de comparer le résultat avec la classification initiale. La classification par les règles et la classification initiale sont confrontées pour vérifier les taux d'accord entre les deux classifications. La matrice de confusion est donnée par le tableau III-12.

		Classification initiale					
		c1	c211	c212	c213	c221	c222
classification par les règles	c1	7					
	c211	2	8	1	2		
	c212		3	19	3	3	5
	c213		1		12		0
	c221		3			16	0
	c222			4		4	53

Tableau III-12 : Matrice de confusion -robustesse des règles de classification-

La reclassification par les règles a permis de bien classer les 146 individus avec un taux d'accord général de 80% que nous considérons comme un taux acceptable. Ce taux peut être amélioré si l'échantillon utilisé pour l'apprentissage comportait plus d'individus. La non-utilisation de la distance entre les moments géométriques invariants n'est donc pas due à la taille de l'échantillon utilisé pour l'apprentissage, mais plutôt au fait qu'elle n'apporte pas une information supplémentaire par rapport aux autres mesures en ce qui concerne la description des écarts géométriques. Cela est confirmé par l'obtention des mêmes règles de classification avec l'augmentation de la taille de l'échantillon d'apprentissage. L'interprétation des règles est donnée dans la section suivante.

### III.3.4.1. Interprétation des règles et contrôle qualité

La classe C222 peut être considérée comme la classe regroupant les entités de bonne qualité puisqu'elle est composée des entités appariées qu'on estime ayant la même forme et occupant la même position. L'utilisation des règles de classification (cf. figure III-36) permet de calibrer les valeurs des mesures utilisées dans un contexte d'appariement et de mesure d'écarts entre les thèmes "Bâti" de la BDTopo® et du Cadastre. On ne peut classer des entités appariées dans la classe C222 que si elles respectent les règles suivantes :

$$\begin{aligned}
 & [(Ds < 0.27) \text{ et } (Dsp < 0.42)] \text{ ou} \\
 & [(Ds < 0.23) \text{ et } (0.42 < Dsp < 0.47)] \text{ ou} \\
 & [(0.27 < Ds < 0.28) \text{ et } (Dsp < 0.47) \text{ et } (Dfa < 0.304) \text{ et } (Dh < 1.75)] \text{ ou} \\
 & [(Ds < 0.24) \text{ ou } (0.47 < Dsp < 0.655) \text{ ou } (Dfa < 0.54) \text{ ou } (Dh < 2.51)] \quad \text{[III-15]}
 \end{aligned}$$

On rappelle que les écarts de position sont estimés par deux mesures (la distance surfacique et la distance de Hausdorff) et que les écarts de forme sont également estimés par deux mesures (la distance entre les fonctions angulaires et la distance entre les signatures polygonales). L'interprétation des règles (équation III-15) permet de conclure que deux entités issues respectivement de la BDTopo® et du Cadastre occupent exactement la même position dans l'espace, si les valeurs de leur distance surfacique et de leur distance de Hausdorff doivent être respectivement inférieures à 0.27 et à 2.5 mètres. Le seuil retrouvé pour la distance de Hausdorff approche une réalité physique qui est l'erreur moyenne quadratique annoncée pour la BDTopo®. Cependant, la valeur du seuil sur les valeurs de la distance surfacique indique qu'il faut avoir au moins 73%

de l'occupation commune de l'espace entre les deux entités (BDTopo® et Cadastre) pour être considérées comme ayant une position correcte l'une par rapport à l'autre.

L'interprétation des seuils sur les mesures de forme permet de retrouver les valeurs retrouvées lors du calibrage des mesures par les techniques de simulations de bruit (cf. §I.3.3, et figure II-32), à savoir un seuil sur les valeurs de la distance entre les fonctions angulaires de 0.54 et un seuil sur les valeurs de la distance entre les signatures polygonales de 0.655. La valeur de seuil sur les valeurs de  $D_{fa}$  rejoint la valeur adoptée par [Arkin & al. 1991] dans un contexte de reconnaissance des formes des objets simples.

La classe C1 regroupe une configuration particulière d'écarts. Les entités participant à la formation de ce type de lien présentent une différence conséquente de forme qui peut être expliquée par une extension (ou une destruction) partielle subie par l'entité (en l'occurrence un bâtiment, dans le présent cas). Ceci reflète un type particulier d'évolution des entités dans les bases de données. Cette évolution peut s'ajouter à la typologie établie par [Badard 1998] concernant les évolutions constatées dans les bases de données en enrichissant la rubrique "Modification géométrique". Ces entités peuvent être retrouvées par la simple application de la règle suivante :

$[(D_{sp} > 0.47) \text{ et } (D_s > 0.34) \text{ et } (D_h > 4.38)]$

Nous signalons encore une fois que les mesures doivent être utilisées d'une manière combinée et qu'une utilisation d'une mesure d'une manière individuelle (même en respectant les valeurs des seuils précédemment énoncées) ne permet pas de résoudre le problème d'identification des différentes classes d'écarts.

Les règles générées, ainsi que le contexte de leur génération sont stockées dans une base de règles pour une éventuelle utilisation future, si on se trouve dans un même contexte d'appariement et de contrôle. La base de règles sera également enrichie par l'introduction de nouvelles règles issues des contrôles des jeux de données.

L'exemple traité précédemment s'intéresse au contrôle de la qualité géométrique des entités simples. Pour les entités complexes, la même démarche sera utilisée mais avec l'utilisation des mesures qui leur sont dédiées, à savoir :

- ✓ Pour les écarts de position : la distance surfacique et la distance de Hausdorff entre les surfaces;
- ✓ Pour les écarts de forme : la distance entre les moments géométriques invariants, la distance entre les moments de Legendre et la distance entre les moments de Zernike.

Le jeu de données utilisé pour détecter les classes des écarts et pour établir les règles de classification ne couvre pas la totalité des configurations. En effet, la méthode doit être réutilisée sur d'autres jeux de données Cadastre-BDTopo® pour obtenir des règles couvrant toutes les configurations d'écarts. Si toutes les configurations sont

détectées, on peut dire que le contexte est totalement couvert et qu'il n'est plus nécessaire de refaire la classification à chaque opération d'appariement et de mesure, mais plutôt d'utiliser directement les règles générées.

On désigne par "contexte" le contexte d'appariement. Si une mise en correspondance et une analyse des mesures et de génération de règles ont été faites entre deux jeux de données de Cadastre et de la BDTopo® et ont permises de la détection de toutes les configurations d'écarts, il sera "inutile" de refaire l'analyse et la génération des règles dans le cas où l'on est amené à appairer et contrôler deux autres jeux de données Cadastre-BDTopo®.

Nous traitons dans le paragraphe suivant un autre exemple d'analyse utilisant la même méthode et s'appliquant à des jeux de données représentant deux couches de l'occupation du sol de la BDTopo® réalisées par deux photointerprètes différents et avec les mêmes spécifications et les mêmes sources de saisie à la même actualité.

### **III.3.5. Application de la méthode des mesures sur des jeux de données "occupation du sol"**

Nous présentons dans ce paragraphe l'application de la méthode développée en §III.3 sur des deux jeux appariés représentant le thème "occupation du sol" de la BDTopo®. On rappelle que les deux jeux de données utilisés ont été saisis par deux photointerprètes différents en utilisant les mêmes sources de saisie, à la même actualité et en se référant aux mêmes spécifications.

En appliquant la méthode d'appariement sur ces deux jeux de données, nous avons obtenu 165 liens mettant en correspondance les entités de ces deux jeux. Ces liens sont répartis en 103 liens simples et 62 liens multiples.

En première approximation, nous avons émis l'hypothèse que si la distance surfacique entre deux entités appariées excède la valeur 0.6, ce lien est considéré comme invalide. Cependant, les liens que nous avons considérés comme invalides n'ont pas été supprimés du processus afin de valider la décision de leur invalidation après une vérification par d'autres mesures. En effet, la distance surfacique est une mesure d'écart de position qui ne peut pas détecter des écarts de forme. Par conséquent, on suppose l'existence des entités dont la distance surfacique dépasse la valeur 0.6 et qui ont la même forme (présence d'un fort biais).

Pour mesurer les écarts de forme et de position entre les entités appariées, cinq mesures ont été employées :

- ✓ Distance surfacique
- ✓ Distance de Hausdorff (entre surfaces)
- ✓ Distance entre les moments de Zernike



- ✓ Distance entre les moments de Legendre
- ✓ Distance entre les moments géométriques invariants.

Pour le calcul des distances entre les moments de Zernike et les moments de Legendre, nous avons utilisé deux types de normalisation des entités (cf. figure II-24 (a) et (b)).

Les mesures des écarts de position et de forme des entités des 22 liens invalidés lors du processus d'appariement renforcent la décision sur l'invalidation de ces liens. Les mesures de la distance de Hausdorff varient entre 32 mètres et 214 mètres. La distance entre les moments géométriques invariants renvoie des valeurs très élevées (> 20 indiquant que leurs formes ne se ressemblent pas). Ces liens proviennent d'une mauvaise interprétation ou d'un oubli de la part de l'un des photointerprètes, puisque la majeure partie de ces liens met en correspondance des entités représentant le thème "constructions légères" avec des thèmes de type "Bois", "Broussaille" ou "Vigne". L'analyse sera donc poursuivie sur les liens restants (143 liens).

Avant d'appliquer la méthode fondée sur les mesures et leur analyse, nous avons voulu analyser les deux jeux de données d'une manière qualitative en utilisant les techniques de croisement des données et établir une matrice de confusion entre les thèmes des deux jeux de données. Les thèmes présents dans les deux jeux de données sont donnés par le tableau suivant :

Numéro	Désignation
1	Bassin
2	Bâtiment industriel
3	Bâtiment religieux
4	Bois
5	Broussaille
6	Cimetière
7	Construction légère
8	Serre
9	Silo
10	Terrain de foot
11	Terrain de tennis
12	Verger
13	Vigne

Tableau III-13 : Postes de légende des données utilisées

Suite à une étape de superposition des deux jeux de données, une comparaison de l'information sémantique a été faite "pixel à pixel". Le résultat est donné par la matrice de confusion suivante<sup>35</sup>:

<sup>35</sup> Les pourcentages sont calculés par rapport à la surface totale occupée par chaque thème du deuxième jeu de données. La ligne et la colonne "Néant" représentent respectivement l'excédent du deuxième jeu de données et l'excédent du premier jeu de données.

		photointerprète #2													
		1	2	3	4	5	6	7	8	9	10	11	12	13	Néant
photointerprète #1	1	51,87%													5,45%
	2		51,91%		0,01%									0,01%	21,36%
	3			46,03%											35,63%
	4	33,59%			93,52%	73,45%							5,92%	3,43%	9,35%
	5				1,04%								0,27%	0,12%	41,40%
	6						98,46%								2,29%
	7		0,73%		0,01%	0,01%							0,08%	0,05%	85,02%
	8							83,91%							75,88%
	9								95,35%						32,03%
	10									82,94%					0,34%
	11										90,67%				22,96%
	12				0,03%	0,03%						50,58%		3,10%	59,03%
	13				0,05%	0,34%							45,33%		32,56%
	Néant	14,54%	47,36%	53,97%	5,36%	26,17%	1,54%	16,09%		4,65%	17,06%	9,33%	43,15%	47,97%	

Tableau III-14 : Matrice de confusion entre les thèmes des jeux de données testés

La matrice de confusion montre qu'il existe une forte valeur de confusion entre le thème "Bois" du premier jeu de données et le thème "Broussaille" du deuxième jeu de données. Ceci nous amène à déduire qu'il y a une erreur d'interprétation des thèmes au moment de la restitution des jeux de données. Une autre confusion existe entre le même thème "Bois" et le thème "Bassin" du deuxième jeu de données. Une inspection visuelle des jeux de données a permis de révéler une erreur due à un oubli, le deuxième opérateur, a omis de saisir un objet "bassin" dans le deuxième jeu de données. Les autres taux de confusions (notamment au niveau de la ligne 4 et de la colonne 4 de la matrice) sont plutôt générés par l'apparition des polygones parasites lors de l'opération de superposition et proviennent à leur tour d'une mauvaise interprétation des frontières du thème "Bois".

La figure III-37 donne une illustration graphique du résultat de la superposition des deux jeux de données.

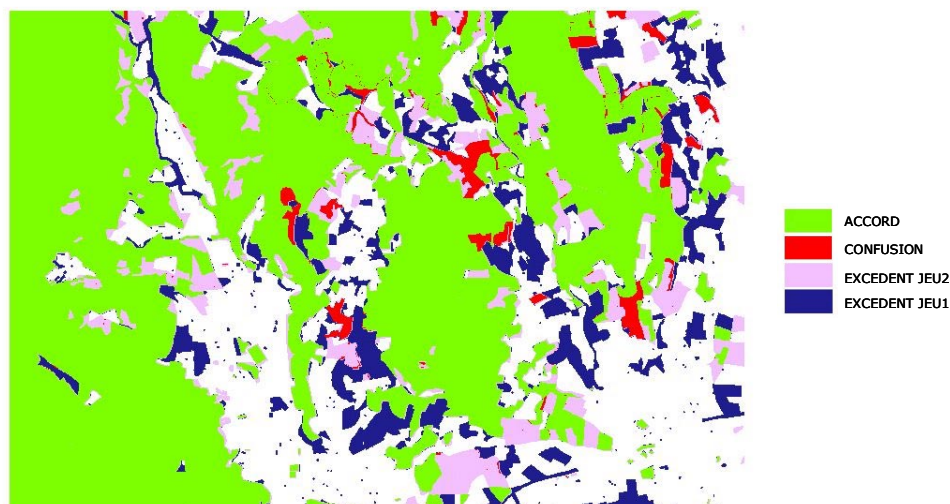


Figure III-37 : Analyse de l'information sémantique

Les techniques de superposition permettent d'avoir une idée générale sur les jeux de données à analyser sans pouvoir évaluer les écarts géométriques entre leurs entités d'une quantitative. En effet, nous appliquons la méthode développée en §III.3., qui nous

permet de détecter les différentes configurations des écarts entre les entités des deux jeux de données d'une part, et de capitaliser cette information pour des utilisations ultérieures, d'autre part.

En un premier temps, les mesures des écarts entre les entités, dont on considère que leurs liens d'appariement sont valides, subiront une analyse en composantes principales afin de décorréler les mesures initiales et poursuivre l'analyse des liens en utilisant des mesures synthétiques complètement décorréliées. La représentation des mesures initiales sur le premier plan principal montre que les distances de Zernike et les distances entre les moments de Legendre (Zer.Sep-Leg.Sep et Zer.2A2-Leg.2A2) suivent la même direction en ayant la même importance (cf. figure III-38). Cette tendance paraît raisonnable, puisque les deux types de moments utilisés peuvent être dérivés l'un de l'autre (comme nous l'avons signalé dans le chapitre II), et ils sont tous les deux porteurs d'une manière "approximative" de la même information.

Les 4 premières composantes principales générées véhiculent 96% de l'information initialement portée par les mesures d'origine et elles présentent une forte variation entre-elles. Les trois dernières sont "presque" à variance égale (et de faible valeur), donc, leur utilisation dans la suite de l'analyse n'apporte pas une information supplémentaire par rapport aux quatre premières. En effet, nous nous contentons de l'utilisation des 4 premières composantes principales lesquelles subiront une classification hiérarchique ascendante afin de pouvoir déterminer les classes des entités ayant la même configuration d'écart.

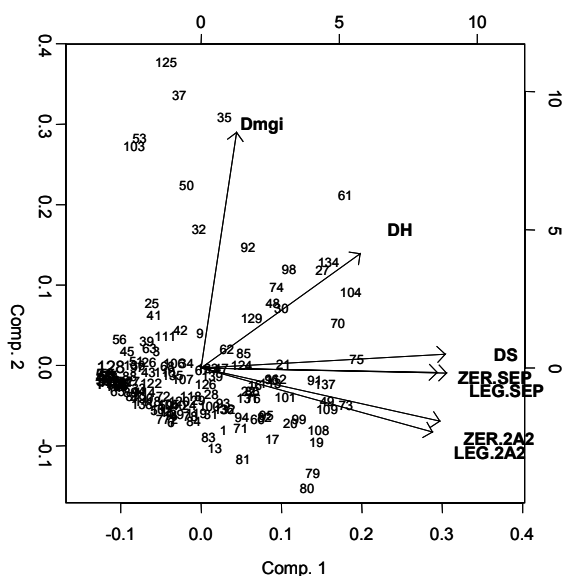


Figure III-38 : Représentation des mesures initiales sur le premier plan principal

L'arbre de la classification des valeurs des composantes synthétiques (CP1→CP4) est donné par la figure III-39.

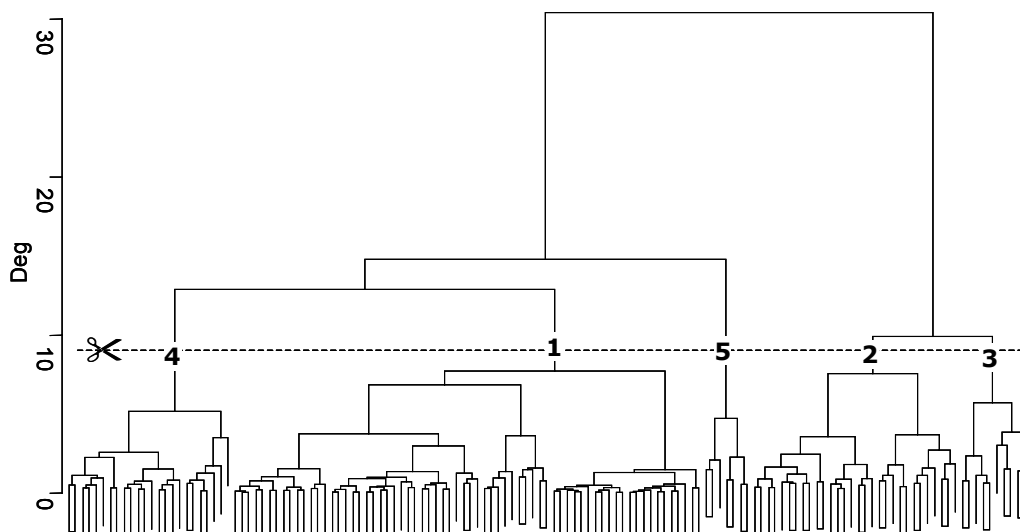
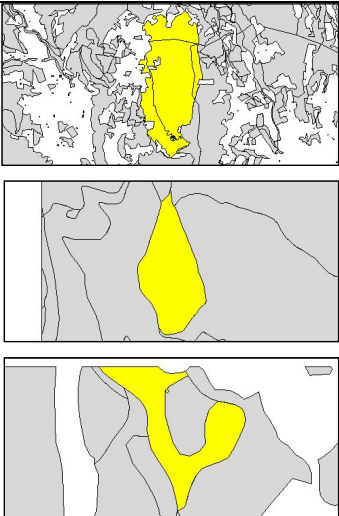
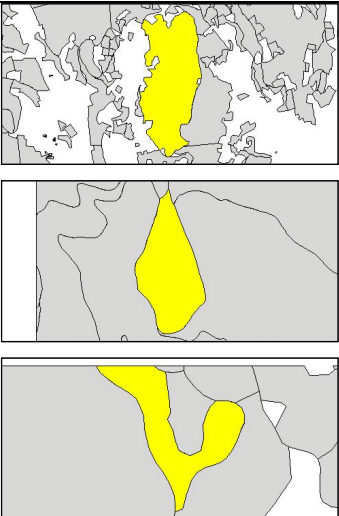


Figure III-39 : Arbre de classification

Nous rappelons que le découpage de l'arbre de classification se fait d'une manière empirique, par le choix de plusieurs seuils de coupure et l'analyse visuelle des classes générées à chaque seuil. Le choix d'un seuil est dit non concluant si nous n'arrivons pas à distinguer visuellement la différence en terme de configurations d'écart entre les classes générées. Le résultat de la coupure de l'arbre de classification est illustré sur la figure III-39, permettant de générer 5 classes dont nous donnons l'interprétation dans le tableau ci-après.

On note, pour le présent test, qu'il est prudent de ne pas procéder à une génération de règles de classification (comme pour l'exemple donné en §III.3.4., vu le nombre réduit des individus analysés (143), pouvant biaiser le résultat en générant des règles non robustes.

Classe	Exemples		Désignation
	Jeu de données #1	Jeu de données #2	
C1			Exactitude de forme et de position. Les exemples donnés illustrent des parcelles représentant des thèmes "naturels", mais la majeure partie de cette classe est composée des entités des thèmes artificiels (bâtiment, terrain, église, etc.)


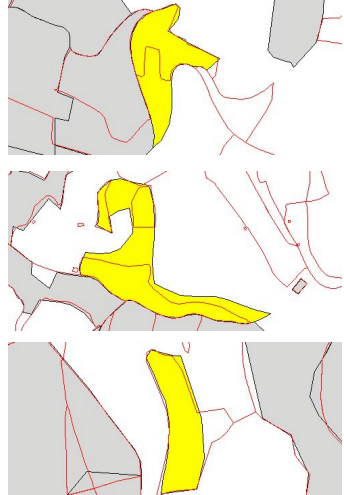
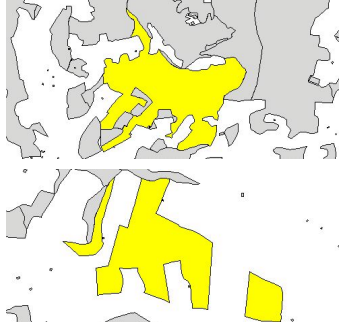
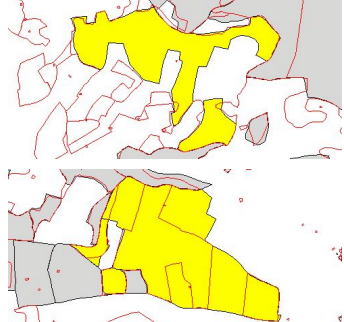
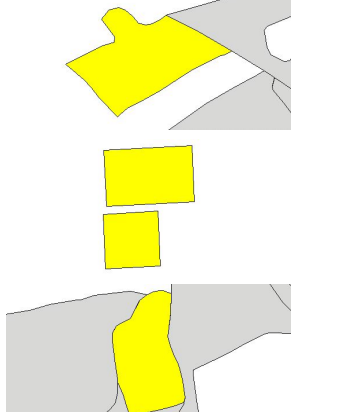
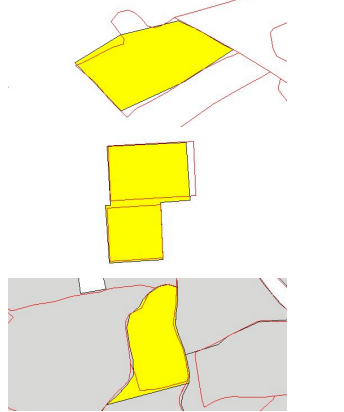
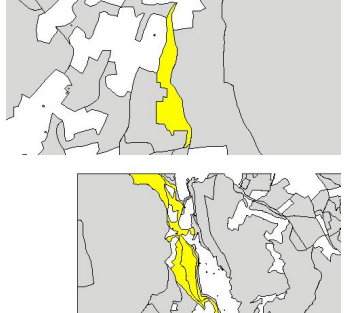
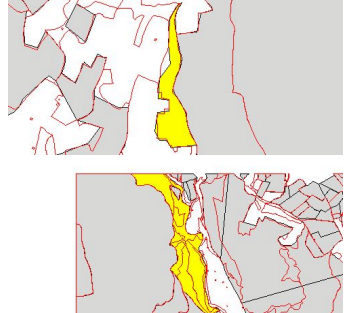
C2			<p>Occupation partielle de l'espace (moins de 50% d'occupation commune entre les thèmes) avec un fort écart de forme. Oubli d'un détail conséquent ou mauvaise identification de l'étendue de l'entité</p>
C3			<p>Occupation partielle (espace commun occupé par les deux entités &gt; 50%) avec une forte déformation. Oubli d'une partie de l'entité ou mauvaise interprétation des limites de l'entité</p>
C4			<p>Ecart moyen en position avec une déformation moyenne en forme. Oubli d'un petit détail de l'une des deux entités appariées. Effet d'agrégation.</p>
C5			<p>Ecart moyen en position avec une forte déformation de la forme. Limites des thèmes "mal" identifiées.</p>

Tableau III-15 : Interprétation des classes d'écart

Nous signalons à ce niveau que l'interprétation est faite en tenant compte du contexte du présent exemple. En d'autres termes, il est impossible de transposer les règles générées dans le cadre du test Cadastre-BDTopo®. Les mesures permettant de

définir une classe comme une classe contenant des entités correctes en forme et en position dans un contexte donné, peuvent définir autrement cette même classe dans un autre contexte. Cependant, la méthode d'analyse proposée reste générique, seuls les seuils et les règles de classification changent d'un contexte à un autre.

L'utilisation combinée de la classification avec l'information sémantique permet de générer des cartes de qualité. Ce type de cartes peut être considéré comme une alternative mieux élaborée [Bel Hadj Ali 2001b] que les cartes de qualité proposées par [Bel Hadj Ali & Vauglin 1999] et elles sont réalisées sur la base de l'utilisation d'une seule mesure.

La figure III-40 illustre un exemple d'une carte de qualité.





Figure III-40 : Exemple d'une carte de qualité



### III.4. SYNTHÈSE

Ce chapitre a été organisé en deux parties. La première partie a présenté une méthode d'appariement des données géographiques surfaciques et la deuxième une méthode d'analyse des mesures entre les entités appariées pour détecter les écarts de forme et de position entre elles.

L'appariement des données géographiques est considéré comme une étape nécessaire, voire primordiale, avant toutes actions de contrôle de la qualité des données. L'appariement des données géographiques peut se faire selon plusieurs critères (sémantiques, topologiques et géométriques). La méthode développée dans le cadre de cette thèse utilise uniquement l'information géométrique pour la mise en correspondance des entités géographiques issues de deux bases de données différentes, mais représentant les mêmes phénomènes du monde réel. Deux entités sont considérées *a priori* comme candidates à l'appariement, si leur intersection est non nulle. L'utilisation de cette hypothèse peut passer à côté de quelques liens d'appariement dans le cas où les deux entités présentent un très fort biais en position de manière à ne pas se croiser. Les essais sur des jeux de données réels ont montré que de telles configurations sont relativement rares. Un ensemble de mesures a été également donné dans cette partie. Ces mesures sont destinées à supprimer les liens parasites et à doter la méthode d'appariement d'une autonomie décisionnelle au niveau de la validité des liens d'appariement. La méthode permet également la détection des liens d'appariement multiples mettant en correspondance un ensemble d'entités d'une base de données avec l'ensemble homologue d'une autre base de données.

La robustesse de la méthode a été testée en l'utilisant pour appairer des jeux de données divers et variés. Les jeux de données ont été choisis de manière à couvrir une large palette des problèmes rattachés à l'appariement : en utilisant des jeux de données saisis avec les mêmes spécifications à la même actualité avec la même source de données; aux jeux de données de spécifications différentes; à des actualités différentes et avec des sources de données différentes. Les tests ont montré une robustesse de la méthode pour appairer des données de type "Bâtiment" représentées dans l'espace géographique. De plus, ils n'assurent pas un pavage complet de l'espace facilitant ainsi la tâche de suppression des liens parasites. Par ailleurs, les tests sur des jeux de données assurant un pavage complet ou semi-complet ont permis de soulever quelques problèmes liés, soit aux mauvaises interprétations des limites des thèmes, soit aux différences dans le découpage de l'espace pour limiter les thèmes. Ces problèmes ont permis de revoir les valeurs des filtres à utiliser pour éviter les liens parasites pouvant induire des faux appariements. Le résultat de l'appariement de deux jeux de données d'occupation du sol saisis avec les mêmes spécifications, les mêmes données sources et à la même actualité est très satisfaisant puisqu'on arrive à un appariement près de 80% sans l'intervention humaine. Les tests réalisés sur des jeux de données représentant des données forestières, acquises à des dates différentes et saisis en utilisant des sources de données sans les mêmes caractéristiques ont montré qu'on arrive à identifier les

structures géométriques "homologues" entre les deux jeux de données. Cependant, il se trouve qu'à l'issue de l'appariement de ce type de données, on trouve des appariements multiples d'une taille conséquente essentiellement générés par la différence de granularité et du niveau de détails entre les deux bases. L'appariement de ce type de données pourrait être affiné par l'utilisation de l'information sémantique, à l'exception d'un seul cas où l'appariement est réalisé pour contrôler la qualité sémantique d'une base de données par rapport à une autre.

La deuxième partie de ce chapitre a présenté une méthode fondée sur une utilisation combinée des mesures entre les entités appariées pour décrire les écarts de forme et de position. Partant du principe que l'utilisation d'une seule mesure est incapable de détecter tous écarts sans confusions, nous avons opté pour une utilisation de plusieurs mesures dont chacune apporte une information complémentaire aux autres. Dans un premier temps, une analyse statistique des mesures utilisées est faite pour analyser l'existence d'une éventuelle corrélation entre les mesures et utiliser les mesures les plus pertinentes. L'analyse des mesures a été faite par la méthode dite "analyse en composantes principales" en générant des mesures synthétiques ayant une corrélation nulle entre-elles. L'analyse des mesures synthétiques pour la détection des différentes configurations des écarts géométriques entre les entités appariées est difficile à interpréter. D'une manière heuristique, on peut considérer que deux liens dont les mesures sont comparables doivent avoir une même configuration d'écarts géométriques.

L'utilisation des techniques de la classification hiérarchique permet la présentation du résultat sous la forme d'un arbre de classification que nous avons découpé pour détecter les différentes classes d'écarts. Le découpage de l'arbre est fait d'une manière itérative et le résultat est stocké afin de conserver la hiérarchie entre les différentes classes générées. L'information sur la hiérarchie entre les classes constitue une information capitale pour agréger l'information sur la qualité, ainsi que sa présentation selon l'utilisation des données [Faïz & Boursier 1994; Bel Hadj Ali 2001b].

Chaque lien -individu- (et par conséquent, chaque entité) est décrit par un ensemble de mesures. La représentation de ces individus dans l'espace à  $n$  dimensions dont les axes sont matérialisés par les métriques utilisées montre qu'ils forment un nuage de points fortement groupés. Cette configuration rend très difficile une discrimination des classes d'une manière nette, ce qui explique la confusion et l'indécision dans l'affectation de quelques entités dans des classes différentes alors que leurs mesures sont proches. Ce problème est largement connu dans le domaine de l'apprentissage sous le nom de "*l'overfitting*" [Mustière 2001].

La méthode proposée dans cette partie consiste donc à faire subir aux mesures synthétiques une classification non supervisée pour déterminer les classes des écarts. Ces classes ont été interprétées en leur donnant une signification physique en fonction des entités qu'elles contiennent. En conclusion de ce chapitre, nous proposons le schéma suivant (cf. figure III-41) pour illustrer le fonctionnement global de la méthode.

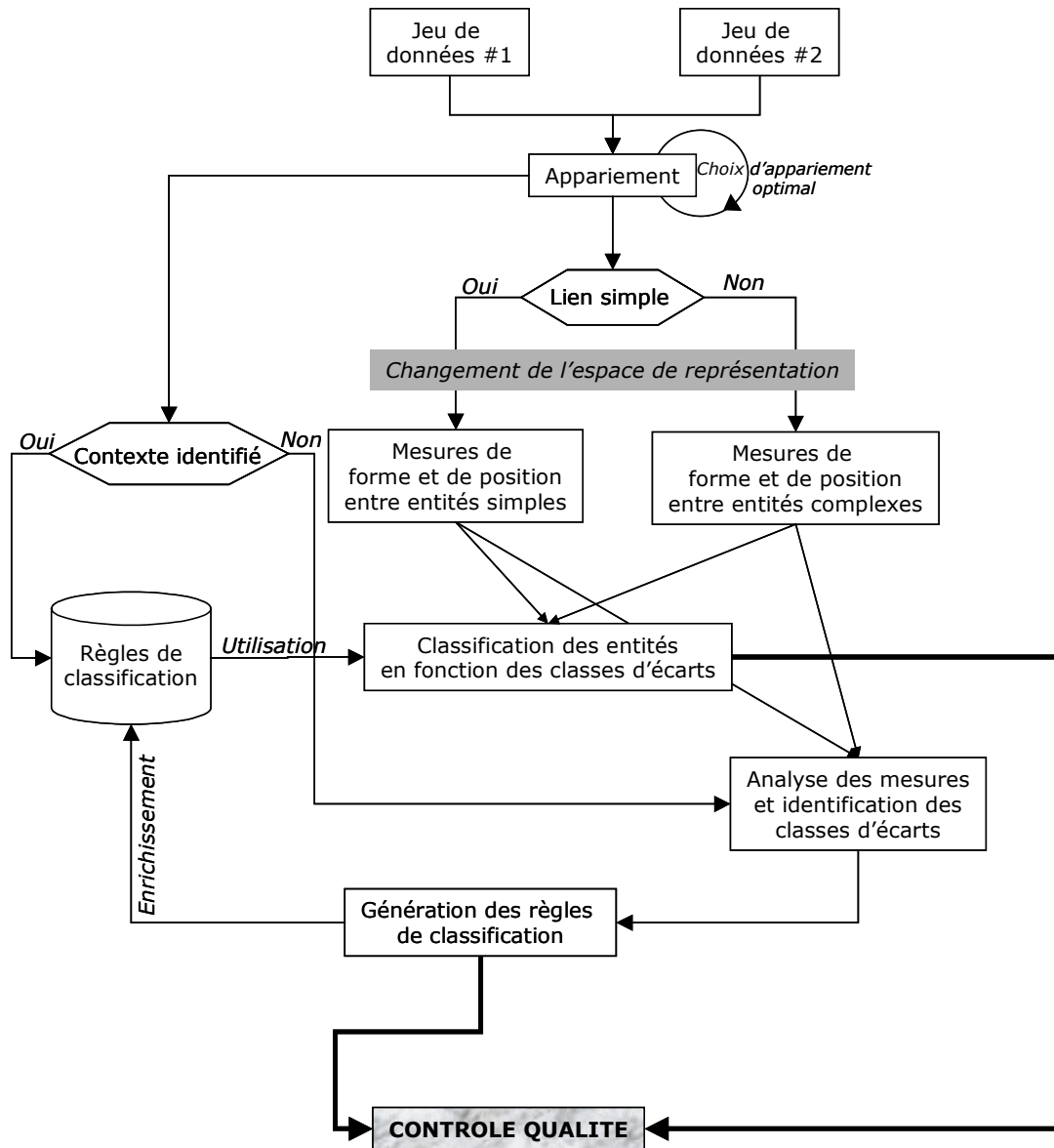


Figure III-41 : Fonctionnement global de la méthode (de l'appariement au contrôle qualité)

La méthode présentée par la figure a été également mise en œuvre comme un module en complément du prototype d'appariement. L'interface graphique de ce module est présentée à l'annexe D.

## **CONCLUSION GENERALE**

L'information géographique est de plus en plus disponible sous une forme numérique, permettant plus de facilité pour sa manipulation, sa gestion, son stockage et son analyse. En contre partie, d'autres problèmes ont surgi touchant plusieurs domaines d'application de l'information géographique, notamment ceux sur la qualité des données numériques. Les travaux rapportés dans cette thèse avaient pour but d'étudier la qualité de la géométrie des entités surfaciques dans les bases de données géographiques. L'objectif visé est donc de proposer des outils et des méthodes pour décrire les écarts de forme et de position des entités surfaciques dans les BDG.

Notre travail s'est déroulé en trois parties principales qui correspondent aux trois chapitres de ce rapport.

Le premier chapitre "*Qualité des données géographiques*" a permis de présenter les concepts existants en matière de qualité des données géographiques en mettant l'accent sur la qualité géométrique. Nous avons retracé un état de l'art sur la qualité géométrique des primitives ponctuelles et linéaires en présentant les outils qui la manipulent. Les outils proposés pour les contrôles ponctuel et linéaire sont insuffisants pour rendre compte des écarts de la géométrie des entités surfaciques, tant sur le plan de la position que sur le plan de la forme. Cette insuffisance est essentiellement due à la définition de l'outil de mesure lui-même (puisque la majeure partie des outils est développée pour mesurer les écarts de position) ou à la faiblesse descriptive de la représentation en liste de points, représentant l'entité surfacique par son simple contour. Nous avons montré que la représentation des primitives surfaciques par une liste ordonnée de points (représentation en mode vecteur) est insuffisante pour rendre compte de toutes les caractéristiques géométrique d'une entité donnée, puisqu'elle ne révèle de la géométrie qu'un aspect purement ponctuel. En conséquence, les mesures qui utilisent la représentation en mode vecteur sont également insuffisantes pour décrire les écarts de forme des entités surfaciques.

Le deuxième chapitre "*Représentations et mesures*" est donc consacré à la recherche de nouveaux espaces de représentations et s'est attaché à les doter de mesures appropriées à la description des écarts géométriques. En effet, un ensemble d'espaces métriques a été proposé permettant de mieux représenter la géométrie des entités surfaciques et de mieux cerner leur qualité.

Le troisième chapitre a présenté l'application des mesures définies dans le deuxième chapitre par le développement d'une méthode d'appariement des bases de données surfaciques. Nous avons montré que l'utilisation d'une seule mesure est insuffisante pour décrire d'une manière globale les écarts géométriques. A cet effet, il est nécessaire de combiner les mesures afin de pouvoir détecter les différentes configurations des écarts en tenant compte des écarts de forme et de position en même temps.

## Apports et résultats

### Changements de représentation

Pour la définition de nouvelles représentations, nous avons choisi de classer les entités surfaciques en deux catégories : les entités simples ou élémentaires et les entités complexes représentant les polygones avec trous ou les agrégats des polygones disjoints ou non.

La représentation par une liste de points peut être suffisante pour être utilisée dans un but de description des écarts de position des entités surfaciques élémentaires en les traitant à travers leurs contours. Mais, elle reste insuffisante pour la description de la forme. En effet, nous avons proposé trois nouvelles représentations de la géométrie des entités surfaciques simples. La première représentation utilise les angles entre les segments du contour et permet de décrire le polygone par une fonction en escalier (cf. §II.3.1.2.). Cette représentation a montré tout son intérêt pour représenter des entités décrivant des objets géographiques "artificiels" de type bâtiment, par contre, son utilisation sur des polygones décrivant des objets géographiques "naturels" s'avère problématique à cause de la présence d'un nombre conséquent de segments par entités. Ceci se traduit par des micro-variations sur le plan angulaires qui peuvent renvoyer des résultats erronés pour décrire la forme. Pour cela, nous avons défini une deuxième représentation que nous avons appelé "signature polygonale". Elle se présente comme une fonction représentant les distances euclidiennes entre le centre de masse du polygone et les points composant son contour (cf. §II.3.1.3.). Cette représentation permet de pallier le manque constaté pour la première représentation. La troisième représentation proposée procède par un changement de l'espace cartésien vers l'espace des fréquences (cf. §II.3.1.4.). L'entité est donc représentée par un ensemble de valeurs des amplitudes et des angles de phase des fréquences.

Les représentations proposées respectent le critère de l'unicité (cf; §II.2.1) sous certaines conditions, ce qui leur procure une robustesse au niveau de la description de la forme des entités surfaciques élémentaires. Par ailleurs, ces représentations demeurent inopérantes pour décrire les entités complexes.

Pour les entités complexes, nous avons proposé de procéder par une étape de discrétisation des entités surfaciques (représentation en mode *raster*) avant de proposer de nouvelles représentations, dans le but d'utiliser la totalité de l'information géométrique de l'entité surfacique (portée par le contour et par l'intérieur). En effet, une méthode a été proposée pour le choix du pas optimal d'échantillonnage pour mieux représenter l'entité analysée sans dégrader sa forme par rapport à sa représentation d'origine.

Les représentations alternatives proposées dans cette thèse pour décrire les entités complexes reposent sur l'utilisation des techniques des moments mathématiques (cf. §II.3.2.2.). En effet, l'utilisation des moments permet de prendre en compte toutes les

propriétés géométriques d'une entité donnée (telles par exemple : aplatissement, élongation, symétrie, etc.). Trois types de moments ont été utilisés, en l'occurrence les moments géométriques, les moments de Legendre et les moments de Zernike. Les moments géométriques sont définis dans une base non orthogonale rendant difficile l'analyse de leur représentativité. Par contre, les moments de Legendre et de Zernike sont définis dans des bases orthogonales puisqu'ils s'appuient sur les polynômes des mêmes noms qui constituent eux aussi une base orthogonale de part leurs définitions. La propriété d'orthogonalité assure une bijection entre l'espace des moments et l'espace des formes ce qui permet de faire l'aller-retour entre les deux espaces, d'analyser la représentativité de ces moments et de définir l'ordre optimal des moments à utiliser pour mieux représenter les entités surfaciques par les moments. Nous avons proposé dans cette thèse une méthode utilisant la reconstruction des entités à partir de leurs valeurs de moments et apportant une réponse à la question de la représentation d'une entité par un ensemble fini de valeurs de moments. La représentation dans l'espace des moments peut également être utilisée pour les entités simples.

Après avoir représenté les entités surfaciques dans les nouveaux espaces, nous avons procédé à une étape de mesure en dotant chacun de ces espaces d'une métrique (ou plusieurs) permettant la détection des écarts géométriques entre les entités mesurées. Les mesures définies ont été testées, dans un premier temps, sur des données synthétiques et bruitées d'une manière maîtrisée afin de tester leur robustesse et leur fidélité pour rendre compte des déformations. Dans un deuxième temps, les mesures ont été testées sur des jeux de données réelles sur lesquelles nous avons introduit un bruit progressif en le simulant d'une manière réaliste par l'utilisation d'un modèle statistique de description des erreurs géométriques dans les bases de données. Les résultats obtenus ont montré que les indicateurs et les métriques définis évoluent dans le même sens que l'amplitude de bruit. Ces tests ont permis de calibrer les mesures dans le cas de leur utilisation dans le contexte d'estimation d'erreurs.

### **Appariement des données surfaciques**

Pour s'affranchir des techniques de sondage et de choix d'échantillons représentatifs, nous avons développé dans le cadre des travaux de cette thèse une méthode d'appariement qui permet la mise en correspondance des entités surfaciques provenant de deux jeux de données différents, mais représentant le même phénomène physique.

La méthode s'appuie, en une première approximation, sur le fait que deux entités sont supposées représenter le même phénomène réel si elles occupent une partie commune de l'espace. Cette hypothèse permet de retrouver toutes les entités surfaciques candidates à l'appariement. D'autres filtres ont été proposés afin d'éliminer tous les liens qui ne reflètent pas une réalité physique (entités qui occupent en commun un espace très réduit par rapport à leurs tailles respectives). La méthode est dotée d'outils lui

permettant de détecter les liens complexes qui mettent en correspondance un ensemble d'entités d'un jeu de données avec un ensemble d'entités d'un autre jeu de données. Nous avons également doté la méthode d'un outil lui permettant d'analyser les liens multiples et d'affiner l'appariement afin d'aboutir aux liens optimaux.

La méthode utilise uniquement l'information géométrique et des mesures prenant en compte les surfaces et les contours. Cette méthode requiert, de la part de l'utilisateur, un seul paramètre qui est un seuil de coupure sur une fonction mesurant l'inclusion relative entre les entités candidates à l'appariement (cf. §II.4.2.3.).

La méthode proposée a été implémentée et sa robustesse a été testée sur des jeux de données choisis de manière à couvrir une large palette de problèmes liés à l'appariement. Les tests ont montré qu'en utilisant la méthode proposée, on peut aboutir à un appariement à plus de 90% en "tout automatique" pour les jeux de données représentant des thèmes "artificiels" (de type "bâtiment", par exemple). Les tests d'appariement des données représentant des thèmes "naturels" (de type occupation de sol) ont permis de soulever de nouveaux problèmes. Ces problèmes consistent essentiellement en la génération des liens multiples de fortes cardinalités, généralement dues à des mises en correspondance invalides entre les entités simples candidates à l'appariement. La résolution de ces problèmes est faite en modifiant le seuil de coupure et en vérifiant visuellement le résultat de l'appariement. En effet, le seuil de coupure sur les valeurs de la fonction d'inclusion varie selon des données utilisées.

Enfin, la méthode proposée n'est pas utilisable uniquement pour des fins de contrôle de qualité (but principal pour lequel a été initialement développée), mais également pour d'autres applications telles que, par exemple, la mise à jour de données ou la mise en œuvre des serveurs multi-échelles. L'utilisation de la méthode pour d'autres applications est aisément envisageable du fait que les résultats de toutes les étapes de l'appariement sont capitalisés et sauvegardés dans un modèle que nous avons développé pour cet effet.

### **Analyse des mesures et définition d'une typologie des écarts**

Les données géographiques sont généralement entachées d'erreurs provenant de plusieurs sources (erreur de saisie, différence dans l'interprétation des limites des thèmes, etc.). Ces erreurs induisent souvent des écarts de position ou des écarts de forme entre les entités qui sont sensées représenter une même réalité physique (résultat de l'appariement). En effet, toutes les entités appariées ont subi une série de mesures afin d'évaluer les écarts de géométrie.

Nous avons montré que l'utilisation d'une seule mesure est insuffisante pour traduire la nature de l'écart entre les entités appariées. En conséquence, l'utilisation combinée de plusieurs mesures s'impose du fait que chaque mesure utilisée est porteuse d'une information particulière sur la description de l'écart géométrique. Par ailleurs, il est difficile d'interpréter "manuellement" les valeurs des mesures pour établir une typologie de l'écart, ce qui nous a poussé à proposer une méthode permettant d'identifier d'une manière automatique les différentes classes d'écart entre les entités des jeux de



données appariées. Ceci revient à analyser un tableau de données de M lignes (représentant les liens d'appariement obtenus) et de N colonnes (représentant les mesures utilisées). Chaque lien est donc identifié par l'ensemble des mesures entre les entités qui le composent.

Les expérimentations que nous avons effectuées ont révélé l'existence d'une corrélation entre quelques mesures utilisées. Afin d'éviter cette corrélation, nous avons procédé à une analyse en composantes principales, en remplaçant les mesures initiales par des mesures synthétiques (les composantes principales) qui sont complètement décorréliées. En plus de l'annulation de l'effet de redondance de l'information portée par les mesures initiales, l'analyse en composantes principales permet de mieux connaître le comportement des mesures les unes par rapport aux autres et de voir la contribution de chacune d'entre elles dans l'identification des différentes classes d'écart.

La détection des différentes configurations d'écart est réalisée par l'utilisation d'une classification non supervisée sur l'ensemble des mesures synthétiques. Les entités de chacune des classes obtenues ont été analysées visuellement afin de donner une description factuelle des écarts géométriques qui les caractérisent.

Nous avons classé les écarts géométriques entre les entités surfaciques en fixant trois niveaux d'écart pour la forme et la position (nul, faible, moyen et fort). En effet, la combinaison de ces trois niveaux permet d'obtenir 13 configurations d'écart géométriques. Il est à noter qu'il est très rare de détecter la totalité des 13 configurations dans un appariement donné entre deux jeux de données.

Enfin, l'utilisation du résultat de la classification avec les mesures initiales permet de générer des règles de classification. Ces règles sont stockées dans une base "de règles" qui sert par la suite à identifier la nature des écarts géométriques si l'on se trouve à qualifier les entités issues de l'appariement de deux jeux de données dont le contexte est similaire à celui qui a permis de générer les règles de classification.

## **Perspectives**

Nous avons donc, dans les travaux réalisés dans le cadre de cette thèse, pris parti de chercher de nouvelles représentations pouvant traduire au mieux les caractéristiques géométriques des entités surfaciques dans les bases de données géographiques, afin de s'affranchir de la représentation sous la forme d'une liste de points. Les travaux de cette thèse se sont également intéressés au développement d'une méthode de mise en correspondance des entités surfaciques provenant de deux jeux de données différents et à l'utilisation de nouvelles représentations à des fins de qualification géométrique, tant sur le plan de la forme que sur le plan de la position.

La méthode d'appariement développée a été testée sur un ensemble réduit de jeux de données. En effet, le choix des différents seuils a été fait d'une manière empirique. Il se pose alors la question de la validité des valeurs des seuils choisies. Ces questions

nécessitent beaucoup d'expérimentation sur d'autres lots de jeux de données afin d'être résolues et de valider la méthode d'une manière définitive.

Dans le même esprit, des tests doivent être poursuivis pour que la méthode proposée pour détecter les différentes configurations d'écarts (voire la totalité) entre les entités des jeux de données soit utilisable.

La méthode propose la génération des règles de classification qui sont réutilisables dans le cas où l'on se trouve amené à analyser des données similaires à celles participant à la création de ces règles. Par ailleurs, pour chaque contexte d'appariement (exemple: Cadastre-BDTopo®), on doit poursuivre les expérimentations sur plusieurs lots de données afin de couvrir la totalité des 13 configurations d'écarts et de consolider d'une manière définitive la base des règles. Ainsi, les méthodes proposées dans cette thèse et le prototype réalisé peuvent être d'une grande utilité, tant du côté de producteur pour contrôler ses données que du côté de l'utilisateur pour toutes les applications qui nécessitent *a priori* une analyse de la géométrie.

Enfin, nous n'avons gardé des représentations utilisées que les résultats des mesures, sans pour autant conserver les nouvelles représentations dans la base de données. Ce choix a été adopté afin de ne pas surcharger les bases de données à analyser. Par ailleurs, une étude plus approfondie de ces représentations et des mesures rattachées pourrait être envisagée dans la perspective de développer une nouvelle génération de SIG. Cette nouvelle génération pourrait utiliser une nouvelle approche introduite par [Goodchild 1999] consistant à utiliser les résultats des mesures ainsi que la définition des fonctions et des règles pour dériver les coordonnées des entités géographiques à partir de ces mesures (*Measurement based GIS*). Ceci doit permettre, à terme, la construction d'un SIG hiérarchisé à différentes échelles, intégrant totalement les modèles d'erreurs et permettant d'analyser l'impact de la propagation des erreurs de position et de la forme à travers les opérations d'analyse. Une étude pourrait également être envisagée, en étudiant les autocorrélations des écarts par l'utilisation des mesures proposées dans le cadre de ce travail, afin de mieux définir la précision relative des primitives surfaciques.

## **ANNEXES**

## ANNEXE A : CALCUL DES MOMENTS GEOMETRIQUES PAR L'UTILISATION DES CONTOURS

Les moments géométriques d'ordre  $p+q$  d'une entité surfacique quelconque définie sur un domaine compact  $\xi$  sont donnés par :

$$m_{p,q} = \iint_{\xi} x^p y^q f(x,y) dx dy \quad [A.1]$$

$f$  désigne la fonction d'intensité des pixels de l'entité surfacique. Dans le contexte de notre étude, les entités sont traitées comme étant des images binaires. Par conséquent la fonction  $f$  prend deux valeurs : soit 1 si le pixel est à l'intérieur de l'entité et 0 ailleurs.

Pour le calcul pratique des moments, la région  $\xi$  est échantillonnée, en un ensemble de pixels dont l'intensité de chacun est égale à l'unité. L'intégrale de l'équation A.1 peut être simplement calculée par une simple sommation sur la région  $\xi$ .

La méthode du calcul des moments d'une entité surfacique par l'utilisation de son contour est fondée sur le théorème de Green-Riemann. Ce théorème permet de calculer l'intégrale d'une fonction sur le domaine bi-dimensionnel  $\xi$  en le réduisant à une intégrale le long de la frontière du domaine  $\xi$ . Ce théorème est énoncé de la manière suivante :

*Soient  $P$  et  $Q$  deux fonctions continues et dérivables sur le domaine  $\xi$ , et soit  $c(t)$  la frontière du domaine  $\xi$ . Si  $b$  est dérivable et orientée dans le sens trigonométrique direct, une intégrale sur le domaine  $\xi$  peut être réduite en une intégrale le long de la frontière  $b$  de  $\xi$  de la manière suivante :*

$$\iint_{\xi} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dx dy = \int_c P dx + Q dy \quad [A.2]$$

Il est clair que pour le calcul de l'intégrale d'une fonction arbitraire  $F$  sur le domaine  $\xi$  (comme c'est le cas des moments où la fonction  $F(x, y) = x^p y^q$ ), elle doit être décomposée en deux fonctions :  $\partial P/\partial y$  et  $\partial Q/\partial x$ . On note que la décomposition n'est pas unique et que le choix de la décomposition est essentiellement arbitraire.

### Calcul des moments des entités polygonales simples

#### Introduction

Soit, d'une manière générale, la courbe  $c(t) = (x(t), y(t))$ ,  $t \in [t_1, t_2]$ , et soit  $f$  une fonction continue sur un domaine  $c$  inclus dans  $\mathbb{R}^2$ . Les deux intégrales suivantes existent et sont définies par :

$$\int_c f(x, y) dx = \int_{t_1}^{t_2} f(x(t), y(t)) x'(t) dt \quad [A.3]$$

$$\int_c f(x, y) dy = \int_{t_1}^{t_2} f(x(t), y(t)) y'(t) dt \quad [A.4]$$

Soient  $P(x, y)$  et  $Q(x, y)$  deux fonctions continues, l'égalité suivante peut avoir lieu :

$$\int_c P(x, y) dx + Q(x, y) dy = \int_c P(x, y) dx + \int_c Q(x, y) dy \quad [A.5]$$

Les intégrales, le long d'une courbe, présentent également d'autres propriétés importantes. Nous en rappelons quelques-unes que nous allons utiliser par la suite.

Soient  $c_1(t)$ ,  $t \in [t_1, t_2]$  et  $c_2(t)$ ,  $t \in [t_2, t_3]$  deux courbes avec  $c_1(t_2) = c_2(t_2)$ , et soit  $c = c_1 \cup c_2$ , l'intégrale curviligne le long de la courbe  $c$  :

$$\int_c f(x, y) dx = \int_{c_1} f(x, y) dx + \int_{c_2} f(x, y) dx \quad [A.6]$$

Si l'orientation de la courbe  $c$  est inversée, le signe de l'intégrale de la fonction  $f$  le long de la courbe  $b$  sera également inversé :

$$\int_c f(x, y) dy = - \int_{c'} f(x, y) dy \quad [A.7]$$

### Cas des polygones simples

Les équations précédemment présentées (A.3 à A.7) représentent les principaux outils à utiliser pour calculer les moments géométriques d'une entité à travers l'utilisation de son contour.

Soit un polygone  $P$  défini par  $n$  points  $p_i = (x_i, y_i)$ ,  $i \in \{0, \dots, n\}$  avec  $p_0 = p_n$ . Le contour  $c$  du polygone peut être considéré comme une courbe linéaire continue par morceaux et est représentée par une union de  $n$  segments de la manière suivante :

$$c(t) = \bigcup_{i=1}^n c_i(t); \quad [A.8]$$

avec  $c_i(t)$ ,  $\forall t \in [0, 1]$  s'écrit sous la forme :

$$c_i(t) = t p_i + (1-t) p_{i-1} \quad [A.9]$$

En effet, les fonctions des coordonnées, ainsi que leurs dérivées nécessaires pour le calcul de l'intégrale curviligne sont données par :

$$x_i(t) = t x_i + (1-t) x_{i-1} \quad [A.10]$$

$$y_i(t) = t y_i + (1-t) y_{i-1} \quad [A.11]$$

$$x'_i(t) = x_i - x_{i-1} \quad [A.12]$$

$$y'_i(t) = y_i - y_{i-1} \quad [A.13]$$

Cependant, n'importe quelle intégrale curviligne le long de la frontière  $c(t)$  peut être calculée de la manière suivante :

$$\int_c P dx + Q dy = \sum_{i=1}^n \int_{c_i} P dx + Q dy \quad [A.14]$$

### Calcul des moments

Le moment d'ordre  $p+q$  d'une région  $\xi$  est donné par l'équation A.1. L'utilisation de l'équation A.2 nécessite la décomposition de  $x^p y^q$  en  $\partial Q/\partial x$  et  $\partial P/\partial y$ . Pour des raisons de simplicité<sup>36</sup>, nous choisissons la décomposition suivante :

$$\frac{\partial Q}{\partial x} = x^p y^q \text{ et } \frac{\partial P}{\partial y} = 0 \quad [A.15]$$

et donc nous aurons :

$$P(x, y) = 0 \text{ et } Q(x, y) = \frac{1}{p+1} x^{p+1} y^q \quad [A.16]$$

Alors, le moment  $m_{p,q}$  d'ordre  $p+q$  de la région  $\xi$  peut être calculé de la manière suivante :

$$m_{p,q} = \iint_{\xi} x^p y^q dx dy = \int_c \frac{1}{p+1} x^{p+1} y^q dy \quad [A.17]$$

En utilisant l'équation A.14, l'intégrale de l'équation A.17 peut être calculée en sommant les intégrales curvilignes le long de tous les segments composant le contour du polygone. Chaque terme de cette sommation est donné par :

$$\begin{aligned} & \int_{c_i} \frac{1}{p+1} x^{p+1} y^q dy \\ &= \frac{1}{p+1} \int_0^1 x_i(t)^{p+1} y_i(t)^q y'_i(t) dt \\ &= \frac{1}{p+1} \int_0^1 (tx_i + (1-t)x_{i-1})^{p+1} (ty_i + (1-t)y_{i-1})^q (y_i - y_{i-1}) dt \\ &= \frac{1}{p+1} (y_i - y_{i-1}) \int_0^1 \left( \sum_{k=0}^{p+1} C_k^{p+1} x_i^k x_{i-1}^{p+1-k} t^k (1-t)^{p+1-k} \right) \left( \sum_{l=0}^q C_l^q y_i^l y_{i-1}^{q-l} t^l (1-t)^{q-l} \right) dt \\ &= \frac{1}{p+1} (y_i - y_{i-1}) \int_0^1 \sum_{k=0}^{p+1} \sum_{l=0}^q C_k^{p+1} C_l^q x_i^k x_{i-1}^{p+1-k} y_i^l y_{i-1}^{q-l} t^{k+l} (1-t)^{p+q+1-k-l} dt \\ &= \frac{1}{p+1} (y_i - y_{i-1}) \sum_{k=0}^{p+1} \sum_{l=0}^q C_k^{p+1} C_l^q x_i^k x_{i-1}^{p+1-k} y_i^l y_{i-1}^{q-l} \int_0^1 t^{k+l} (1-t)^{p+q+1-k-l} dt \\ &= (y_i - y_{i-1}) \sum_{k=0}^{p+1} \sum_{l=0}^q a_{k,l}^{p+1,q} x_i^k x_{i-1}^{p+1-k} y_i^l y_{i-1}^{q-l} \end{aligned}$$

avec

$$a_{k,l}^{p+1,q} = \frac{1}{(p+q+2)(p+1)} \frac{C_k^{p+1} C_l^q}{C_{k+l}^{p+q+1}}$$

<sup>36</sup> Nous faisons remarquer à ce niveau que le choix de la décomposition n'est pas unique. Le choix, que nous avons fait, a pour majeure argumentation la simplification du calcul. Par ailleurs, nous pensons que le choix d'une autre décomposition aboutira aux mêmes résultats finaux, puisque nous avons effectué une étape de canonication pour éviter la dissymétrie de calcul selon les axes des abscisses et des ordonnées.

Tout calcul fait, on obtient :

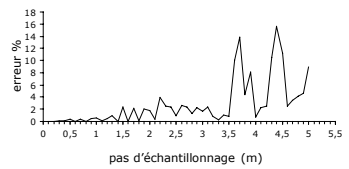
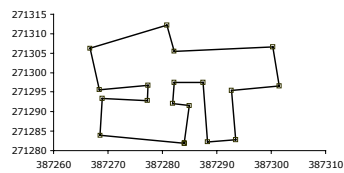
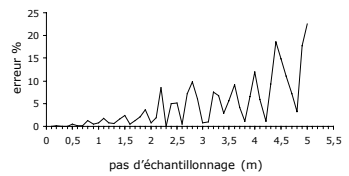
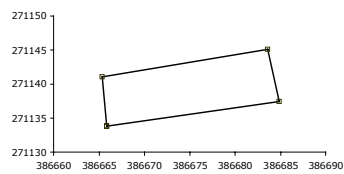
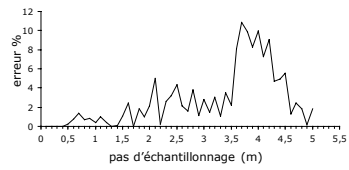
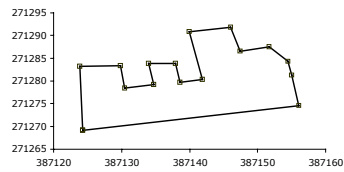
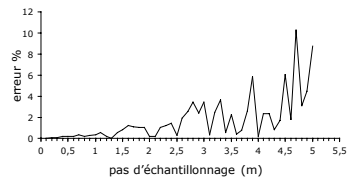
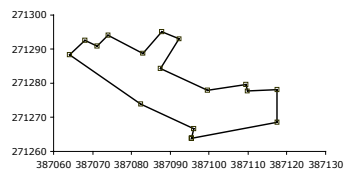
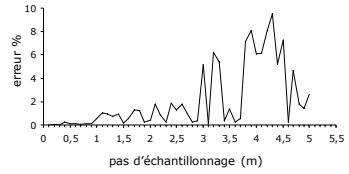
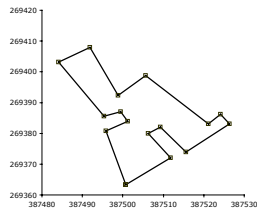
$$m_{p,q} = \frac{1}{(p+q+2)(p+q+1)C_p^{p+q}} \sum_{i=1}^n (x_{i-1}y_i - x_i y_{i-1}) \sum_{k=0}^p \sum_{l=0}^q C_l^{k+1} C_{q-l}^{p+q-k-1} x_i^k x_{i-1}^{p-k} y_i^l y_{i-1}^{q-l}$$

Pour normaliser les moments, il suffit de les diviser par la valeur de la surface donnée par :

$$a = \frac{1}{2} \sum_{i=1}^n x_{i-1}y_i - x_i y_{i-1}$$

Bien que cette méthode ait le mérite de traiter les polygones dans leur format vecteur, sans passer par une étape d'échantillonnage, elle ne peut traiter que les polygones élémentaires, d'une part, et elle présente une instabilité algorithmique dès qu'on l'utilise pour calculer les moments d'ordre élevé, d'autre part.

## ANNEXE B : INFLUENCE DU CHOIX DU PAS D'ECHANTILLONNAGE -EXEMPLES-





## ANNEXE C : PROTOTYPE D'APPARIEMENT

**Interface graphique du prototype d'appariement surfacique**

*En entrée, le prototype demande les deux jeux de données à appairier (Dataset#1 et Dataset#2).*

*Une valeur de seuil est également demandée pour supprimer les liens dus aux intersections parasites. Cette valeur peut être introduite manuellement en se basant sur le document des spécifications (s'il existe), ou bien elle sera détectée automatiquement comme la moitié de la surface de plus petit polygone entre les deux jeux de données à appairier.*

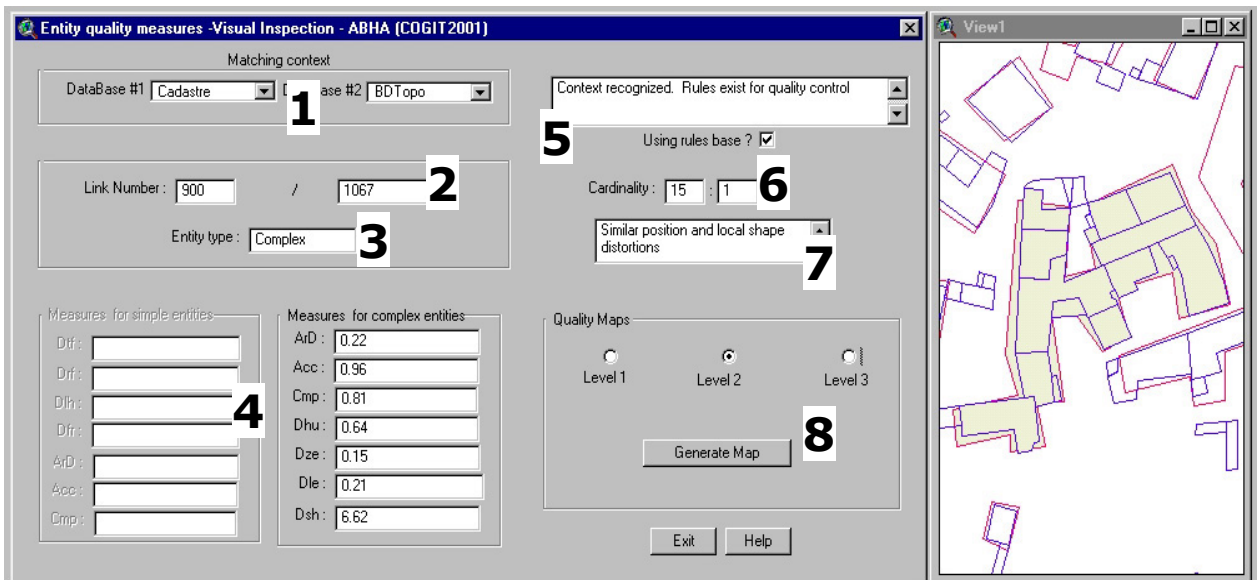
*La glissière sur l'interface graphique permet à l'utilisateur de définir une valeur de seuil de coupure sur les valeurs de la fonction d'inclusion. Les valeurs sur la glissière varient de 0 à 0.5 avec un pas de 0.05.*

*La première case à cocher en bas de l'interface donne la possibilité à l'utilisateur d'engager un ensemble de mesures de forme et de position sur les entités appariées.*

*La deuxième case à cocher en bas de l'interface permet à l'utilisateur de générer un fichier d'historique du processus de l'appariement dans lequel seront consignées toutes les informations utiles sur les jeux de données et les valeurs des filtres utilisés.*

*La partie droite de l'interface, "Context" et "Matching tables", permet de définir la structure de stockage des résultats de l'appariement.*

## ANNEXE D : MODULE DE CONTROLE DE LA QUALITE PAR L'UTILISATION DES REGLES DE CLASSIFICATION



**Module de contrôle visuel des entités appariées**

*En introduisant les deux bases de données appariées (1), le module peut ainsi déterminer le contexte de l'appariement en cherchant dans la base des contextes si le contexte en cours a été ultérieurement étudié. Si le contexte est reconnu, l'information sera affichée en (5) et le module attaquera la base des règles relatives à ce contexte.*

*Le module propose également à l'utilisateur de visualiser les liens d'appariement établis un par un. En effet, l'utilisateur introduit le numéro du lien (2), et c'est ainsi que le module détecte la nature du lien (simple ou complexe (3)) et affiche ses cardinalités s'il est complexe (6) et renvoie les mesures des écarts entre les entités appariées (4).*

*Dans le cas où le contexte de l'appariement serait étudié, les règles de classification ont été générées pour identifier les différentes classes d'écarts géométriques. Le module applique les règles sur les mesures pour décrire la nature de la configuration des écarts (7).*

*Des cartes de qualité peuvent être générées (8) en utilisant les résultats de la classification pour présenter l'information de la qualité sous une forme détaillée ou agrégée selon l'utilisateur ou selon l'application visée utilisant les données analysées.*

## **REFERENCES BIBLIOGRAPHIQUES**

- [**Abbas 1994**] Abbas I. *Bases de données vectorielles et erreur cartographique. Problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorff*. Thèse de doctorat, Université de Paris VII, 1994.
- [**Abo Zaid & Horne 1992**] Abo Zaid A.M. et Horne E. *Generalized complex moment descriptors*. In proceeding of the 34<sup>th</sup> Midwest Symposium on Circuits and Systems, Vol. 1, 151-154, 1992.
- [**Adoram & Lew 1999**] Adoram M. et Lew M.S. *IRUS: Image retrieval using shape*. IEEE International Conference on Multimedia Computing and Systems, Vol. 2, 597-602, 1999.
- [**Affholder 1993**] Affholder J.G. *Road modelling for generalization*. NCGIA research initiative 8, Specialist meeting on Formalizing Cartographic Knowledge, Buffalo, Etats unis, 23-36, 1993.
- [**Alesheikh & al. 1999**] Alesheikh A.A., Chapman M.A., Blais J.A.R et Karimi H. *Uncertainty models of GIS objects*. In proceeding of the International Symposium on Spatial Data Quality, 1999, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 308-315, 1999.
- [**Alt & al. 1993**] Alt H., Behrends B., Blömer J. *Approximate matching of polygonal shapes*. Technical report N. B93-10, Serie B – Informatik, Freie Universität Berlin, 1993.
- [**Amrhein & Griffith 1991**] Amrhein C.G et Griffith D.A. *A model for statistical quality control of spatial data in a GIS*. In proceeding of the Canadian Conference on GIS, 91-103, 1991.
- [**Andrews & al. 1994**] Andrews D.S., Snoeyink J., Boritz J., Chan M., Denham G., Harrison J. et Zhu C. *Further comparison of algorithms for geometric intersection problems*. In proceeding of SDH'94, Vol. 2, Waugh T.C. and Healy R.G. (Eds), Edinburgh, Scotland, UK, 709-724, 1994.
- [**Angwin 1991**] Angwin G.T. *Quality control of digital map generation*. In proceeding of GIS/LIS'94, Vol.2, 715-723, 1991.
- [**Anselin & Getis 1992**] Anselin, L. et Getis A. *Spatial statistical analysis and geographic information system*. The Annals of Regional Science, 26, 19-33, 1992.
- [**Arkin & al. 1991**] Arkin E.M., Chew L.P., Huttenlocher D.P., Kedem, K., Kedem K. et Mitchel J.S.B. *An efficient computable metric for comapring polygonal shapes*. IEEE Trabsactions on Pattern Analysis ans Machine Intelligence, Vol. 13, N. 3, 209-216, 1991.
- [**Aspinall 1996**] Aspinall R.J. *Measurement of area in GIS : a rapid method for assessing accuracy of area measurement*. In proceeding of the GIS Research UK Conference, Canterbury, Kent UK, 135-142, 1996.

- [**Badard 1998**] Badard Th. *Extraction des mises à jour dans les BDG – De l'utilisation de méthodes d'appariement*. Revue Internationale de Géomatique, Vol. 8, N. 1-2, Hermès (Ed), Paris, 121-147, 1998.
- [**Badard 2000**] Badard Th. *Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques*. Thèse de doctorat, Université de Marne-la-vallée, 2000.
- [**Beard & Chrisman 1988**] Beard K. et Chrisman N.R. *Zipper: a localized approach to edgematching*. American Cartographer 15(2), 163-172, 1988.
- [**Bédard 1987**] Bédard Y. *Uncertainties in land information systems databases*. In the proceeding of AutoCarto 8, 175-184, 1987.
- [**Bédard & Vallière 1995**] Bédard Y. et Vallière D. *Qualité des données à référence spatiale dans un contexte gouvernemental*. Université Laval, Québec, 1995.
- [**Behzad & David 1985**] Behzad S. et David J.A. *Uniform resampling of digitized contours*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, N. 6, pp. 674-681, 1985.
- [**Bel Hadj Ali 1997**] Bel Hadj Ali A. *Appariement géométrique des objets géographiques et étude des indicateurs de qualité*. Rapport de DEA, Université de Marne-La-Vallée, 1997.
- [**Bel Hadj Ali & Vauglin 1999**] Bel Hadj Ali A. and Vauglin F. *Geometric matching of polygons in GISs and assesment of geomerical quality of polygons*. In proceeding of the International Symposium on Spatial Data Quality, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 33-43, 1999.
- [**Bel Hadj Ali 2000**] Bel Hadj Ali A. *Mesures entre entités surfaciques – Application à la qualification des liens d'appariement -*. La recherche à l'IGN, Bulletin d'information N. 71, IGN (Eds), 33-54, 2000.
- [**Bel Hadj Ali & Vauglin 2000**] Bel Hadj Ali A. & Vauglin F. *Assessing positional and shape accuracy of polygons in vector GIS*. In proceeeding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 9-12, 2000.
- [**Bel Hadj Ali 2001a**] Bel Hadj Ali A. *Using moments for representing polygons and assessing their shape quality in GIS*. Article submitted to the Journal of Geographical Systems (Accepted), 2001.
- [**Bel Hadj Ali 2001b**] Bel Hadj Ali A. *Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification*. In proceeding of ECSQARU'2001 Workshop on Spatio-Temporal

Reasoning and Geographic Information Systems, Jeansoulin R. et Papini O. (Eds), Toulouse, 2001.

- [**Belkasim & al. 1990**] Belkasim S.O., Shridhar M. et Ahmadi M. *Shape contour recognition using moment invariants*. In proceeding of the 10<sup>th</sup> International Conference on Pattern Recognition, Vol. 1, 649-651, 1990.
- [**Belkasim & al. 1996**] Belkasim S.O., Shridhar M. et Ahmadi M. *Efficient algorithm for fast computation of zernike moments*. In proceeding of the 39<sup>th</sup> Midwest Symposium on Circuits and Systems, Vol. 3, 1401-1404, 1996.
- [**BI 1997**] Bulletin d'information de l'Institut Géographique National. *Qualité d'une base de données géographique: concepts et terminologie*. IGN (Ed.), Vol. 67, 1997.
- [**Blakemore 1983**] Blakemore M. *Generalization and error in spatial data bases*. In Cartographica, Vol. 21 (2/3), 131-139, 1983.
- [**Bolstad & al. 1990**] Bolstad P.V, Gessler P. et Lillesand T.M. *Positional uncertainty in manually digitized map data*. International Journal of Geographical Information Systems, 4(4), 399-412, 1990.
- [**Bonin 2000**] Bonin O. *New Advances in error simulation in vector geographical databases*. In proceeeding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 59-65, 2000.
- [**Borowski & Borwein 1989**] Borowski E.J. et Borwein J.M. *Dictionnaire des mathématiques*. Collins (Ed.), ISBN 06006434347-6, 1989.
- [**Brown & Baran 1995**] Brown J. et Baran J. *Are you conflated? Integrating TIGER and other data sets through automated network conflation*. In GIS-T 95, Cambridge: Gis/Trans LTD, 1995.
- [**Buttenfield 1991**] Buttenfield B. *A rule for describing line feature geometry*. *Map generalization*: Buttenfield B. & McMaster R. (Eds), Harlow Essex England: Longman scientific, Chap. 3, 150-171, 1991.
- [**CEN 1999**] CEN/TC 287. Env 12657: "*Geographic information –Data description-Metadata*", Rev. 165, CEN/TC287, 1999.
- [**Chamussy & al. 1994**] Chamussy H., Charre J., Dumolard P. et Durand M.G. *Initiation aux pratiques statistiques en géographie*. Groupe Chadule, Masson (Ed.), 1994.
- [**Chassery & Montauvert 1991**] Chassery J.M. et Montauvert A., 1991. *Géométrie discrète en analyse d'images*. Hermès (Ed.), 1991.
- [**Chavent 1992**] Chavent M. *Analyse de données symboliques, une méthode divisive de classification*. Thèse de Doctorat, Université de Paris-9 Dauphine, 1992.

- [**Cheng & Molenaar 1999**] Cheng T. et Molenaar M. *Syntactic representation of three fuzzy object models*. In proceeding of the International Symposium on Spatial Data Quality, 1999, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 506-516, 1999.
- [**Cheung & Eisenstein 1978**] Cheung R. et Eisenstein B. *Feature selection via dynamic programming for text-independant speaker identification*. In IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. ASSP-26, N. 5, 397-403, 1978.
- [**Cheung & Ip 1998**] Cheung K.K.T et IP H.H.S. *Image retrieval in digital library based on symmetry detection*. In proceeding of the International Symposium on computer Graphics, 366-372, 1998.
- [**Chen 1993**] Chen C.C. *Improved moment invariants for shape descimination*. Pattern Recognition, Vol. 26, No. 5, 683-686, 1993.
- [**Chesnokov 1982**] Chesnokov S.V. *Determinacy analysis of socio-economic data*. Nauka (Ed.), Moscow, 1982.
- [**Chrisman 1982**] Chrisman N.R. *A theory of cartographic error and its measurment in digital bases*. In proceeding of AutoCarto (Utrecht: EGIS Foundation), 5, 159-168, 1982.
- [**Chrisman 1989**] Chrisman N.R. *Modeling error in overlaid categorical maps*. In Accuracy of Spatial data bases, Goodchild M. and Gopal S. (Eds), 21-34, 1989.
- [**Chrisman 1983**] Chrisman N.R. *The role of quality in the long term fonctionning of geographical information system*. In AutoCarto 6, Ottawa, Ontario CA, 1983.
- [**Chrisman & Lester 1991**] Chrisman N.R. et Lester M. *A diagnostic test for error in categorical maps*. In AutoCarto 10, Vol. 6: Technical papers, Baltimore, 1991.
- [**Chrisman & al. 1992**] Chrisman N.R., Dougenik J.A et White D. *Lessons for the design of polygon overlay processing from the odessey Whirlpool algorithm*. In the proceeding of SDH'92, Vol. 2, 401-410, 1992.
- [**Clementini & Di Felice 1997**] Clementini E. et Di Felice P. *A global framework for qualitative shape description*. Geoinformatica, 1, 11-27, 1997.
- [**CNIG 1993**] Conseil National de l'Information Géographique, Groupe de travail. *Qualité des données géographiques échangées*. 1993.
- [**Cohen & Guibas 1997**] Cohen S.D. et Guibas L.J. *Partial matching of planar polylines under similarity transformations*. In proceeding of 8th Annual ACM-SIAM Symposium on Discrete Algorithms, 777-786, 1997.
- [**Coster & Chermant 1989**] Coster M. et Chermant J.L. *Précis d'analyse d'images*. Presses du CNRS, Chap. 9, 291-339, 1989.

- [**Couget 1997**] Couget P. *Etude d'un outil de bruitage de la qualité des données géographiques*. Mémoire de DESS AIST, Vauglin F. (Ed), 1997.
- [**Dai & Jing 1999**] Dai X.L. et Jing L. *An object-based approach to automated image matching*. In the proceeding of the IEEE international symposium on Geoscience and Remote Sensing, Vol. 2, 1189-1191, 1999.
- [**De Groeve & Lowell 1998**] De Groeve, T. et K.E. Lowell. *Super ground truth as a foundation for a model to represent and handle spatial uncertainty*. In proceedings of 3rd International Symposium on Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources, Lowell K. and Jatton A. (Eds), ISBN 1-57504-119-7, Quebec, 189-193, 1998.
- [**De Jong 1990**] De Jong S.M. *Integration of remotely sensed and GIS data to determine SPOT classification accuracy*. In proceeding of EGIS'90, Vol. 1, Amsterdam NL, 517-525, 1990.
- [**Delattre 2000**] Delattre S. *Modélisation des primitives surfaciques: approches par moments – Application à la qualification des liens d'appariement*. Mémoire de DESS Mathématiques et traitement de signal, Université du Littoral côte d'opale, Laboratoire COGIT, Bel Hadj Ali & Vauglin (Eds.), 2000.
- [**Devogele 1997**] Devogele Th. *Processus d'intégration et d'appariement de bases de données géographiques – Application à une base de données routières multi-échelles*. Thèse de doctorat, méthodes informatiques, Université de Versailles, 1997.
- [**Devogele 2000**] Devogele Th. *Using Distances for linear accuracy measurements*. In proceeding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 175-178, 2000.
- [**Dimitrijevíc 2000**] Dimitrijevíc I. *Etude des liens d'appariement topologique et géométrique dans les bases de données géographiques*. Rapport de DESS Intelligence artificielle, Université Pierre et Marie Curie, Paris VI, Vauglin F. (Ed.), 2000.
- [**Edwards 1994**] Edwards G. *Characterising and maintaining polygons with fuzzy boundaries in geographic information systems*. Sixth International Symposium on Spatial Data Handling, Vol. 1, Edingburgh, 223-239, 1994.
- [**Edwards & Lowell 1996**] Edwards G. et Lowell K.M. *Modelling uncertainty in photointerpreted boundaries*. Photogrammetric Engineering and Remote Sensing, Vol. 62, N. 4, 377-391, 1996.
- [**Edwards 1997**] Edwards G. *Reasoning about shape using the tangential axis transform (TAT) or the shape "grain"*. In proceeding of AAI'97, Rhode Island, 27-28, 1997.



- [Egenhofer & Herring 1990] Egenhofer, M. J. et Herring, J. R. *A mathematical framework for the definition of topological relationships*: Proceedings of the 4th International Symposium on Spatial Data Handling, 803-813, 1990.
- [Egenhofer & Franzosa 1991] Egenhofer, M. J. et Franzosa, R. D. *Point-set topological spatial relations*: International Journal of Geographic Information Systems, Vol. 5, 161-174, 1991.
- [Egenhofer & Mark 1995] Egenhofer M.J. et Mark D.M. Naive geography. In Frank A.U and Kuhn W. (Eds), *Spatial information theory: a theoretical basis of GIS – International conference COSIT'95*, Berlin, 1-15, 1995.
- [Ehrlichlaeger 1996] Ehrlichlaeger C.R. *Modelling elevation uncertainty in geographical analysis*. In proceeding of the International Symposium on Spatial Data Handling, Delft NL, 9B.1-9B.25, 1996.
- [Eiter & Mannila 1994] Eiter T. et Mannila H. *Computing discrete Fréchet distance*. Technical report N. CD6TR94/96, Technische Universität Wien, 1994.
- [Ezer & al. 1994] Ezer N., Anarim E. et Sankur B. *A comparative study of moment invariants and fourier descriptors in planar shape recognition*. In Proceeding of the 7<sup>th</sup> Mediterranean Electronical Conference, Vol. 1, 242-245, 1994.
- [Faïz & Boursier 1994]. Faïz O.S. et Boursier P. *GeoQual: a data model for handling the quality of geographic information*. In proceeding of 2<sup>nd</sup> ACM Workshop on advances in geographic information systems, Gaithersburg, USA, 1994.
- [Faïz 1996] Faïz, O. S. *Modélisation, exploitation et visualisation de l'information qualité dans les bases de données géographiques*. Thèse de doctorat, Université de Paris-Sud, UFR scientifique d'Orsay, 1996.
- [Flusser & Suk 1994] Flusser J. et Suk T. *A moment-based approach to registration of images with affine geometric distortion*. In IEEE Transactions on Geoscience and Remote Sensing, Vol. 32, N. 2, 382-387, 1994.
- [Fritsch 1997] Fritsch E. *Représentation de la géométrie et des contraintes cartographiques pour la généralisation du linéaire routier*. Thèse de doctorat, Science de l'Information Géographique, Université de Marne-La-Vallée, 1997.
- [Fonte & Lodwick 2000] Fonte C.C. et Lodwick W.A. *Procedures for incorporating uncertainty information in a geographical information system*. In proceeding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 213-216, 2000.
- [Fouqué 1999] Fouqué L. *Simulation d'erreurs dans une base de données géographiques*. Mémoire de DESS de mathématiques appliquées, statistiques et modèles stochastiques, Université de Rennes I, 2000.

- [**Gabay & Doytsher 1994**] Gabay Y. et Doytsher Y. *Automatic adjustment of line maps*. In the proceeding of GIS/LIS'94, Phoenix-Arizona, 333-341, 1994.
- [**Gabay & Doytsher 1995**] Gabay Y. et Doytsher Y. *Automatic feature correction in mergin line maps*. In proceeding of the 1995 ACSM/ASPRS Annual Convention & Exposition Technical Papers (Betesda: American Congress on surveying and Mapping/ American Society for photogrammetry and Remote Sensing), Vol. 1, 404-410, 1995.
- [**Gillmann 1985**] Gillmann D. *Trinagulations for rubber-sheeting*. In Proceeding of Auto-Carto 7 (Betesda: American Congress on surveying and Mapping/ American Society for photogrammetry and Remote Sensing), 191-199, 1985.
- [**Greenberg & al 1996**] Greenberg S., Guterman H. et Rotman S.R. *Rotation-invariant MLP classifiers for automatic aerial image recognition*. In proceeding of Electrical and Electronics Engineers in Israel, 8<sup>th</sup> convention, 2.2.4/1-2.2.4/5, 1996.
- [**Griffith 1989**] Griffith D.A. *Distance calculations and errors in geographic databases*. In Accuracy of spatial data bases, Goodchild M.F. and Gopal S. (Eds), 81-90, 1989.
- [**Goodchild 1999**] Goodchild M.F. *Measurement-based GIS*. In proceeding of the international Symposium on Spatial Data Quality, 1999, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 1-9, 1999.
- [**Goodchild & Cova 1995**] Goodchild M.F. et Cova T.J. *Mean objects: extending the concept of central tendency to complex spatial objects in GIS*. In proceeding GIS/LIS'95, ASPRS/ACSM, Nashville, TN, 354-364, 1995.
- [**Goodchild 1991**] Goodchild M.F. *Issue of the quality and uncertainty*, Advances in cartography, Müller Ed., Barking, Essex-Elsevier, 113-139, 1991.
- [**Goodchild & Hunter 1997**] Goodchild M.F et Hunter G. *A simple positional accuracy measure for linear features*. Technical communication, International Journal of Geographical Information Systems, 11(3), 299-306, 1997.
- [**Hai & Hai 1996**] Hai L. et Hai-Zou L. *Chinese signature verification with moment invariants*. In proceeding of IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, 2963-2968, 1996.
- [**Haining 1990**] Haining R.P. *Spatial data analysis in the social and environmental sciences*. Cambridge: Cambridge University Press, 1990.
- [**Hangouët 1998**] Hangouët J-F. *Approche et méthodes pour l'automatisation de la généralisation cartographique; application en bord de ville*. Thèse de doctorat, Science de l'Information Géographique, Université de Marne-La-Vallée, 1998.

- [**Harvey 1994**] Harvey F. *Defining unmovable nodes/segments as part of vector overlay: the alignment overlay*. In proceeding of SDH'94, Vol. 1, Waugh T.C. and Healy R.G. (Eds), Edinburgh, Scotland UK, 159-176, 1994.
- [**Harvey & Vauglin 1997**] Harvey F. et Vauglin F. *No fuzzy creep! A clustering algorithm for controlling arbitrary node movement*. In proceeding of Auto-Carto 13, technical paper, ACSM ASPRS, Annual convention, Vol. 5, Seattle, 317-326, 1997.
- [**Harvey & al. 1998**] Harvey F., Vauglin F. et Bel Hadj Ali A. *Geometric matching of areas: comparison measures and association links*. In proceeding of the 8<sup>th</sup> International Symposium on Spatial Data Handling, SDH'98, Vancouver CA, Poiker T.K. and Chrisman N. (Eds.), 557-568, 1998.
- [**Hausdorff 1937**] Hausdorff F. *Set theory*. Chelsea, New York, 1957, traduction de la troisième édition de Mengenlehre, 1937.
- [**Hergoz 1989**] Hergoz A. *Modeling reliability on statistical surfaces by polygon filtering*. In Accuracy of Spatial data bases, Goodchild M.F. and Gopal S. (Eds), 209-218, 1989.
- [**Hottier 1997**] Hottier Ph. *Notes de cours sur le contrôle de localisation par la comparaison à une référence*. DEA Science de l'Information Géographique, Ecole Nationale des Sciences Géographiques, 1997.
- [**Hu 1962**] Hu M.K. *Visual problem recognition by moment invariant*, IRE Trans. Inform. Theory, Vol. IT-8, 179-187, 1962.
- [**Hunter & Goodchild 1993**] Hunter G.J. et Goodchild M.F. *Managing uncertainty in spatial databases: Putting theory into practice*. In the journal of the urban and regional Information systems association, Vol. 5, N. 2, 55-62, 1993.
- [**Hunter & al. 1994**] Hunter G.J., Goodchild M.F. et Robey M. *A toolbox for assessing uncertainty in spatial databases*. In the proceeding of the Annual conference of the Australasian urban and regional information systems association, AURISA'94, Sydney, 1994.
- [**Hunter & Goodchild 1995**] Hunter G.J. et Goodchild M.F. *Dealing with error in spatial databases: A simple case study*. Photogrammetric Engineering & Remote Sensing, Vol. 61, N. 5, 529-537, 1995.
- [**IGN 1997**] Institut Géographique National. *Spécifications détaillées de la Base de données Topographique*. Version 2, Rev. 4, IGN/SIT/Equipe produit BDTopo, Réf. SIT/97/0155, IGN, Paris, 1997.
- [**IGN 1999**] Institut Géographique National. *Spécifications de contenu BDCarto*. Version 3.1, Edition 4, IGN/SIT/Produit BDTopo, Réf. SDT/99/-232, IGN, Paris, 1999.

- [ISO 1994] International Organization for Standardization, Norme internationale – *Management de la qualité et assurance de la qualité- vocabulaire*. Numéro de référence ISO8402:1994(E/F/R), 2<sup>ème</sup> édition, 1994.
- [Jen & Boursier 1994] Jen T.J et Bouriser P. *A model for handling topological relationships in 2D environment*. In the proceeding of the 6<sup>th</sup> International Symposium on Spatial Data Handling, 73-88, 1994.
- [Johnson & Wichern 1998] Johnson R.W. et Wichern D.W. *Applied multivariate statistical analysis*. Prentice hall (Ed.), Upper Saddle River, New Jersey, 1998.
- [Jones & al. 2000] Jones C.B, Ware M. et Miller D.R. *Bayesian probabilistic methods for change detection with area-class maps*. In proceeeding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 329-336, 2000.
- [Joos 1994] Joos G. *Quality aspects of geo-informations*. In proceeding of EGIS'94, EGIS foundation, 1147-1153, 1994.
- [Juran & al. 1974] Juran J.M, Gryna F.M.J et Bingham R.S. *Quality control handbook*. McGraw-Hill, New-York, 1974.
- [Kaufman & Rousseeuw 1990] Kaufman L. et Rousseeuw P.J. *Finding groups in data : an introduction to cluster analysis*. Wiley (Ed.), New York, 1990.
- [Kauppinen & al. 1995] Kauppinen H., Seppanen T. et Pietikainen M. *An experimental comparaison of autoregressive Fourier-based descriptors in 2D shapes classification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, N. 2, 201-207, 1995.
- [Khagendra & John 1992] Khagendra, T., et John, B. *Accuracy of spatial data used in geographic information systems*. Photogrammetric Engineering and Remote Sensing, 58(6), Review article, 835-841, 1992.
- [Khotanzad & Hong 1990] Khotanzad A. et Hong Y.H. *Invariant image recognition by zernike moments*. IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 12, N. 5, 489-497, 1990.
- [Kidner 1996] Kidner B. D. *Geometric signatures for determining polygon equivalence during multi-scale GIS update*. Second Join European Conference & Exhibition on Geographical Information, Barcelona, Spain, IOS Press, 238-247, 1996.
- [Kiiveri 1997] Kiiveri H.T. *Assessing, representing and transmitting positional uncertainty in maps*. International Journal of Geographical Information Science, Vol. 11, N. 1, 33-52, 1997.

- [**Kim & Kim 1999**] Kim W.Y. et Kim Y.S. *Robust rotation angle estimator*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, N. 8, 768-773, 1999.
- [**Kohonen 1986**] Kohonen T. *Learning vector quantization for pattern recognition*. Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland, 1986.
- [**Lanter & Veregin 1992**] Lanter D.P. et Veregin H. *A research paradigm for propagating error in layer-based GIS*. Photogrammetric Engineering and Remote Sensing journal, Vol. 58, N. 6, 825-833, 1992.
- [**Laurini & Milleret-Raffort 1993**] Laurini, R., et Milleret-Raffort, F.. *Les bases de données en géomatique*. Hermes (Ed.), Paris, 1993.
- [**Laurini 1998**] Laurini R. *Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability*. International Journal of Geographical Information Science, 12, 373-402, 1998.
- [**Lemarié 1996**] Lemarié C. *Etat de l'art sur l'appariement*. Rapport technique DT/960002/S-RAP, IGN, Service de la Recherche, Saint Mandé, France, 1996.
- [**Le Men & Jamet 1990**] Le Men H. et Jamet O. *Interprétation automatique de l'occupation du sol sur image SPOT*, Symposium International de Cartographie Thématique Dérivée des Images Satellitaires, Saint Mandé, 1990.
- [**Le Men & Jamet 1993**] Le Men H. et Jamet O. *Qualité de processus d'interprétation et qualité des résultats: Un exemple en cartographie d'occupation du sol*. In proceeding of ACT'93, Tunis, 1993.
- [**Le Men 1994**] Le Men H. *Géométrie de l'occupation du sol dans la BDCarto®*. Note interne IGN/DT-940392, 1994.
- [**Leung & Yan 1998**] Leung, Y. et Yan J. *A locational error model for spatial features*. In International Journal of Geographical Information Science, Vol. 12, N. 6, 607-620, 1998.
- [**Leung & Yan 1997**] Leung Y. et Yan J. *Point-in-polygon analysis under certainty and uncertainty*. Geoinformatica, 1, 93-114, 1997.
- [**Liao & Pawlak 1996**] Liao S.X. et Pawlak M. *On image analysis by moments*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, N. 3, 254-266, 1996.
- [**Mark & al. 1999**] Mark D.M., Smith B. et Tversky B. *Ontology and geographic objects: An empirical study of cognitive categorization*. In proceeding of International Conference COSIT'99, Spatial Information Theory, Freksa C. (Ed.), Springer lecture note in computer Science, 283-298, 1999.

- [**Matheron 1963**] Matheron G. *Traité de Géostatistique Appliquée*. Tome I, Mémoires du Bureau de Recherches Géologiques et Minières, No 14, Edition Technip., Paris, (1962)
- [**McAlpine & Cook 1971**] McAlpine J.R. et Cook B.G. *Data reliability from map overlay*. In proceeding of Australian and New Zealand Association for the Advancement of Science, 43<sup>rd</sup> Congress, Section 21-Geographical science, 1971.
- [**Milenkovic 1989**] Milenkovic V. *Verifiable implementations of geometric algorithms using finite precision in geometrical reasoning*. Kapur D. and Mundy J. (Eds), Cambridge, 1989.
- [**Milgram 1993**] Milgram M. *Reconnaissance des formes: méthodes numériques et connexionistes*. Collection A2I, Armand Collin (Ed), ISBN : 2-200-21290-9, 1993.
- [**Mirkin 1987**] Mirkin B. *method of principal cluster analysis*. Automation and Remote control, N. 48, 1379-1388, 1987.
- [**Molenaar 1998**] Molenaar M. *An introduction to the theory of spatial object modeling*. Taylor & Francis (Eds), London UK, 1998.
- [**Mokhtarian & Mackworth 1992**] Mokhtarian F. et Mackworth A.K. *A theory of multiscale, curvature-based shape representation for planar curves*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, N. 8, 789-805, 1992.
- [**Mukundan & Ramakrishnan 1998**] Mukundan R. et Ramakrishnan K.R. *Moment functions in image analysis : theory and applications*. World Scientific (Ed.). ISBN 981-02-3524-0, 1998.
- [**Mustière 2001**] Mustière S. *Apprentissage supervisé pour la généralisation cartographique*. Thèse de doctorat, Université de Paris VI, 2001.
- [**Nassery & Faez 1996**] Nassery P. et Faez K. *Signature pattern recognition using pseudo zernike moments and a fuzzy logic classifier*. In Proceeding of International Conference on Image Processing, Vol. 2, 197-200, 1996.
- [**NIST 1992**] National Institute of Standards and Technology. *Spatial Data Transfer Standard (SDTS)*, FIPS PUB 173, 1992.
- [**Norheim 1998**] Norheim A.R. *Distinguishing positional and attribute error in two old growth forest mapping projects*. In proceeding of the 8th International Symposium on Spatial Data Handling, Poiker Th.K. & Chrisman N. (Eds.), Vancouver CA, 161-171, 1998.
- [**OGC 1999**] Open GIS Consortium. The OpenGIS<sup>TM</sup> Abstract Specification, Topic 9: *Quality*, Version 4, 1999.

- [**Ohta & Kanade 1985**] Ohta Y. et Kanade T. *Stereo by intra and inter scanline search using dynamic programming*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7(2), 139-154, 1985.
- [**Okabe & al. 1992**] Okabe A., Boots B. et Sugihara K. *Spatial tessellations, concepts and applications of voronoi diagrams*. Wiley & sons (Ed.), 1992.
- [**Parademetriou 1992**] Parademetriou R.C. *Reconstructing with moments*. In proceeding of 11<sup>th</sup> IAPR International Conference on Pattern Recognition, Vol. III, Conference C: Image, Speech and Signal Analysis, 476-480, 1992.
- [**Pejnovic & al. 1992**] Pejnovic P., Butorovic L. et Stojiljkovic Z. *Object recognition by invariants*. In proceeding of the 11<sup>th</sup> International conference on Pattern Recognition, Vol. II, Conference B: Pattern Recognition Methodology and Systems, 434-437, 1992.
- [**Perantonis & Lisboa 1992**] Perantonis S.J. et Lisboa P.J.G. *Translation, rotation and scale invariant pattern recognition by high-order neural networks and moment classifiers*. IEEE Transactions on Neural Networks, Vol. 3, Issue 2, 241-251, 1992.
- [**Perkel 1956**] Perkel, J. *On epsilon length*. Bulletin de l'académie polonaise des sciences, N. 4, 399-403, 1956.
- [**Phalakarn 1991**] Phalakarn B. *Evaluation de la qualité des processus de segmentation d'image par mise en correspondance à une référence*. Thèse de doctorat, Science de l'information géographique, Université Paris 7, 1991.
- [**Plazanet 1996**] Plazanet C. *Enrichissement des bases de données géographiques: Analyse de la géométrie des objets linéaires pour la généralisation cartographique (Application aux routes)*. Thèse de doctorat, Science de l'information géographique, Université de Marne-La-Vallée, 1996.
- [**Pullar 1991**] Pullar D. *Spatial overlay with inexact numerical data*. Auto-Carto 10, technical paper, ACSM ASPRS, Annual Convention, Vol. 6, 1991.
- [**Pullar 1993**] Pullar D. *Consequences of using a tolerance paradigm in spatial overlay*. In proceeding of Auto-Carto 11, 288-296, 1993.
- [**Quinlan 1993**] Quinlan J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann (Ed.), San Mateo, 1993.
- [**Reiners 1998**] Reiners T. *Maximum likelihood clustering of large data sets using a multilevel, parallel heuristic*. Diploma thesis, Institute of economics, department of business administration, business computer science, and information management, University of California at Davis, 1998.

- [Regnauld 1998] Regnauld N. *Généralisation du bâti : structure spatiale de type graphe et représentation cartographique*. Thèse de doctorat en informatique, Université de provence-Aix-Marseille 1, 1998.
- [Rosen & Saalfeld 1985] Rosen B. et Saalfeld A. *Match criteria for automatic alignment*. In Proceeding of Auto\_Carto 7 (Beteda: American Congress on surveying and Mapping/ American Society for photogrammetry and Remote Sensing), 1-20, 1985.
- [Rothe & al. 1996] Rothe I., Süsse H. et Voss K. *The method of normalization to determine invariants*. IEEE Transactions on Pattern Analysis and Machine Intellignce, Vol. 18, N. 4, 366-376, 1996.
- [Rui & al. 1998] Rui Y., She A., et Huang Th.S. *A modified fourier descriptor for shape matching in MARS*. Image Databases and Multimedia Search , Series on Software Engineering and Knowledge Engineering Vol 8, Chang S. K. (Ed.), World Scientific Publishing House in Singapore, 165-180, 1998.
- [Rushton 2000] Rushton, G. *Apportionning losses in locationnal efficiency among elements in a set of location decision*. In the proceeding (abstracts) of the first international conference on goeographic information system, Savannah-USA, 253, 2000.
- [Saalfeld 1988] Saalfeld A. *Automated map compilation*. International Journal of Geographical Information Systems, 2, 217-228, 1988.
- [Salmeron & Milgram 1986] Salmeron E. et Milgram M. *Utilisation de la relaxation pour la mise en correspondance des segments d'une carte et d'une image aérienne*. Semaine Internationale de l'Image Electronique, 2<sup>ème</sup> Colloque Image, Nice, 32-38, 1986.
- [Saporitti & al. 1993] Saporitti N., Daures J.F et Zumsteeg A. *Etudes menées sur la qualité des donnése et contrôle qualité des bases de données vecteur*. Journées Géographiques et Défense, CNIT Paris la défense, 80-110, 1993.
- [Servigne 1993] Servigne S. *Bases de données géographiques et photos aériennes: de l'appariement à la mise à jour*. Thèse de doctorat, INSA, Lyon, 1993.
- [Shen & Shen 1996] Shen J. et Shen D. *Image characterization by fast calculation of low-order Legendre moments*. IEEE, In proceeding of ICPR'96, 1144-1149, 1996.
- [Shi 1994] Shi W. *Modelling positional and thematic uncertainties in integration of remote sensing and geographic information systems*. PhD thesis, International Institute for Aerospace Survey and Earth Sciences (ITC), ISBN 90 6164 099 7, 1994.
- [Shi & Guo 1999] Shi W. et Guo W. *Modeling topological relationships of spatial objects with uncertainties*. In proceeding of the International Symposium on



Spatial Data Quality, 1999, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 487-495, 1999.

[**Smith 1996**] Smith B. *Mereotopology: A theory of parts and boundaries*. Data and knowledge engineering, Elsevier (Ed.), N. 20, 287-303, 1996.

[**Smith 1999**] Smith B. *Agglomerations*. In proceeding of International Conference COSIT'99, Spatial Information Theory, Freksa C. (Ed.), Springer lecture note in computer Science, 267-282, 1999.

[**Suan & al. 2000**] Suan P.K., Chu T.H. et Aileen A. *Area representation errors associated with rasterization*. In prooceding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 337-342, 2000.

[**Teague 1980**] Teague M. *Image Analysis via the General Theory of Moments*. Journal of Optical Soc. Americ., Vol. 70, N. 8, 920-930, 1980.

[**Teh & Chin 1988**] Teh C.H. et Chin R.T. *On image analysis by the moments*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, N. 4, 496-513, 1988.

[**Tello 1995**] Tello, R., 1995. *Fourier descriptors for computer graphics*. In IEEE Transactions on Systems, Man. And Cybernetics, Vol. 25, N. 5, 861-865, 1995.

[**Tong & al. 2000**] Tong X., Shi W. et Liu D. *Error distribution, error tests and processing of digitized data in GIS*. In prooceding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 642-646, 2000.

[**Tveite 1999**] Tveite H. *An accuracy assessment method for geographical line data sets based on buffering*. International Journal of Geographical Information Science, Vol. 13, N. 1, 27-47, 1999.

[**Vauglin 1997**] Vauglin F. *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires*. Thèse de doctorat, Science de l'information géographique, Université de Marne-La-Vallée, 1997.

[**Vauglin & Bel Hadj Ali 1998**] Vauglin F. et Bel Hadj Ali A. *Geometric matching of polygonal surfaces in GISs*. ASPRS-RTI Annual Conference, Tampa FL, March 30- April 3, USA, 1511-1516, 1998.

[**Veltkamp & Hagedoorn 1999**] Veltkamp R.C et Hagedoorn M. *State-of-the-art in shape matching*. Technical report, UU\_CS-1999-27, Utrech University, The Netherlands, 1999.

[**Venn 1881**] Venn J. *Symbolic logic*. MacMillan, London, première édition, 1881.

- [**Veregin 1994**] Veregin H. *Accuracy tests for polygonal features*. Technical report, Department of geography, Kent State University, Ohio, 1994.
- [**Wagner 1988**] Wagner D.F. *A method of evaluating polygon overlay algorithms*. In the technical papers of 1988 ACSM-ASPRS annual convention, Vol. 5, St Louis, 1988.
- [**Walter & Fritsch 1999**] Walter V. et Fritsch D. *Matching spatial data sets : a statistical approach*. International Journal of Geographical Information Science, Vol. 13, N. 5, 445-473, 1999.
- [**Ward 1963**] Ward J.H. *Hierarchical grouping to optimize an objective function*. Journal of American Statistical Association, N. 58, 236-244, 1963.
- [**Worboys 1998**] Worboys M. *Computation with imprecise geospatial data*. Computers, Environment and Urban Systems, 22, 85-106, 1998.
- [**Zhang & Tulip 1990**] Zhang G. et Tulip J. *An algorithm for the avoidance of sliver polygons and clusters of points in spatial overlay*. In the proceeding of the 4<sup>th</sup> International Symposium on Spatial Data Handling, Zurich, 141-150, 1990.
- [**Zaslavsky 1995**] Zaslavsky I. *Analysis of association between categorical maps in multi-layer GIS*. In proceeding of GIS/LIS'95, Nashville, TN, Vol. 2, 1066-1074, 1995.
- [**Zhou & Lee 1994**] Zhou F. et Lee Y.C. *Polygon uncertainty modeling and representation*. In proceeding of Canadian Conference on GIS, Ottawa, Vol. 1, 185-191, 1994.

## PUBLICATIONS

- [Harvey & al. 1998] Harvey F., Vauglin F. et Bel Hadj Ali A. *Geometric matching of areas: comparison measures and association links*. In proceeding of the 8<sup>th</sup> International Symposium on Spatial Data Handling, SDH'98, Vancouver CA, Poiker T.K. and Chrisman N. (Eds.), 557-568, 1998.
- [Vauglin & Bel Hadj Ali 1998] Vauglin F. et Bel Hadj Ali A. *Geometric matching of polygonal surfaces in GISs*. ASPRS-RTI Annual Conference, Tampa FL, March 30- April 3, USA, 1511-1516, 1998.
- [Bel Hadj Ali & Vauglin 1999] Bel Hadj Ali A. and Vauglin F. *Geometric matching of polygons in GISs and assesment of geomerical quality of polygons*. In proceeding of the International Symposium on Spatial Data Quality, Shi W., Goodchild M.F. and Fisher P. (Eds), Hong Kong Polytechnic University, 33-43, 1999.
- [Bel Hadj Ali 2000] Bel Hadj Ali A. *Mesures entre entités surfaciques – Application à la qualification des liens d'appariement -*. La recherche à l'IGN, Bulletin d'information N. 71, IGN (Eds), 33-54, 2000.
- [Bel Hadj Ali & Vauglin 2000] Bel Hadj Ali A. & Vauglin F. *Assessing positional and shape accuracy of polygons in vector GIS*. In prooceding of the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), Amsterdam NL, 9-12, 2000.
- [Bel Hadj Ali 2001a] Bel Hadj Ali A. *Using moments for representing polygons and assessing their shape quality in GIS*. Article submitted to the Journal of Geographical Systems (Accepted), 2001.
- [Bel Hadj Ali 2001b] Bel Hadj Ali A. *Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification*. In proceeding of ECSQARU'2001 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems, Jeansoulin R. et Papini O. (Eds), Toulouse, 2001.

### Atef BEL HADJ ALI – Curriculum Vitæ

Né le 16 Mai 1969 à Monastir (Tunisie). Ingénieur Cycle long en Télécommunications et Télémechanique, Ecole de l'Air, Sfax, Tunisie (Lauréat, avec attribution du prix de la présidence de la République Tunisienne 1992). Titulaire d'un D.E.S.S. Télédétection de l'université de Pierre et Marie Curie, Paris VI (1993) et Titulaire d'un D.E.A. Sciences de l'Information Géographique de l'Université de Marne-La-Vallée (1997). Ingénieur (Chef de projet) actuellement en poste au Centre National de Télédétection (Tunisie).

Marne la Vallée, Septembre 2001