Modèles et méthodes pour l'information spatio-temporelle évolutive

Christine Plumejeaud

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Christine PLUMEJEAUD

Thèse dirigée par **M. Jérôme Gensel**
et codirigée par **M. Claude Grasland**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
et de **Mathématiques, Sciences et Technologies de l'Information, Informatique**

Modèles et méthodes pour l'information spatio-temporelle évolutive

Thèse soutenue publiquement le **22 septembre 2011**,
devant le jury composé de :

Monsieur Jean-Pierre Giraudin
Professeur d'informatique, Université Pierre Mendès-France, Président
Madame Thérèse Rougé-Libourel
Professeur d'informatique, Université Montpellier II, Rapporteur
Monsieur Christophe Claramunt
Professeur d'informatique, Institut de Recherche de l'Ecole Navale, Rapporteur
Madame Anne Ruas
Directrice de Recherche en Géomatique, Institut Géographique National, Examinatrice
Madame Sophie de Ruffray
Professeur de géographie, Université de Rouen, Examinatrice
Monsieur Jérôme Gensel
Professeur d'informatique, Université Pierre Mendès-France, Directeur de thèse
Monsieur Claude Grasland
Professeur de géographie, Université Paris VII, Co-Directeur de thèse

« La pensée est le labeur de l'intelligence, la rêverie en est la volupté. »

(Victor Hugo)

Remerciements

Je souhaite exprimer ma plus grande reconnaissance aux membres de mon jury qui ont accepté d'évaluer mes travaux. Je remercie d'abord mon Président de jury, Jean-Pierre Giraudin, Professeur d'informatique et responsable de la filière informatique au CNAM de Grenoble, et qui, à ce titre, m'a suivi dès le début dans cette grande aventure. Je le remercie également pour ses judicieux conseils pédagogiques lorsqu'il était mon tuteur de monitorat.

Je remercie ensuite mes rapporteurs, les Professeurs d'informatique Thérèse Rougé-Libourel et Christophe Claramunt, dont les remarques ont permis d'améliorer la qualité de ce travail.

Merci aussi à Sophie de Ruffray, Professeur de Géographie à l'université de Rouen, et à Anne Ruas, directrice de recherche en Géomatique à l'Institut Géographique National, pour leur examen attentif de ce travail. Je remercie aussi Anne pour avoir cru en moi et m'avoir emmené dans une nouvelle aventure de recherche à l'IGN.

Je tiens à exprimer mon immense reconnaissance à mes directeurs de thèse, les Professeurs Jérôme Gensel et Claude Grasland. Je les remercie pour m'avoir placé face à un problème aussi passionnant que difficile, et pour la confiance qu'ils ont su porter en moi. Ces travaux ont bénéficié de leur complémentarité, tant sur le plan humain que scientifique. Ainsi, Jérôme, ta patience, ta gentillesse et ta résistance sont des vertus que je m'efforce de cultiver. Avant de réagir au quart de tour, je repense désormais à ce que tu ferais. Claude, ton talent pour jeter des ponts au dessus du vide, vers les autres et vers d'autres idées, ont su donner un souffle créateur à ma thèse. Je prends comme modèle ton ouverture et ta façon de cultiver des relations professionnelles enrichissantes.

En dehors de ce jury, mes premiers remerciements vont à Hélène Mathian qui, en plus d'être une collègue très compétente, est une personne sur qui j'ai toujours pu compter. Hélène, tu as su m'écouter, m'aiguiller de bon conseils et me redonner courage dans les moments difficiles : je te remercie pour ta grande humanité et ta bienveillance.

Cette recherche pluridisciplinaire m'a amené à travailler avec de « vrais » informaticiens (ceux qui compilent du code, n'est-ce pas Serge ? ;-)), comme avec des géographes, des statisticiens, voire même des archéologues, et c'est une chance exceptionnelle. Ainsi, je remercie ces professionnels de la géostatistique pour leur conseil et les bonnes bières partagées ensemble : Martin Charlton, Edzer Pebesma (dites « monsieur R »), et Tomislav Hengl. J'ai une pensée affectueuse pour les ingénieurs statisticiens et cartographes de RIATE, dits « les parisiens », dont la gaité a éclairé nos rencontres de travail. Cependant, il y a un géographe en particulier à qui vont mes pensées, c'est Guillaume Vergnaud, car non seulement notre collaboration a été des plus fructueuses, mais aussi nous avons partagé de très bons moments ensemble à Grenoble et Barcelone. Guillaume, j'espère que tu vas bien. Merci aussi aux informaticiens, ceux de chez Mescal en particulier, comme Jean-Marc Vincent ou Grégory Mounié, dont l'intérêt pour mes travaux me convainc encore plus que la Géomatique a un grand avenir devant elle, fait de collaborations, d'échanges et d'amitiés interdisciplinaires.

Après ce petit détour par les chemins interdisciplinaires, j'en reviens à mon équipe et mes collègues au quotidien.

– Je remercie les permanents de l'équipe Steamer. Merci à Marlène et Paule-Annick, pour leur coopération constante et leur générosité. Merci surtout, Marlène, pour ton regard sur les métadonnées et ta relecture approfondie de cette partie du travail, qui m'a donné le courage de terminer. Je remercie aussi les nouveaux venus, Danielle et Philippe, pour leur gentillesse, mais aussi le regard scientifique qu'ils ont su porter sur mes travaux, ne serait-ce qu'un bref instant.
– Merci à mes collègues d'équipe, mais surtout à José, pour LaTeX bien-sûr ;-), et aussi pour tous nos ennuis consolés, et nos joies partagées.
– Merci aux autres, les thésards (Raffaella, Betul, Mouna, Angela, ...), les post-doc (Sandro, Sidonie, et Mathieu), les ingénieurs (Anton, Laurent G., Laurent P., Benoit, Gael, Benjamin, Bruno, ...) et les stagiaires (Dounia, Marion, Rong-Rong, Socrates, ...).

Cette thèse n'aurait pas été la même sans Dounia ni Anton. Avec leur aide efficace, des prototypes ont vu le jour, des idées ont pu se concrétiser, et d'autres naître. Je les remercie aussi de leur grande honnêteté, et de leur amitié sincère. Merci aussi à Benoit pour avoir repris le flambeau sur HyperCarte, et m'avoir soulagée d'une tâche omniprésente, celle de faire vivre tous les Hyper* au quotidien. Enfin, Alban, Nadine et les moyens informatiques du LIG m'ont dépatouillé suffisamment souvent pour que je les remercie de leur aide, et de l'intérêt vif qu'ils ont porté à cette recherche.

Cette thèse s'est déroulée dans un grand laboratoire, plein de gens, très affairés, souvent pré-occupés, mais qui gardent quand même l'envie de partager des moments ensemble. Je remercie Charlotte et Sattisvar, dont le goût des autres ne s'est jamais démenti, pour la solidarité dont ils savent faire preuve. Grâce à eux, j'ai eu envie de participer au lancement de LIG Synergy, avec Rémi, Javier, Yves, Emeric et les autres. Notre barbecue reste un succès mémorable pour moi. L'aventure continue pour cette association, et j'espère qu'elle se poursuivra longtemps.

Enfin, en parlant d'association, mon rôle de secrétaire de l'AI CNAM PST, Association des Ingénieurs du CNAM et de la Promotion Supérieure du Travail a été très gratifiant, et la réunion du bureau de l'association était toujours un événement très agréable. Je remercie Eric Boniface, mon Président d'association, et André Plisson, directeur du CNAM de Grenoble, pour la grande confiance qu'ils ont bien voulu m'accorder. Je te dois un grand merci Eric pour une raison plus personnelle : tu m'as déniché un mari fantastique !

Je ne peux pas remercier tous les membres de ma famille comme je l'aurais aimé, mais je dis merci à ma tante Bernadette, qui a su prendre régulièrement des nouvelles du front et de mon cœur, et m'a offert des beaux moments d'air pur et de plaisir partagé. Je dois aussi un petit quelque chose à mes amis qui m'ont encouragée, comme Jean-Philippe et Laurence, pour que je continue mes progrès, et à Nathalie, qui n'a pas ménagé sa peine pour me voir finir cette thèse.

Merci enfin à Cyril, mon mari, qui pour l'instant n'a pas vraiment vécu les meilleurs jours avec moi (on sait bien que la rédaction d'une thèse n'est pas une sinécure pour l'entourage). Merci pour ton amour et ta patience.

Bientôt la lune de miel !

Table des matières

Table des figures

Liste des tableaux

Prolégomènes - Définition de l'objet d'étude

Cette thèse porte sur la définition de modèles et de méthodes pour la gestion de l'information spatio-temporelle évolutive. Nous nous intéressons à l'information statistique territoriale, et aux problèmes que la gestion d'une telle information peut poser.

Ce chapitre introduit le lecteur à l'ensemble des notions relatives à l'information statistique territoriale qui seront abordées dans cette thèse. Ce chapitre situe notre approche dans le cadre pluridisciplinaire de la *géomatique*, et précise quels seront les objectifs poursuivis dans ce travail.

A Constitution de l'information statistique territoriale

Ce chapitre apporte des définitions et des précisions au sujet des modalités de collecte et d'analyse des données qui sont au coeur de notre sujet d'étude : la statistique territoriale.

A.1 Statistique ou statistiques ?

Très fréquemment, les termes de « statistique » ou « statistiques » sont employés lorsque l'on parle de données socio-économiques. Au pluriel ou au singulier, ces termes n'ont pas le même sens. Selon la définition extraite d'un manuel sur la science statistique [Saporta 06], qui s'appuie sur l'*encyclopeadia universalis*, et indique que « Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste en leur recueil, leur traitement et leur interprétation. », on peut distinguer la science statistique des données (les statistiques) qu'elle traite.

Concernant la science statistique, et dans le cadre de la constitution d'un système d'information géographique dont la vocation serait l'aide à l'aménagement du territoire, la définition proposée par Jean-Claude Régnier [1] nous semble pertinente :

> « La statistique est la science qui procède à l'étude méthodique à partir de modélisations mathématiques, des modes d'utilisation et de traitement de données, c'est-à-dire de l'information, dans *le but de conduire et d'étayer une réflexion ou de prendre une décision* en situation concrète soumise aux aléas de l'incertain. »

1. http://jean-claude.regnier.pagesperso-orange.fr/joao_claudio/statisti/statistique_def.htm

Il faut rappeler qu'au sujet de la science statistique elle-même, en 1922, un des plus grands statisticiens de la première moitié du XXème siècle, Ronald A. Fisher, écrivit

> « L'objet de la méthode statistique est la réduction des données. Une masse de données doit être remplacée par un petit nombre de quantités représentant correctement cette masse, et contenant autant que possible la totalité de l'information pertinente contenue dans les donnée d'origine. Cet objectif est accompli par la construction d'une population infinie hypothétique. La statistique comporte des problèmes de spécification apparaissant à travers *le choix de la formalisation mathématique de la population*, des problèmes d'estimation, impliquant *le choix de méthodes de calcul de quantités dérivées de l'échantillon*, que nous appellerons statistiques, construites pour estimer les valeurs des paramètres de la population hypothétique, et en des problèmes de distribution. »

Deux des aspects essentiels de la problématique que nous abordons ici sont donc soulignés : la description et la formalisation des données d'une part, et, d'autre part, le choix de méthodes ou modèles de calculs appropriés. L'estimation procède d'abord d'un choix de méthode de calcul, et ce choix repose sur des hypothèses qui sont émises d'après la connaissance que l'on a des données. Dans cette thèse, le principal aspect qui sera traité concerne d'abord et avant tout donc l'accès à la connaissance des données, par une bonne description, et éventuellement l'usage de modèles, de méthodes ou d'outils adaptés à leur exploration.

On peut parler de statistiques en général : par exemple, le prix moyen d'une voiture sur le marché ou les préférences de consommateurs de produits de haute technologies sont des informations (quantitatives ou qualitatives) qui se réfèrent à un groupe d'individu (les voitures, les consommateurs) qui ne sont pas nécessairement localisés sur l'espace géographique. Cependant, cette thèse traite de la modélisation de la statistique *territoriale* qui produit au contraire des données ancrées sur des sous-portions de l'espace géographique. Dans Brunet, [Brunet 92], la statistique est définie justement par rapport à cet aspect territorial :

> « Activité qui vise à collecter des mesures sur les populations de toutes sortes - mais surtout des populations humaines, dans leurs caractéristiques, leurs activités, leurs production. La statistique aide au gouvernement de l'Etat (d'où son nom). »

Ici, c'est en vérité de la statistique publique dont il est fait état. On peut encore se référer à la définition qu'en donne l'INSEE (Institut National de la Statistique et des Etudes Economiques) :

> « Les statistiques publiques regroupent l'ensemble des productions issues des enquêtes statistiques [...] et de l'exploitation, à des fins d'information générale, de données collectées par des administrations, des organismes publics ou des organismes privés chargés d'une mission de service public. »

A.1.1 Objectif des statistiques

L'un des aspects de la définition de Brunet porte sur l'objectif, la raison d'être des données statistiques puisqu'elle « *aide au gouvernement de l'Etat* ». Cet objectif est également celui que rappelle la Direction de la Statistique Européenne, EUROSTAT [2], mais ici pour la gouvernance de l'entité supranationale que constitue l'Europe :

2. `http://epp.eurostat.ec.europa.eu/portal/page/portal/about_eurostat/corporate/introduction`

> « La mission d'Eurostat consiste à fournir à l'Union européenne un service d'information statistique de haute qualité. [...]. Il s'agit d'un rôle clé. Les démocraties ne sont pas en mesure de fonctionner correctement si elles ne peuvent pas s'appuyer sur des statistiques fiables et objectives. D'une part, celles-ci sont nécessaires aux responsables au niveau communautaire, national, local et aux chefs d'entreprises pour prendre leurs décisions. D'autre part, elles permettent à l'opinion publique et aux médias de se faire une idée précise de la société contemporaine et d'évaluer les résultats, notamment de l'action politique. Bien sûr, les statistiques nationales demeurent importantes au niveau des États membres. Les statistiques de l'Union Européenne, quant à elles, sont indispensables pour toute décision et évaluation au niveau européen. »

Les statistiques apportent des réponses à de nombreuses questions. La société évolue-t-elle conformément aux promesses des hommes politiques ? Le chômage augmente-t-il ou diminue-t-il ? Les émissions de CO2 sont-elles plus importantes qu'il y a dix ans ? Quel est le nombre de femmes actives ? Quelle est la santé de l'économie dans votre pays par rapport aux autres États membres ?

Comme nous le verrons par la suite, cet objectif a de nombreux impacts au niveau de la description des données, car tous les États ne poursuivent pas les mêmes objectifs. De plus, ces objectifs sont susceptibles de changer au cours du temps, et donc les données produites ne sont pas forcément équivalentes. Ceci en dépit des efforts incontestables pour atteindre un meilleur niveau de compatibilité des données statistiques entre états, efforts coordonnées par des entités supra-nationales comme EUROSTAT en Europe ou l'Organisation des Nations Unies (ONU).

A.1.2 Qui produit l'information statistique ?

La responsabilité de la production de statistique(s) est du ressort exclusif des États. L'INSEE est le producteur officiel de données en France. Ainsi, les données diffusées par l'ONU ou bien la CIA (*Central Intelligence Agency*) (*via* le *World Fact Book* [3]) sont, en vérité, une compilation de données produites au niveau des États. De même, en Europe, EUROSTAT, n'est pas habilitée à la collecte des données, [Terrier 00] :

> « L'Europe s'est dotée d'une Direction de la Statistique, EUROSTAT, mais cette Direction ne procède en direct à aucune enquête ni autre collecte statistique propre auprès des citoyens ou des entreprises. Elle se contente de rassembler et d'utiliser les statistiques produites et fournies par les instituts nationaux. Son rôle consiste donc essentiellement à faire converger les concepts, les nomenclatures, les méthodes et la temporalité des enquêtes afin de disposer de statistiques à peu près comparables entre les pays. »

La division statistique de l'ONU [4] vise comme Eurostat à améliorer la diffusion, l'harmonisation et la qualité des données statistiques, mais ne collecte pas de données. Parfois, à l'intérieur des États, comme le Royaume-Uni [5], certaines régions conservent leur prérogatives concernant la production de statistiques : l'Écosse, le pays de Galles, l'Irlande du Nord, et l'Angleterre possèdent leur propre institut de collecte des données. Ici apparaissent les premières sources d'incompatibilité entre données : jamais collectées par un seul et même producteur, elles sont collectées au niveau des États, ou à des échelons de gouvernance inférieurs.

3. https://www.cia.gov/library/publications/the-world-factbook/
4. http://unstats.un.org/unsd/default.htm
5. http://www.statistics.gov.uk/hub/index.html

Au niveau étatique, cette information possède plusieurs sources : les registres de l'Etat civil. les recensements, ou les enquêtes statistiques. L'information est donc issue de trois sources, distinctes mais complémentaires. Bien qu'en France la tradition statistique soit ancienne, pour de nombreux pays, la tenue de registres d'État Civil, et de recensements fiables n'est pas systématique. Par exemple, le Nigéria a eu pendant très longtemps de gros problèmes pour organiser des recensements fiables [Locoh 95]. Les définitions et les précisions que nous apporterons concernant cette information sont donc d'abord liées au contexte national français. Afin de relativiser et bien montrer que nous sommes dans une perspective de collecte sur des aires géographiques dépassant largement la France (l'Europe en général, le Maghreb aussi, voir le monde), certains points spécifiques à d'autres pays sont soulignés au passage.

A.1.3 Comment est produite l'information statistique ?

A.1.3.1 L'État Civil Les registres de l'Etat civil constituent la source de données la plus régulière (groupement par année des données). L'état civil, régi par un cadre législatif, existe en France depuis la Révolution française. De cette époque date l'enregistrement systématique des naissances, des mariages et des décès dans des registres communaux ; le maire, officier d'état civil, est responsable de leur tenue (auparavant, les actes d'état civil existaient mais sous la forme de registres paroissiaux déposés dans les évêchés). C'est sur ce socle qu'a été bâti le système de recueil de données sur les naissances, les reconnaissances d'enfants, les mariages, les divorces et les décès enregistrés en France, utiles pour l'analyse de la situation démographique et de son évolution [Le Bras 93]. L'INSEE publie une note méthodologique [6] détaillant comment ces données sont utilisées pour l'établissement de statistiques annuelles.

La constitution de l'Etat Civil d'un individu implique l'usage de règles (concernant par exemple le statut marital ou la nationalité) qui dépendent de chaque État. En vue de coordonner les approches gouvernementales, et de résoudre certains point délicats concernant, par exemple, la reconnaissance du décès d'un enfant à la naissance, la validité d'un mariage, la question de la transexualité, une Commission Internationale de l'État Civil (CIEC [7]) a été créée après la Seconde Guerre mondiale. Sont membres certains pays d'Europe comme l'Allemagne, la Belgique, le Royaume-Uni, etc. Cependant, par exemple, les Etats-Unis n'en font pas partie, et pour eux, l'obligation de tenir des registres d'Etat civil n'a été imposée qu'à partir de 1920.

A.1.3.2 Les recensements Les recensements sont la deuxième source de données, dont la période de production est généralement bien plus longue. Le recensement de la population a pour objectif le dénombrement des logements et de la population résidant et la connaissance de leurs principales caractéristiques : sexe, âge, activité, professions exercées, caractéristiques des ménages, taille et type de logement, modes de transport, déplacements quotidiens. La tradition est d'effectuer un recensement complet de la population au moins tous les 10 ans, voir en dessous. Par exemple, la Constitution Américaine exige un recensement de la population tous les 10 ans aux Etats-Unis. En France, institué en 1801, le recensement s'est déroulé tous les 5 ans jusqu'en 1936. De 1946 à 1999, les intervalles inter-censitaires ont varié de 6 à 9 ans. Un recensement exhaustif coûte cher, il exige du personnel, et des conditions spécifiques pour être valable.

Dans un souci d'économie, la loi du 27 février 2002, relative à la démocratie de proximité, a donc modifié en profondeur les méthodes de recensement en France. Depuis janvier 2004, le comptage tra-

6. http://www.insee.fr/fr/methodes/sources/pdf/etat_civil_les_sources.pdf
7. http://www.ciec1.org/index.htm

ditionnel est remplacé par des enquêtes de recensement annuelles. Les communes de moins de 10 000 habitants continuent d'être recensées exhaustivement, comme lors des précédents recensements mais une fois tous les 5 ans au lieu de tous les 8 ou 9 ans. Les communes de 10 000 habitants ou plus font désormais l'objet d'une enquête annuelle auprès d'un échantillon de 8 % de la population, dispersé sur l'ensemble de leur territoire. Au bout de 5 ans, tout le territoire de ces communes est pris en compte et les résultats du recensement sont calculés à partir de l'échantillon de 40 % de leur population ainsi constitué.

A.1.3.3 Les enquêtes statistiques Une enquête statistique consiste à observer une certaine population (élèves d'une classe, personnes âgées de 20 à 60 ans dans une région donnée, familles dans une région donnée, exploitations agricoles, appartements, travailleurs, etc.) et à déterminer la répartition d'un certain caractère statistique (note obtenue, taille, nombre d'enfants, superficie, nombre de pièces, secteur d'activité, etc.) dans cette population. Concernant la temporalité des enquêtes, il n'existe aucune règle en la matière, la fréquence peut être variable (annuelle, mensuelle, etc.), les enquêtes peuvent également être menées de façon dite « continue », comme l'enquête emploi en France à partir de 2003 [Goux 03]. Les règles pour la conduction d'enquêtes statistiques sont complexes car il s'agit d'interroger un échantillon représentatif de la population ciblée, comme d'établir les modalités de l'enquête (sondage par téléphone, visite à domicile, etc.) ainsi que le questionnaire.

A.1.4 Le secret statistique

L'information statistique est soumise à la règle du *secret statistique*, défini dans la loi n° 51-711 du 7 juin 1951 :

> « Le secret statistique interdit, pendant une durée de soixante-quinze ans et sauf dérogation [...], toute communication de données ayant trait à la vie personnelle et familiale, et plus généralement, aux faits et comportements d'ordre privé recueillies au moyen d'une enquête statistique. Pour leur part, les renseignements d'ordre économique ou financier ne peuvent être communiqués à quiconque pendant une durée de vingt-cinq ans, sauf dérogation [...]. »

Le secret statistique est un principe de protection très fort des individus et des entreprises ciblés par les enquêtes, recensements ou sondages, et il impose des règles strictes au niveau de la diffusion des données. Ces règles interdisent la publication de données qui permettraient une identification indirecte des répondants et de leur réponse, concept appelé « impossibilité d'identification ». Par exemple, si un tableau donne la répartition par âge et situation matrimoniale et que les personnes d'un certain âge (par exemple 50 à 59 ans) ont toutes le même état matrimonial (par exemple, divorcées), le secret statistique n'est plus respecté dans ce tableau, et ce dernier n'est donc pas diffusable. Pour les entreprises, on ne publie aucun résultat qui concerne moins de trois entreprises, ni aucune donnée pour laquelle une seule entreprise représente 85% ou plus de la valeur obtenue.

Ces règles *limitent donc la finesse des informations au niveau de la diffusion* et ont aussi comme conséquence qu'en réalité les chiffres publiés sont des agrégats statistiques. Pratiquement, les données sont donc agrégées dans des tableaux, regroupant les décomptes d'individus par lieu, et par catégorie sociale, professionnelle, ethnique, etc. Les nomenclatures servant à établir ces regroupements, d'ordre spatial ou thématique, sont arbitraires, variables dans le temps, variables dans l'espace.

A.1.5 Transformation des données statistiques

Avant d'être diffusée, la donnée brute (la micro-donnée) est généralement transformée. Les processus de transformation sont complexes. Ils peuvent être le fait des producteurs de données, mais également des organismes habilités à diffuser ces données. Ainsi, des organismes comme l'OCDE (Organisation pour la Coopération et le Développement Economique) n'hésitent pas à modifier les données brutes pour améliorer leur cohérence, ou d'autres à adapter les chiffres à une certaine vision politique du monde (le « *World Factbook* » produit par la CIA aux Etats-Unis en est un bon exemple).

L'exemple du chômage montre à quel point ces processus de transformation peuvent faire diverger les résultats produits, même lorsque les indicateurs se basent sur des définitions communes. Ainsi, en dépit d'une tentative d'harmonisation européenne symbolisée par le partage d'une définition commune définie par l'Organisation Internationale du Travail, l'INSEE et Eurostat publient des chiffres de chômage différents pour la même unité (la France) à la même date : ainsi, le taux de chômage publié par l'INSEE en février 2008 (8,4 %) diffère de celui estimé par Eurostat (8,8 %). Pour les deux instituts, un chômeur est une personne qui n'a pas eu d'activité rémunérée supérieure à une heure pendant une semaine, et qui peut prouver sa recherche d'emploi. Cependant, les méthodes de calcul, de pondération et de correction des chiffres à partir de l'enquête emploi trimestrielle diffèrent entre l'INSEE et Eurostat. Une explication détaillée de ces méthodes est publiée dans des documents publics [8]. Par exemple, Eurostat explique qu'il s'appuie sur les chiffres de l'enquête emploi publiée par l'INSEE, mais qu'il les réajuste ensuite :

> « Les séries mensuelles sur le chômage et l'emploi sont calculées dans un premier temps au niveau de quatre catégories (hommes et femmes de 15 à 24 ans, hommes et femmes de 25 à 74 ans) pour chaque État membre. Ces séries sont ensuite corrigées des variations saisonnières et tous les agrégats au niveau national et européen sont calculés. [...] Pour la Suède et la Finlande, la tendance-cycle a été utilisée à la place des données corrigées des variations saisonnières jugées trop volatiles. [...] »

Il faut noter ici que le texte explique comment sont transformées les données (par l'usage de la tendance-cycle en Suède par exemple), sans pour autant produire la formule exacte permettant de revenir à la source, c'est-à-dire au niveau des données brutes. Le chômage illustre aussi l'évolution des méthodes de calcul et de mesure dont sont l'objet les indicateurs. [Goux 03] explique que l'INSEE fait évoluer régulièrement sa méthodologie de calcul du chômage, décrite et justifiée dans des documents accessibles en ligne [9].

Par ailleurs, les indicateurs statistiques sont parfois le produit de réflexions théoriques qui visent à produire une représentation synthétique d'un ensemble de facteurs mesurés par des données brutes. Une étude récente de l'[UMS 2414 RIATE 08] portant sur le déclin démographique en Europe, et présentée devant le parlement européen en 2008, donne un exemple de ce type d'indicateur. Dans le texte de l'étude se trouve la définition d'un indicateur synthétique de vieillissement :

> « Il suffit alors d'effectuer le rapport entre l'âge moyen d'une population et son espérance de vie en bonne santé pour en déduire un *indicateur synthétique de vieillissement* exprimé sous la forme d'un pourcentage du potentiel d'activité de la population qui a été consommé. »

8. INSEE : `http://www.insee.fr/fr/methodes/sources/pdf/eeencontinu.pdf`
EUROSTAT : `http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-29012010-AP/FR/3-29012010-AP-FR.PDF`

9. `http://www.insee.fr/fr/methodes/sources/pdf/estimations_chomageBIT_enquete_emploi.pdf`

Cette définition est accompagnée d'un schéma, reproduit dans la figure 1, en vue de faciliter sa compréhension.

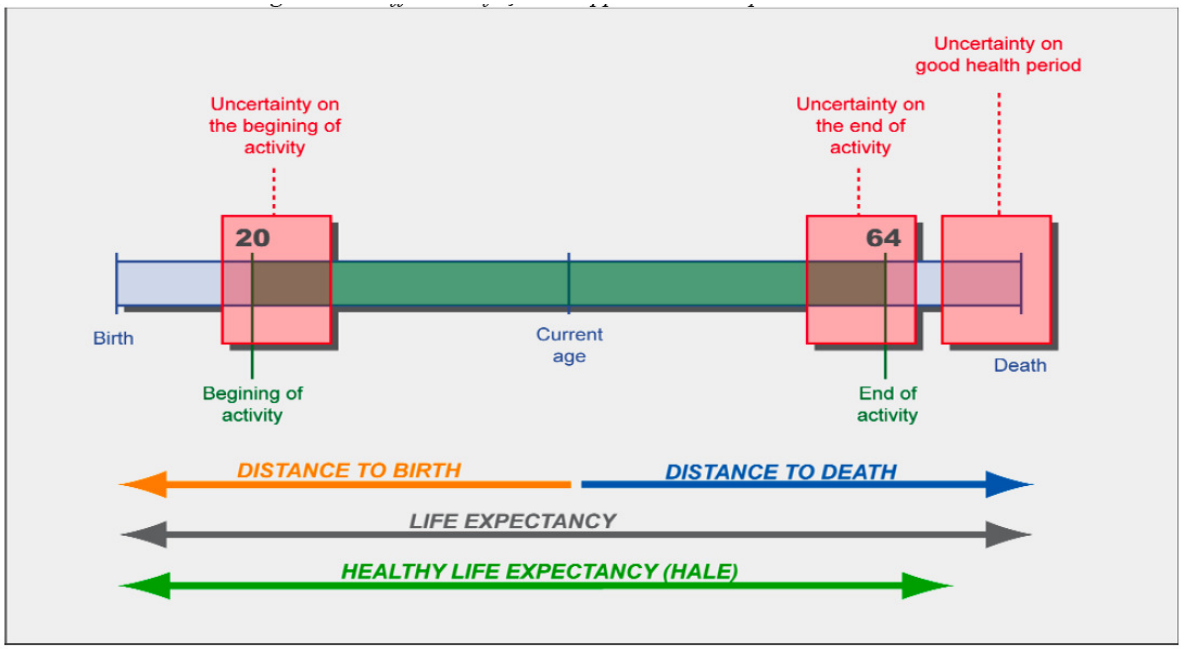


FIGURE 1 – Différentes façons d'appréhender l'espérance de vie, d'après [UMS 2414 RIATE 08].

A travers cet exemple, il apparaît que la représentation du calcul d'un indicateur composite à partir d'indicateurs de base est possible (ici, c'est le ratio de deux indicateurs de base qui fournit l'indicateur). En effet, si la relation mathématique entre l'indicateur synthétique de vieillissement (noté V) et l'espérance de vie en bonne santé est connue, il devient possible d'estimer des valeurs manquantes de V en employant des indicateurs bruts (espérance de vie, âge moyen). Par ailleurs, la connaissance de la formule de calcul permettrait de déduire que l'indicateur V est un bon *proxy* [10] pour calculer l'espérance de vie d'une population, lorsque son âge moyen est connu. Dans ces conditions, si les transformations sont clairement exprimées, ainsi que les éléments sur lesquels elles opèrent, l'analyse de similarité est quasiment directe.

Cependant, les formules permettant de fabriquer des indicateurs composites sont rarement si simples. Un guide méthodologique résumant les bonnes pratiques est publiée par l' [OCDE 08], pour l'ensemble des opérations qui doivent être mises en œuvre : sélection, normalisation, estimation de données manquantes, pondération et agrégation des données. Bien qu'effectivement les opérations soient plus complexes, l'étude de ce document fait apparaître l'usage récurrent d'une certaine terminologie pour ces opérations : les mots « correlation », « moyenne », « variance », « hypothèse » reviennent très fréquemment. La description des transformations recourt généralement à un vocabulaire de la statistique avancé pour lequel il n'existe pas, à l'heure actuelle, de formalisation informatique permettant de retranscrire aisément ces manipulations dans un format structuré exploitable par des automates. Ainsi, ces manipulations sont décrites dans des documents textuels, non structurés, le plus souvent dans la langue du producteur des données.

10. *proxy* est un des termes employés en statistiques comme synonyme de variable auxiliaire aidant à retrouver une variable inconnue.

A.2 La spatialisation de l'information statistique

L'information statistique est relative à des *zones de recensement*. Bien que localisée, ni sa représentation spatiale, ni son géoréférencement ne sont immédiats. En effet, au contraire de relevés de température de Météo France qui s'effectuent en un lieu géoréférencé, cette information peut être diffusée avec seulement un code associée à l'unité de recensement, sans que les limites exactes de cette unité de recensement soient produites. En particulier, si l'on remonte dans le passé, aucune limite précise des paroisses, qui jusqu'à la Révolution française servaient à l'enregistrement de l'Etat civil, ne peut être établie [Motte 03]. Ces zones de recensement sont diverses et interviennent à différents niveaux d'échelle. Par exemple, le site de l'INSEE recense au moins quatre différents types de zones de recensement (ou zonages) :

- Les unités urbaines
- Le zonage en aires urbaines et aires d'emploi de l'espace rural
- Les zones d'emploi
- Le découpage infra-communal en IRIS

Pour la statistique publique, c'est l'*îlot statistique*, plus petite unité spatiale de dénombrement, qui sert de base à la collecte des données, puis à leur diffusion. Ce principe est international, mais peut avoir plusieurs appellations, suivant les aires géographiques ou les périodes temporelles. Par exemple, au Canada, jusqu'en 2006, l'Îlot de Diffusion (ID) est le terme consacré. En France, l'îlot était l'unité géographique de base pour la statistique et la diffusion des recensements de la population jusqu'à celui de 1999. Mais la petitesse de l'îlot posant des problèmes de confidentialité, la Commission Nationale de l'Informatique et des Libertés (CNIL) a imposé des zones plus grandes. Désormais, la brique de base pour la diffusion des résultats du recensement rénové, qui débute avec le millésime 2006, est le quartier IRIS (pour " Ilots Regroupés pour des Indicateurs Statistiques ") et l'îlot est abandonné. Avec le découpage du territoire en IRIS, l'INSEE vise une taille de 2 000 habitants par maille élémentaire. Cette taille est censée protéger le secret statistique, mais avec la conséquence que le territoire n'étant pas peuplé de façon uniforme, les IRIS ne sont pas de surfaces comparables. Constitués par regroupement d'îlots, les règles et modalités qui président à la constitution des frontières des IRIS ne sont pas complètement objectives non plus : ces frontières sont établies sur avis de commissions, non forcément indépendantes des enjeux locaux liés à ces IRIS. Également, les frontières des IRIS changent dans le temps.

En réalité, un nombre nettement plus important de zonages (ou de maillages) co-existent sur un même espace géographique. La figure 2, inspirée d'un essai de typologie établi par [Le Gléau 99], présente une liste d'exemples de découpages co-existant simultanément sur le territoire français. Chacun d'entre eux est susceptible de servir d'espace de collecte ou d'espace d'analyse pour l'information statistique.

Plusieurs types de maillages sont superposés mais pas nécessairement exactement emboîtés dans un espace géographique : le maillage de la propriété correspond, par exemple, au dessin des parcelles du cadastre, c'est en général le plus fin ; le maillage des circonscriptions administratives est généralement hiérarchisé et emboîté, une subdivision de niveau inférieur étant partie intégrante d'une unité de niveau supérieur (mais ce n'est pas toujours le cas, il peut exister des enclaves d'un niveau dans un autre ou bien un maillage incomplet à un niveau donné). Cette forme d'emboîtement des maillages sur différentes échelles (ou niveaux d'analyse) constitue une *structure multiniveau* [Mathian 01]. Mathian et Piron synthétisent quels sont les avantages sur le plan de l'analyse géographique de modéliser ces niveaux en vue d'une analyse contextualisée [Mathian 01] :

> « Parcourir des niveaux différents produit en effet un changement dans la perception du phénomène étudié : aux niveaux supérieurs, les détails disparaissent au profit de formes globales et d'étendues plus larges, tandis que les niveaux les plus fins révèlent des différenciations spatiales locales. Le découpage joue comme un filtre qui affecte jusqu'à la signification de

	Catégories de zonage	Exemples
zonages de pouvoir	zonages institutionnels	États régions départements arrondissements communes syndicats (SIVU, SIVOM) communautés urbaines communautés d'agglomération communauté de communes cantons circonscriptions législatives
	zonages administratifs spécialisés	districts scolaires secteurs sanitaires régions militaires zones ANPE
	zonages d'intervention	zones de fonds structurels de l'Union Européenne zones éligibles à la PAT parcs naturels zones urbaines sensibles
zonage de savoir	zonage en "tâches"	Unités urbaines Z.P.I.U. typologies des espaces à dominante urbaine aires urbaines
	zonages en "mailles"	régions agricoles zones d'emploi bassins d'emploi zones de petite chalandise bassins hydrologiques

FIGURE 2 – Différentes catégories de zonages.

la mesure du phénomène [Raynal 96]. Le changement d'échelons d'observation peut avoir une dimension heuristique, il permet de :

– définir les différents niveaux d'organisation d'un phénomène ;
– mettre à jour de nouvelles structures spatiales et de nouveaux objets spatiaux ;
– s'intéresser à la pertinence d'un découpage géographique, raisonné ou arbitraire, en tenant compte de la variabilité interne des unités spatiales. »

A.2.1 Agrégation spatiale

Les mesures attachées aux entités des structures multiniveau résultent le plus souvent, de *procédures d'agrégation statistique*. Lorsque celles-ci sont implicites, elles sont sources d'ambiguïtés dans la construction des données, l'emploi de la méthode et l'interprétation des résultats. Ainsi, comme l'illuste D'Aubigny [D'Aubigny 94], voir figure 3, on peut distinguer dans une structure multiniveau

– le *niveau élémentaire* (ici niveau h = 0), dont les entités sont appelées atomes. C'est le niveau le plus fin de la hiérarchie pour lequel l'information est disponible ;
– le *niveau de collecte* (niveau h = -1), qui est le niveau auquel les mesures ont été réellement prises.

Lorsque l'on a accès au niveau de collecte, ces deux niveaux peuvent être confondus. Mais, en géographie, lorsque les unités étudiées sont des entités spatiales, ces niveaux sont le plus souvent distincts. Les mesures associées aux entités spatiales (niveau élémentaire) sont effectuées sur des entités de niveau inférieur (niveau de collecte), qui ne sont pas introduites dans le système d'information multiniveau.

Par exemple, dans l'analyse des dynamiques des communes, l'atome est la commune. Mais si l'analyse porte sur les populations des communes, les résidants constituent, de manière implicite, un niveau inférieur alors non mesurable.

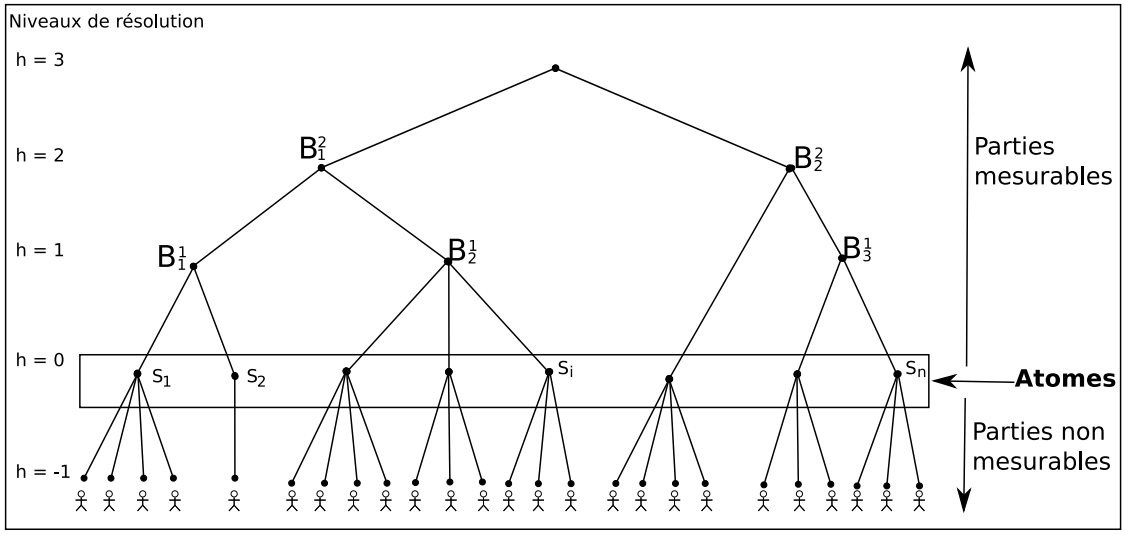


FIGURE 3 – Mesurabilité et hiérarchie de niveaux (d'après [D'Aubigny 94]).

Le passage d'un niveau à un autre met en oeuvre des procédures d'agrégation qui transforment l'information, celle-ci étant analysée non plus suivant l'unité élémentaire dont elle est issue, mais suivant une unité agrégée. On s'expose dès lors aux risques de l'*erreur écologique* : l'erreur écologique est une erreur d'interprétation qui consiste à inférer des résultats obtenus à un niveau inférieur à celui de l'analyse. Ce qui signifie accorder à tous les individus d'une population aggrégée les même comportements, ignorant les particularités et motivations propres de chacun. La tentation de s'abstraire du maillage pour échapper aux généralisations qu'amènent les procédures d'agrégation statistique est donc forte. Cependant, analyser les individus par rapport à leurs caractéristiques propres, mais indépendamment de leur environnement et des différents contextes qui influencent leurs comportements produirait le biais inverse, celui de l'*erreur atomiste*.

Par ailleurs, les mailles présentent souvent une grande hétérogénéité de taille et de forme ; ces découpages sont remis en cause, car ils introduisent des biais importants sur les organisations et relations spatiales observées et sur leurs interprétations [Openshaw 81], [Fotheringham 91], [Grasland 00]. Il s'agit du problème désigné en anglais par l'expression *Modifiable Areal Unit Problem* (*MAUP*) ou plus généralement le *Change Of Support Problem (COSP)*, qui a donné lieu à divers développements méthodologiques. Ces approches sont l'objet de discussions quant à leurs apports et contraintes respectives [Gotway 02], [Grasland 06], car elles dépendent de la nature des données traitées ainsi que des hypothèses de modélisation sous-jacentes au phénomène étudié.

Sans entrer dans ces considérations et pour résumer, le MAUP est un problème portant sur l'interprétation statistique qui peut être faite de données issues de maillages. Il apparaît que la forme et l'échelle du support des données jouent un rôle très important sur l'interprétation statistique des données. Le premier biais d'interprétation est un effet de l'agrégation : on constate que les résultats de l'analyse de la répartition spatiale des données varient en fonction de la taille des unités. Par exemple, le traitement d'un jeu de données communales au niveau de son découpage initial ne donne pas des résultats identiques à celui qui serait fait en regroupant les données au niveau départemental. Ce qu'on appelle communément l'« effet

d'échelle » (« *scale effect* ») s'explique par le fait que la variance de l'échantillon initial (communal dans l'exemple) est mathématiquement plus élevée que la variance de la moyenne de ces échantillons. Le second biais, dénommé « *effet du zonage* » (« zoning effect »), dérive du mode de regroupement des données à une échelle fixée : [Openshaw 79] prouve que la corrélation entre des données spatiales varie en fonction du découpage territorial (comparaison faite entre entités de surfaces comparables). La figure 4 résume la situation.

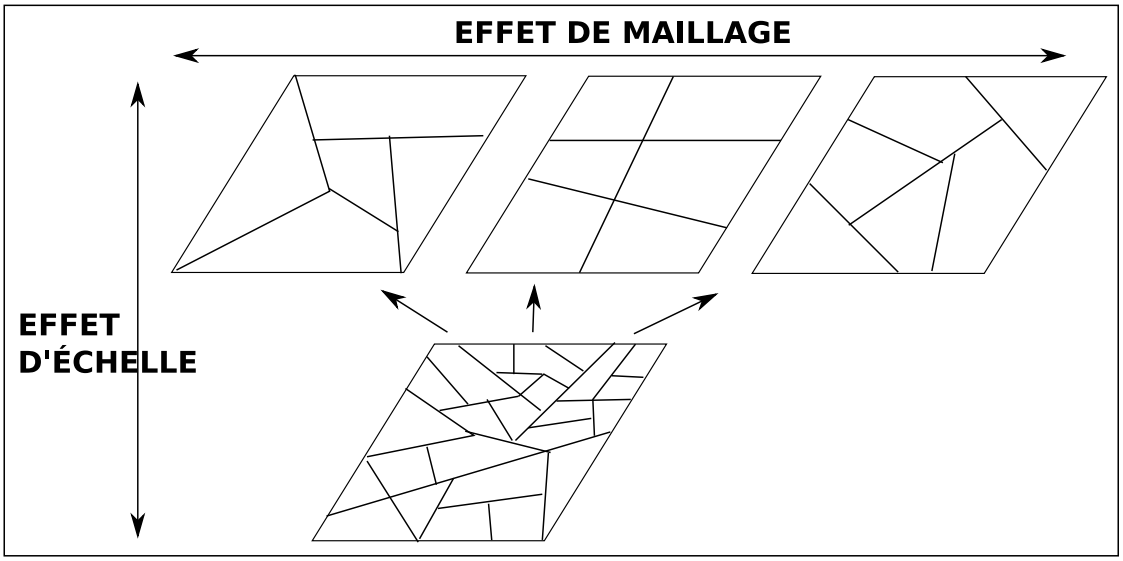


FIGURE 4 – L'effet d'échelle et de découpage (d'après [Fotheringham 91]).

A.2.2 Unités de recensement : zonages ou maillages ?

L'INSEE emploie le terme de *zonage* pour définir le support spatial de collecte des données. Ce terme est à mettre en relation avec la notion de *maillage* [Pumain 10], autre terme qui a été plus longuement débattu en géographie. Le mot « maillage », renouvelé depuis les années 1950, désigne au sens premier un filet ou un réseau, et renvoie au terme de maille. Le terme maille semble dériver du latin *macula* qui désignait à la fois la boucle et la tâche - le terme de maille pourrait également dériver de *mezg*, tisser, nouer, qui a donné en anglais le terme *mesh* (filet) [Brunet 92]. Cette définition souligne *le caractère dual du maillage* : dans le cas d'un filet, la maille peut désigner soit le réseau des fils entrelacés (sens réticulaire), soit l'espace vide dont la boucle définit le contour (sens territorial). Sur Hypergéo, l'encyclopédie électronique en ligne consacrée à l'épistémologie de la géographie, le maillage est ainsi défini [Pumain 04] :

> « Une partition d'une zone géographique divisée en unités contiguës dont la forme et la taille peuvent être régulières ou irrégulières. Par exemple, l'ensemble des limites des unités administratives qui couvrent un territoire forment un maillage polygonal, généralement irrégulier. Mais un maillage peut aussi désigner un réseau dont l'objectif est la desserte complète d'un territoire, comme le maillage du réseau électrique ou celui des zones d'accès à la téléphonie cellulaire. Le terme est ambigu, et s'emploie pour désigner parfois l'ensemble des divisions de l'espace (grandes mailles, petites mailles), et parfois le réseau des frontières ou des limites qui les constituent (maillage serré, maillage lâche). »

En mathématiques, le maillage est aussi un terme technique qui désigne la discrétisation spatiale d'un milieu continu : c'est une partition, c'est à dire une famille de sous-ensembles exhaustifs et disjoints, c'est-à-dire encore tels que tout point appartient à un de ces sous-ensembles et à un seul. Ainsi défini, le terme de maillage est équivalent à ceux de partition ou de relation d'équivalence. En effet la relation « x appartient à la même classe que y » est réflexive, symétrique et transitive, définissant de la sorte des classes d'équivalence entre les éléments qui appartiennent à la même maille territoriale.

Au sens géographique, le maillage n'est pas seulement un découpage servant à la collecte des données, comme semble l'être le zonage. Brunet [Brunet 92] propose la définition suivante :

> « Ensemble des filets qui situent les lieux dans les mailles de l'appropriation et de la gestion du territoire, et principe de partition opératoire et socialisé de l'espace. Le maillage va de la parcelle à l'Etat à travers toute l'échelle géographique. »

Il développe ainsi son point de vue :

> « L'espace est " parti " de mailles. C'est l'une de ses caractéristiques fondamentales. Les processus d'appropriation produisent, par définition, des partitions. La maîtrise du territoire, et de ses ressources tant humaines que physiques, nécessite sa partition dès lors que l'on atteint une certaine masse et un certain degré de complexité. Il s'agit en effet :
>
> 1. de partager entre les familles le sol, pour exploiter ses ressources : cela fait les parcelles, les concessions, les exploitations agricoles.
> 2. d'assurer une base aux groupes élémentaires en lesquels se divise un peuple : cela fait les finages de villages, les territoires des tribus.
> 3. de disposer de relais du pouvoir, en leur attribuant une étendue qu'ils aient les moyens de maîtriser ; c'est alors une question de distance et de masse : le pouvoir " se rapproche " des citoyens en morcelant le territoire en niveaux successifs. Ce découpage facilite aussi bien les inventaires et les bilans que la police et le contrôle de l'application des lois.
>
> La première voie de la partition trouve son expression achevée dans le *cadastre*. La seconde et la troisième dans les *circonscriptions administratives*. [...] »

Dérivant de cette définition, le géographe Claude Grasland [Grasland 98] propose de définir un maillage territorial comme « *une partition simultanée de l'espace et de la société en sous-ensemble deux à deux disjoints* » : l'aspect social est ici mis en avant pour mettre en évidence le rôle structurant que joue le maillage dans la société. Ainsi, il note qu'un maillage constitue à la fois « un niveau d'observation des sociétés et de leur espaces, mais aussi un niveau potentiel d'organisation de la vie en société ».

Dans la littérature statistique, le terme *zonage* est souvent préféré à celui de maillage, considérant qu'un maillage est un zonage particulier : « une partition du territoire sans omission ni recouvrement. », [Le Gléau 99], [Terrier 00], [Terrier 05]. Le zonage n'a pas vocation à être une partition complète de l'espace. Soit le zonage est un cas particulier de maillage, soit le maillage est un cas particulier de zonage. On bien encore, on peut considérer les termes comme synonymes [Pumain 10].

Brunet [Brunet 92] explique que zonage est en réalité un anglicisme (traduction de « *zoning* ») qui « prend en compte des espaces quelconques, dits à tort « zones » : *non aedificandi, non altius tollendi* et toutes sortes d'espaces classés selon leur population, leur utilisation du sol, etc. ». En anglais, zonage est traduit par « *zoning* » et maillage par « *mesh* », mais « *mesh* » renvoie alors au sens de réseau. La littérature d'analyse spatiale emploie le terme de « zones » pour les unités spatiales d'un maillage et de aussi le terme « *support* » pour tout type de support des données géographiques (maillé, grillé, ponctuel). Dans la littérature statistique anglo-saxonne, dont l'article suivant dresse un état de l'art complet concernant

les problèmes de changement de support, [Gotway 02], le terme « *mesh* » est absent : on parle d' « *areal zoning system* » ou bien de *zoning*.

En conclusion, et comme illustré dans la figure 5, nous choisissons d'adopter dans la suite de ce manuscrit le terme de zonage, pour parler d'un découpage territorial (ou partition territoriale), couvrant complètement ou non l'espace géographique étudié. Le terme maillage sera dédié aux découpages territoriaux formant une partition complète et sans recouvrement de l'espace géographique étudié. Toutefois, les considérations afférentes à la signification et les fonctions d'un maillage sont en réalité les mêmes en ce qui concerne les zonages : ils restent des découpages de l'espace, ou des nomenclatures géographiques, [Pumain 10].

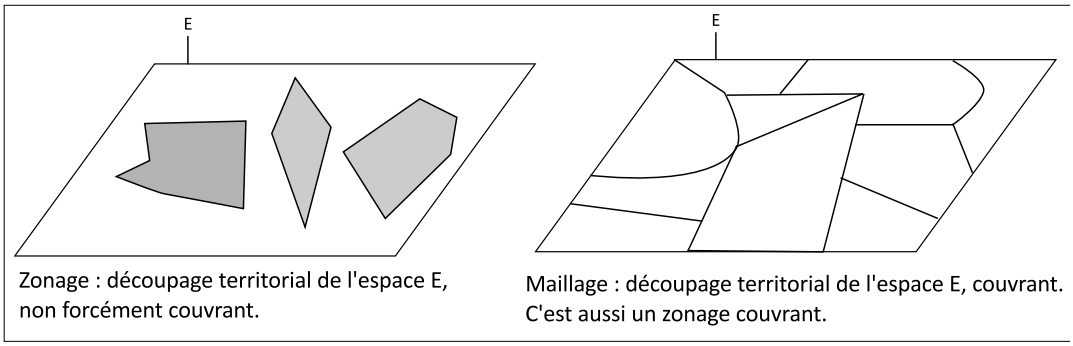


FIGURE 5 – Zonage ou maillage de l'espace.

A.2.3 Critique des découpages territoriaux

On peut s'interroger sur la signification d'un découpage territorial, dans le sens ou « celui-ci est généralement établi par un pouvoir en fonction de certains objectifs » [Grasland 98]. Grasland, comme d'autres avant lui, [Raffestin 80], insiste sur le fait qu'un découpage territorial tel que le maillage est avant tout *l'expression d'un pouvoir* :

> « Une propriété centrale des maillages territoriaux est de fournir à travers les recensements et les registres administratifs une description exhaustive des membres d'une société et de l'espace qu'ils occupent. Mais cette exhaustivité est purement formelle puisqu'elle repose sur l'utilisation de plusieurs grilles de collecte de l'information (sociologiques ou géographiques) qui ne donnent à connaître que la répartition globale des individus et des surfaces dans le cadre d'agrégats dont la pertinence reste dans la plupart des cas à démontrer. Les catégories sociales d'une part (activités, religions, ethnies, etc.) et les catégories spatiales d'autre part (quartiers, communes, départements, régions, etc.) constituent des modes d'observation de la société et de l'espace qui sont produits par un pouvoir en fonction de certains objectifs de contrôle ou de gestion et qui véhiculent l'idéologie sur laquelle repose ce pouvoir. »

L'exemple des recensements au Nigéria [Locoh 95] démontre l'aspect hautement politique des découpages et des recensements associés : sur l'espace d'un siècle, le Nigéria, d'abord colonie britannique depuis 1860 puis indépendant dès 1960, possède une histoire du recensement très chaotique, à tel point que le seul recensement valide (celui de 1992) à découvert qu'il n'y avait pas 120 millions mais 88,5 millions d'habitants au Nigéria. La raison de cette surprenante surestimation de la population tout au long du vingtième siècle est expliquée par l'empêchement de la tenue de recensements sérieux. Les raisons

de ces empêchements sont multiples, essentiellement d'ordre politique : d'abord durant la colonialisation, la population ne voulait pas être recensée pour éviter les impôts et les taxes afférentes. De plus, sur ce territoire vaste et mal desservi, un sous-effectif des agents de recensement empêchait la tenue d'un recensement complet de la population. Ensuite, après l'indépendance de 1960, il s'agissait pour chaque région du Niger de se voir attribuer une part équitable de la manne pétrolière :

> « Pour atténuer le clivage nord/sud, le gouvernement militaire décida de diviser le pays en douze États Fédérés. La période de prospérité liée au *boom* pétrolier qui suivit la guerre fut marquée par un accroissement considérable du budget fédéral. Chacun des douze États entendait bien obtenir une part équitable des investissements fédéraux. Le gouvernement fédéral décida que ces derniers seraient, pour une large part, fonction de l'effectif de la population de chaque État. »

Ainsi, dans ce redécoupage (voir figure 6) et le recensement de 1973 contesté puis invalidé qui s'en suivi, l'enjeu politique que constitue la forme du maillage et le recensement associé se révèle très clairement.

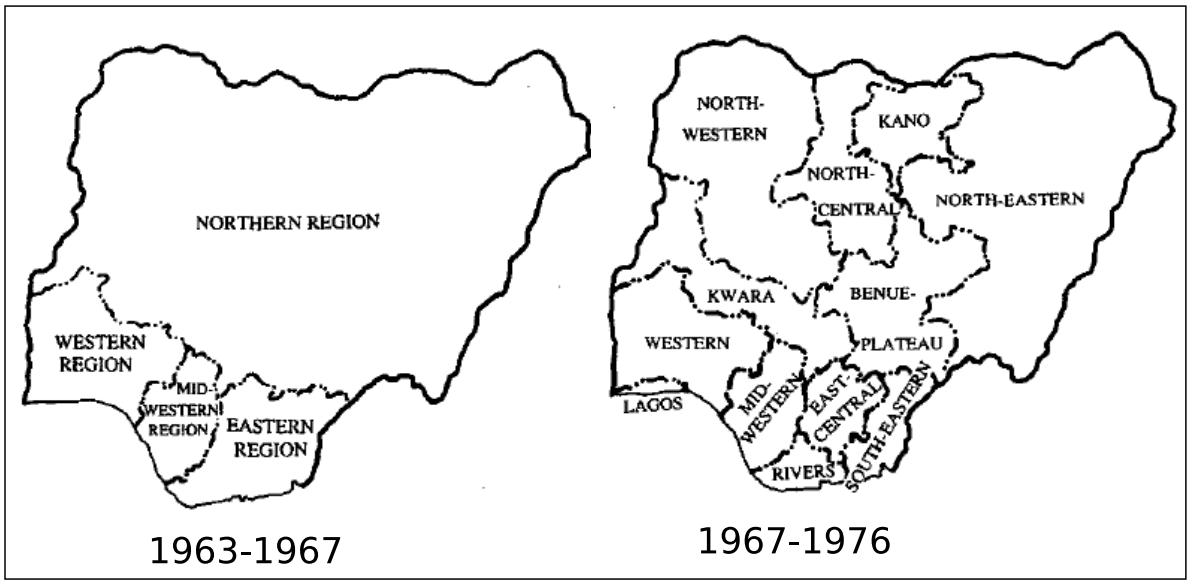


FIGURE 6 – Redécoupages du Nigéria en 1967

L'avis des statisticiens rejoint ici l'avis des géographes sur ces découpages de l'espace. En distinguant les *zonages de pouvoir* des *zonages de savoir*, Terrier rapporte bien que certains zonages sont effectivement pour fonction l'exercice d'un pouvoir [Terrier 98]. Les zonages de savoir seraient ceux qui auraient pour fonction l'amélioration d'un savoir en permettant de clarifier certaines connaissances et d'affiner les observations faites sur un territoire plus vaste. Il apparait encore cependant que les données utilisables dans ces zonages de savoir ont le plus souvent été produites dans des zonages de pouvoir. Terrier introduit également les *zonages de l'avoir* à propos desquels il dit, page 126 [Terrier 05] :

> « La définition et la pratique des découpages de l'avoir est très ancienne. Jean-Marie Delarue fait remonter l'existence des premiers zonages à la fondation de Rome et à la légende de Rémus et Romulus. Cette dernière ne nous dit-elle pas que Romulus a tracé un sillon pour déterminer l'enceinte de Rome ; son frère Rémus l'a franchi et il a été tué pour cela. C'est dire qu'on peut mourir pour des zonages. Il faut rappeler que ces jumeaux avaient été élevés par une louve et que les loups sont des animaux qui marquent leur territoire. »

Ainsi, il existe une multitude de zonages, correspondant à des objectifs particuliers, dont les limites suscitent toujours des débats passionnés.

Les critiques des découpages portent donc essentiellement sur le choix *a priori* d'agrégats qui relèvent d'une idéologie et d'une volonté de contrôle. Les sociologues sont également nombreux à poser la question de la pertinence des catégories qui servent à structurer (« mailler ») une population suivant des critères thématiques. Ils sont rejoints dans cette critique par les géographes, comme Grasland [Grasland 98] qui note après le sociologue Alain Chenu [Chenu 97], que les catégories statistiques sont également une forme de maillage de la population, subdivisant une population en classes d'individus identiques relativement à des critères établis par le pouvoir. Ces critères, qui définissent les catégories, sont tout aussi discutables que les critères employés pour mailler l'espace.

Par ailleurs, le scientifique n'accède généralement qu'à des agrégats de données, organisées comme nous venons de l'expliquer suivant des maillages, maillages qui ne conviennent pas forcément à l'objectif de l'étude visé par le scientifique. Le scientifique regrette donc l'absence de données individuelles, non qu'elles n'existent pas, mais parce que l'accès à ces données est protégé, en raison du secret statistique.

Enfin, dernière critique, le maillage est un mode particulier d'appréhension de la réalité qui privilégie l'appartenance stricte et exclusive par rapport à un autre mode de lecture fondé sur la multiplicité et le flou des appartenances des éléments aux classes. Le maillage tend ainsi à nier la notion d'*imprécision* et d'*incertitude* qui est inhérente à la plupart des dénombrements et des classifications en sciences sociales [Moles 95]. Même lorsque le chercheur est conscient du caractère essentiellement flou de son objet d'étude, il tend à privilégier des modes d'appréhension rigide de la réalité. Ceci est particulièrement flagrant en géographie régionale, comme l'a souligné par exemple C. Rolland-May dès 1984, [Rolland-May 84].

> « Exception faite des unités spatiales délimitées par l'homme pour des raisons politiques, administratives, juridiques, militaires ou autres motivations de domination spatiale, *il est souvent difficile, voire impossible, de fixer à un espace géographique une limite nette, linéaire, continue*. Le géographe se trouve le plus souvent en présence de marges, bordures, espaces « périphériques » ou autres zones de transition. [...] Paradoxalement, ces caractères, tout en étant reconnus au début de toute analyse spatiale, sont en quelque sorte occultés lors de cette analyse. La régionalisation spatiale par exemple, cherche à mettre en évidence une limite unique, nette, linéaire d'un espace géographique, même si cette limite est « arbitraire » (George P., 1970), artificielle ou résulte d'une synthèse plus ou moins arbitraire aussi de critères divers. De même, la géographie inductive quantitative admet implicitement la notion de précision en constituant la « matrice d'information chrono-spatiale ». Chaque unité spatiale est supposée appartenir entièrement et sans ambiguïté à l'ensemble étudié. Il nous semble ainsi que, tout en reconnaissant la notion d'imprécision spatiale, le contexte cartésien qui sous-tend toute science nous pousse à respecter la loi du tiers-exclus c'est-à-dire à adopter dans notre réflexion, nos méthodes d'analyse spatiale, de régionalisation ou de classification, l'idée qu'un élément spatial ne peut appartenir qu'à un espace et un seul. Dans cette optique, un espace imprécis, « plus ou moins » bien délimité, n'est pas susceptible d'une étude scientifique, car il ne se prête pas à une telle structure binaire de pensée et de réflexion. »

Par exemple, un individu a généralement une occupation de l'espace au cours de la journée qui alterne au moins entre son domicile à son lieu de travail, sans compter les lieux de loisir. Mais le recensement alloue à l'individu une place fixe, celle de son domicile. Ce mode cartésien d'appréhension de l'espace répond dans la plupart des cas à un souci de simplicité et d'efficacité. Il est directement lié au développement des méthodes statistiques fondées sur le dénombrement et l'échantillonnage, ainsi qu'au transfert de ces méthodes vers les sciences sociales.

A.3 Représentation des données statistiques

Cette section présente comment se modélisent et se représentent les statistiques territoriales dans les domaines de la statistique et de l'analyse spatiale, modélisation qui implique des modalités d'échange de l'information particulières entre acteurs du domaine statistique.

A toute *unité spatiale* (qui désigne tout objet localisé, ville ou région, ou cellule de base d'une grille, ou maille administrative) est associée une information spatiale (les données relatives à la localisation et à la forme d'une unité spatiale) et une information sémantique (données relatives aux caractéristiques et aux propriétés de l'unité spatiale). Les informations sont alors structurées en deux grands types de tableaux, qui ne se prêtent pas au même types de manipulations : les « tableaux d'information géographique » et les « tableaux d'échanges » [Pumain 97]. Les seconds sont utilisés pour décrire les relations entre unités spatiales, essentiellement toutes les sortes d'échanges entre des lieux ou des zones. Cependant cette thèse ne traite pas des problèmes relatifs à la modélisation des données d'échange. Nous nous intéressons donc en priorité aux tableaux d'information géographique, et à leur spécialisation en tableau de contingence.

A.3.1 Les tableaux d'information géographique

Un tableau d'information géographique associe à des unités spatiales, repérées par un identifiant (nom ou code) des attributs qui précisent des caractéristiques mesurées sur ces unités spatiales, (voir tableau 1). Dans ces tableaux, une ligne correspond à une unité spatiale, pour laquelle chaque colonne du tableau correspond à un attribut différent. Les attributs (encore appelés caractères ou variables) peuvent être qualitatifs (nominaux ou ordinaux) ou quantitatifs (mesurés sur des échelles d'intervalle ou de rapport).

TABLE 1 – Tableau d'information géographique, d'après [Pumain 97].

Code de l'unité spatiale	Variables ou attributs X_1 ... X_j ... X_p
1 . . i . .	X_{i1} ... X_{ij} ... X_{ip}

L'information spatiale (les limites du polygones ou les coordonnées de la cellule associées à cet identifiant) n'est pas intégrée directement dans ces tableaux, et très souvent, l'information est produite à part.

Un tableau est un « *jeu de données* », c'est-à-dire qu'il regroupe des variables statistiques, disponibles sur l'aire géographique représentée par l'ensemble des unités spatiales présentes dans le tableau. Un tableau regroupe des variables (ou indicateurs statistiques) qui peuvent être issus de plusieurs sources,

couvrir des temporalités différentes et des espaces différents, ce qui suscite en réalité un certain nombre de problèmes sur lesquels nous reviendrons.

A.3.2 Les tableaux de contingence

Un nombre conséquent d'indicateurs statistiques socio-économiques se présentent sous la forme de tableaux de contingence, qui associent des valeurs à des catégories croisées. Ils correspondent à une agrégation de l'information suivant une dimension thématique. L'information spatiale est parfois implicite : l'unité spatiale pour laquelle ces chiffres sont publiés n'est pas toujours précisée. C'est le cas notamment des données démographiques, qui sont publiées par sexe et par tranche d'âge, ou de la population active telle que les publie l'INSEE sur son site [11] et comme reproduit dans le tableau 2.

Population active (en milliers)	hommes	femmes	ensemble
15 ans ou plus	14 806	13 463	28 269
15-64 ans	14 702	13 394	28 096
15-24 ans	1 487	1 224	2 712
25-49 ans	9 576	8 756	18 332
50-64 ans	3 639	3 413	7 052
dont : 55-64 ans	1 801	1 677	3 478
65 ans ou plus	104	69	173

TABLE 2 – Population active selon le sexe et l'âge en 2009 en France.

Ce caractère multi-dimensionnel de l'information statistique est exposé dans [Rafanelli 90], qui présentent l'objet statistique comme étant un quadruplet $< N, C, S, f >$ où :

- N est le nom de l'indicateur statistique ;
- C est un ensemble fini de catégories (ou dimensions) $C_1, C_2, ..., C_n$, qui ont chacune leur unité de mesure, et un domaine spécifique ;
- S est un attribut résumant la variable quantitative mesurée, qui possède un domaine de valeur, et une unité de mesure ;
- f est la fonction d'agrégation utilisée pour résumer les valeurs (la somme, le compte, le minimum, le maximum ou la moyenne).

L'objet statistique défini dans l'exemple précédent porte sur N, la « population active » utilise le compte comme fonction d'agrégation f, et S correspond au décompte de personnes actives, qui sont classées suivant deux dimensions, e.g. par tranche d'âge (C_1), et sexe (C_2). Nous remarquons que cette modélisation des catégories sociales peut aisément être étendue à une nomenclature géographique, en choisissant, par exemple, les départements comme catégorie. En effet, les catégories ou les unités spatiales forment un maillage, thématique ou spatiale, du territoire.

Il est essentiel de noter que ces catégories thématiques (ou sociales) sont rarement homogènes sur l'ensemble des données collectées, car elles sont spécifiques au lieu ou/et à la période étudiés. Les catégories socio-professionnelles, les recensements à caractère ethnique ou les tranches d'âge sont des exemples de classifications qui relèvent d'un caractère politique et sont par conséquent instables dans le

11. Les données sont disponibles sur `http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATCCF03170`.

temps, hétérogènes dans l'espace. Les exemples abondent :

- Les recensements à caractère ethnique font débat, et l'histoire du recensement de la population aux États-Unis (déroulée par [Gauthier 02]) démontre comment les transformations politiques et sociales d'un pays peuvent amener à reconsidérer les classifications officielles employées.
- L'étude de [Arel 02] démontre comment en Russie, entre 1991 et 2000, des administrations sous influence mènent des recensements qui reposent sur une définition de la nationalité, de l'appartenance ethnique et du statut de migrant adaptés à une certaine vision politique.
- En France, les actifs sont classés en fonction de leur statut professionnel (salarié, chef d'entreprise, indépendant), de la taille de l'entreprise dans laquelle ils travaillent, du secteur de l'activité (primaire, secondaire ou bien tertiaire), du niveau d'études requis pour pratiquer leur profession, etc. Mais ce mode de classification de la population en catégories socio-professionnelles n'a pas d'équivalent européen comme l'explique [Kieffer 02], parce que chaque pays construit ces catégories en fonction de son histoire et de théories spécifiques.

B Une approche pluri-disciplinaire

L'approche proposée pour la gestion de l'information statistique territoriale est pluri-disiciplinaire, au croisement de disciplines comme l'informatique et la géographie, dans un domaine nouveau qu'on appelle la *géomatique*.

B.1 Définition de la géomatique

En pleine évolution, comme discipline, thématique de recherche ou ensemble d'activités, selon le sens qu'on lui confère, la géomatique est en train d'acquérir une dimension plus importante et théorique que ne lui conférait la définition en 1992 de Brunet [Brunet 92].

> « Ensemble des procédures de traitement des données géographiques par ordinateur ; le terme concerne surtout les bases de données numériques servant à la géodésie, au cadastre et aux cartes topographiques, et il est peu employé hors des cénacles spécialisés dans ces domaines. »

Aujourd'hui, la définition que produit la *Revue Internationale de Géomatique* correspond mieux à l'ensemble des activités que recouvre la géomatique :

> « La géomatique est une thématique de recherche pluri-disciplinaire orientée vers la représentation, la modélisation, l'intégration, l'analyse et la visualisation de données géographiques. Les domaines d'application sont très variés : aménagement et planification des territoires, systèmes et services de mobilité, espaces géographiques complexes et dynamiques. La géomatique rassemble plusieurs communautés de scientifiques : géographes, urbanistes, informaticiens, géomaticiens, agronomes, archéologues, etc., autour d'un objectif commun, celui du développement durable de notre environnement et des outils et des méthodes pour le construire. »

B.2 Objectif de la géomatique

Dans les différentes définitions actuelles, l'objectif commun des géomaticiens est d'intégrer les moyens d'acquisition et de gestion des données à références spatiales en vue d'aboutir à une **information d'aide à la décision** [12]. La géomatique est donc une science pluri-disciplinaire qui se met au service de l'aménagement du territoire en facilitant l'acquisition, le traitement et la diffusion des données sur le territoire (aussi appelées « données spatiales », « données géospatiales » ou « données géographiques »).

Ce travail est mené dans l'optique de concevoir des modèles et des méthodes pour aider à l'aménagement du territoire. En conséquence, *les questions relatives à l'accès, la compréhension, et l'évaluation de l'information statistique territoriale sont centrales dans ce travail*.

12. http://www.cc-tarndadou.fr/definition_sig.php

Introduction

Cette thèse se situe à la croisée de deux disciplines, la géographie et l'informatique, dans un domaine émergeant que l'on nomme la géomatique. La géographie, comme science de l'étude de l'organisation spatiale des sociétés humaines, cherche des concepts et des méthodes pour rendre compte du monde dans lequel vivent les hommes, et de ses évolutions. Elle doit pour cela exploiter des collections de données qui proposent des instantanés de nos sociétés à différentes époques, diverses échelles, sur des thèmes variés : l'économie, la démographie, la santé, l'éducation, etc. L'informatique d'aujourd'hui lui apporte des outils et des modèles pour organiser et analyser des collections de données de plus en plus vastes, et hétérogènes, grâce, par exemple, aux Systèmes d'Information Géographique (SIG).

L'information statistique territoriale est, en particulier, le moyen que se donnent les États pour connaître le territoire qu'ils gouvernent. Cette information à références spatiale et temporelle est collectée sur des unités zonales avec une fréquence plus ou moins régulière suivant les époques. Par exemple, en Europe, le chômage est aujourd'hui une variable collectée au niveau des États, tous les mois, à partir d'une enquête emploi auprès d'un échantillon représentatif d'habitants. Le recensement complet s'établit, lui, sur un cycle de cinq années en France, tandis que l'État Civil enregistre pour chaque commune de France les mariages, les décès, et les naissances, ce qui permet l'élaboration des statistiques démographiques annuelles.

La compréhension de l'information portée par ces données est capitale pour appréhender le monde et ses évolutions, anticiper l'avenir, et comprendre le passé. Il s'agit de se donner les moyens de représenter et comparer ces statistiques sur les dimensions spatiale, temporelle, et thématique. Ainsi, la prospective est inscrite aujourd'hui dans l'agenda de recherche et développement des SIG, qui doivent remplir les fonctions suivantes :

- saisie des informations géographiques sous forme numérique (Acquisition),
- gestion de base de données (Archivage),
- manipulation et interrogation des données géographiques (Analyse),
- mise en forme et visualisation (Affichage),
- représentation du monde réel (Abstraction),
- prospective (Anticipation).

La prospective territoriale nécessite avant tout d'accroître l'ensemble des connaissances sur les structures territoriales, les tendances et les impacts des politiques dans un territoire donné. Cette connaissance est fondée sur la possession de données (démographiques, économiques, sociales, environnementales...) détaillées sur l'ensemble d'un territoire, et ce, sur une longue période de temps, afin d'aider les scientifiques à identifier et comprendre les tendances, à identifier d'éventuels problèmes et leurs réponses, et à élaborer et tester des scénarios de politique. Aujourd'hui, la richesse des données disponibles à tous les niveaux géographiques nous permet d'espérer créer un outil pour la prospective.

Cependant, le domaine des SIG a émergé à une époque où les données étaient encore rares. Aujourd'hui, avec l'ouverture d'un grand nombre de systèmes d'informations de statistiques territoriales au grand-public, il s'agit de réviser en profondeur les modèles de gestion et d'analyse des données. En effet, cette richesse des données est une arme à double-tranchant : d'une part, elle ouvre la possibilité de pratiquer des analyses multi-niveau, de combiner des données sur des thèmes variés, mais, d'autre part, il apparaît que les supports, les définitions, les modalités de classification, et le niveau de fiabilité de ces données ne sont pas homogènes, ni dans l'espace, ni dans le temps. Cette hétérogénéité des données constitue le cœur de notre problématique. Ainsi, l'agenda de recherche que dresse [Thomas 05] pour l'analyse et la visualisation de données décrit un des défis les plus importants pour la représentation de données à références spatio-temporelles et multi-échelles. Il s'agit de produire une représentation des liens complexes et évolutifs entre ces données. Cette thèse vise l'élaboration d'un cadre général de traitement de l'information statistique territoriale issue de sources multiples. Elle propose des modèles et des outils pour analyser ces données, dans le cadre de la conception d'un système d'information conçu pour l'aide à la décision.

1.1 Problématique

Les prolégomènes ont détaillé les caractéristiques et modalités de constitution de l'information statistique territoriale, dite aussi « socio-économique » [Frank 01]. Nous résumons dans ce paragraphe les principales causes de son hétérogénéité. L'information statistique territoriale est issue de la collecte de données statistiques par des organismes habilités par les États (les producteurs de données) sur des unités zonales. Les méthodes de collecte, leur fréquence dans le temps, et la nature des données collectées varie suivant les producteurs de données. Les données qui sont diffusées sont bien souvent issues de transformations et de processus d'agrégation statistique qui ont pour rôle de protéger le secret statistique, de synthétiser l'information, mais qui biaisent l'interprétation qui peut être faite de cette information [D'Aubigny 94, Openshaw 79, Openshaw 81]. Ce biais est souvent même volontaire, car déjà, la forme du découpage du territoire initial de collecte comme celui de la diffusion des données n'est pas anodin, il est l'expression d'un pouvoir, politique ou scientifique [Grasland 98, Terrier 05]. De la même façon, les modalités d'agrégation thématique (les catégories socio-professionnelles, les pyramides d'âge, etc.) sont très variables, et discutables [Chenu 97, Arel 02, Kieffer 02], et sont le reflet d'une volonté politique sous-jacente.

Le mode de collecte des données rend donc difficile la constitution de collections de données homogènes dans l'espace et régulières dans le temps, éléments qui sont indispensables à une meilleure qualité de l'analyse. La *variabilité sémantique* [Comber 05, Plumejeaud 11] est un problème aussi difficile , que celui du *changement de support* [Gotway-Crawford 05], connu aussi comme le « *split tract problem* », [Howenstine 93], ou problème des recompositions territoriales en français.

Il n'existe pas de système d'information capable de gérer cette hétérogénéité des données. Sur divers plans cependant, la recherche a proposé des solutions pour prendre en compte certains aspects particuliers de cette hétérogénéité de l'information statistique territoriale.

Par exemple, les différents zonages peuvent présenter une forme d'emboîtement, constituant ainsi des *structures multi-niveaux*, et proposant différents niveaux d'observation (le terme « échelle » est souvent employé). Etudier des phénomènes géographiques sur ces différents échelles permet de filtrer l'information, et de mettre à jour des structures spatiales, et des interactions entre niveaux locaux et glo-

baux de l'espace [Marceau 99, Mathian 01]. Ces structures multi-niveaux évoluent elles-aussi dans le temps. Sur le plan informatique, il existe des travaux visant à modéliser ce type de structure de données : [Rigaux 95, Raynal 96, Grasland 05b]. Cependant, ces travaux n'intègrent pas les changements au cours du temps de ces structures multi-niveaux, changements qui soulèvent des questions intéressantes mais difficiles à résoudre.

L'hétérogénéité des sources de données pose la question de la *qualité* des analyses qui peuvent être faites à partir de cette information. La qualité est un terme qui recouvre plusieurs propriétés de l'information, à la fois relatives aux attentes de l'utilisateur vis à vis les données, (c'est la qualité dite « externe »), comme aux spécifications du système qui délivre ces données, (c'est la qualité « interne ») : les traitements et interprétations effectués à partir des données pourront être qualifiés de fiables, précis, à jour, complets, etc., ou l'inverse. Les travaux s'intéressant à la problématique de la qualité dans les systèmes d'information, qu'ils soient géographiques ([Chrisman 84, Devillers 05, Servigne 05]) ou statistiques ([McCarthy 82, UN/ECE 95, Dean 96, Kent 97]) ont établi la nécessité de créer et gérer des métadonnées décrivant les informations collectées dans les systèmes d'information. Il s'agit d'assurer à la fois l'intéropérabilité syntaxique en se conformant aux standards existants, mais également l'intéropérabilité sémantique avec l'usage de vocabulaires contrôlés [Barde 05]. Cependant, dans le domaine de l'information statistique territoriale, l'usage des métadonnées n'est pas encore systématique. Il est notamment très difficile de rendre compte de la qualité des données et de leur lignage d'une façon suffisamment structurée et simple.

Enfin, dans le domaine de l'exploration de données spatiales [Tukey 77, Anselin 93], de la fouille de données [Zeitouni 00, Guo 09], un ensemble d'outils statistiques ont été mis au point, qui permettent notamment de repérer les valeurs exceptionnelles [Rousseeuw 96]. Ces valeurs exceptionnelles peuvent être des erreurs ou bien des valeurs thématiquement intéressantes, à relier au contexte historique et géographique. L'usage de ces méthodes et de ces outils pourrait se révéler particulièrement intéressant pour l'étude de la qualité des données.

1.2 Contribution

Nous donnons ici les grandes lignes de notre contribution, qui se structure en trois propositions, en réponse à la problématique que nous venons d'exposer. Par ailleurs, ces propositions sont constamment illustrées par des exemples issus de l'espace européen et de la statistique socio-économique, démographique et environnementale qui s'y rapporte. Cette thèse ne traite pas de tous les types de données : ainsi, les données de flux se rapportant aux échanges entre des unités territoriales ne sont pas gérées par ce modèle. Ce travail a été mené dans le cadre du projet européen *ESPON 2013 database* qui traite essentiellement de l'information territoriale issue de la NUTS sur l'espace européen, allant des niveaux locaux représentés par les communes aux niveaux nationaux, et qui vise à couvrir une période d'un siècle, entre 1950 et 2050. Ce cas d'étude se retrouve tout au long des propositions qu'il sert à illustrer et à valider.

1.2.1 Un modèle pour des hiérarchies multiples et évolutives.

Cette première proposition a pour cible le support de l'information statistique territoriale. Le modèle que nous proposons s'appuie sur les nombreux travaux menés dans le domaine des SIG sur la datation des supports. Cependant, il élargit les résultats aux supports organisés de façon hiérarchique. Ce modèle

qui est orienté-objet, se base sur un paradigme identitaire, et possède également une visée explicative qui permet de donner du sens aux changements territoriaux et facilite leur analyse. En effet, il intègre la modélisation des évènements historiques et en particulier des évènements ayant un impact sur le territoire, c'est-à-dire ceux qui causent la modification des contours des unités qui composent le support. Nous proposons alors une méthode de définition et de suivi des identités des unités géographiques au cœur du modèle, ainsi qu'une méthode de mise à jour et de maintenance de ce modèle. En effet, il s'avère que la gestion de l'identité des unités géographiques est un point à la fois crucial et délicat, tout comme l'acquisition des évènements dans le modèle. Par ailleurs, une méthode d'analyse interactive de ces changements est proposée, via des cartes de densité du changement, permettant à un expert de l'aménagement du territoire de mettre en relation ces changements avec ses propres connaissances sur le plan politique, économique et social.

1.2.2 Adaptation de la norme ISO 19115 pour l'information statistique territoriale.

La seconde proposition traite du problème de variabilité sémantique des valeurs statistiques associées au support. La première étape indispensable consiste à décrire ces données au moyen de métadonnées. Plusieurs standards sont candidats à leur structuration : SDMX [13], pour *Statistical Data Model eXchange*, ou la norme ISO 19115. Cependant, en pratique, dans le domaine de l'information statistique, ces standards sont mal compris et peu utilisés. Nous proposons donc un profil adapté du standard ISO 19115, facilitant l'acquisition de ces métadonnées aux producteurs de données. Également, nous proposons de créer un système d'information *actif*, au sens où l'entend l'ONU, [UN/ECE 00], c'est-à-dire capable de traiter les métadonnées au même niveau que les données, intégrant les données comme les métadonnées dans un même stockage physique. Enfin, une première étape vers l'interopérabilité avec le standard émergeant SDMX est franchie avec la traduction de notre profil de la norme ISO 19115 vers SDMX.

1.2.3 Exploration et analyse interactive des données.

Alors que les deux premières propositions organisent les données (support et valeurs) de façon à pouvoir exploiter le potentiel de connaissance que l'information statistique territoriale représente, la troisième proposition explore la mise à disposition d'outils (à la fois techniques et conceptuels) pour analyser et explorer dans un mode interactif ces informations. Nous proposons une plate-forme dédiée aux analyses statistiques et visant à repérer des valeurs exceptionnelles (*outliers* en anglais), et à les mettre en relation avec leur origine, et les modalités de leur production. À travers l'interface, l'utilisateur est invité à se questionner sur le contexte de production de la donnée analysée, d'une part en mettant l'évolution de cette donnée en relation avec les changements territoriaux connus, et d'autre part en accédant directement aux métadonnées qui la décrivent. Enfin, par rapport aux cartes d'écarts territoriaux comme proposées dans HyperAtlas [Grasland 05b], qui permettent de repérer des valeurs exceptionnelles, nous montrons l'intérêt que l'intégration d'un modèle spatio-temporel du support tel que celui proposé peut avoir pour l'analyse de l'évolution de ces écarts.

13. `http://sdmx.org/`

1.3 Plan de la thèse

La première partie de cette thèse, consacrée à l'état de l'art, est composée de trois chapitres. Le premier chapitre de cette partie présente les différentes approches existantes pour modéliser des données à références spatiales ou temporelles en géomatique. Le second chapitre présente l'état d'avancement de la description et du traitement de la sémantique de données statistiques. Le troisième chapitre définit plus complètement ce que recouvre la notion de qualité, et décrit les méthodes statistiques de reconstruction de séries temporelles comme de recherche de valeurs exceptionnelles, et les différents types de logiciels qui les mettent en oeuvre.

Dans la deuxième partie de cette thèse nous présentons nos propositions : le premier chapitre décrit un modèle pour l'information spatio-temporelle évolutive basé sur un paradigme identitaire, et indexé par les évènements du changement. Le second chapitre propose des méthodes de description de l'information thématique statistique, qui étendent le précédent modèle avec un ensemble d'informations descriptives, des métadonnées. Le troisième chapitre présente notre proposition pour l'analyse et l'exploration interactive de cet ensemble d'informations hétérogènes dans une plate-forme basée sur notre modèle, intégrant des outils d'analyse statistique.

Enfin, nous concluons cette thèse en résumant les contributions de notre travail à la modélisation et l'analyse de l'information statistique territoriale, qui est une information à références spatiale et temporelle. Nous abordons également les perspectives que cette thèse offre, soit pour la poursuite de travaux sur l'information statistique territoriale, soit dans le cadre plus général de la modélisation spatio-temporelle avec l'adaptation de nos solutions à d'autres problématiques. Nous discutons les limites de celles-ci, et proposons quelques pistes qui seraient à explorer.

Première partie

Etat de l'Art

Chapitre 2

Approches pour la modélisation spatio-temporelle

Ce travail se situe dans le domaine de la géomatique et concerne plus particulièrement la gestion du changement au cours du temps de données à références spatiales. Dans notre problématique, les données sont des statistiques territoriales qui reflètent les évolutions permanentes de la société, sur le plan économique, démographique ou social. Notre recherche vise à produire des méthodes permettant d'analyser ces évolutions ; néanmoins, ces statistiques sont collectées sur des zonages en perpétuelle mutation. La littérature regorge d'exemples et de cas d'étude ([Howenstine 93], [Gregory 02], [Martin 03], [Ben Rebah 08]) montrant à quel point ces zonages changent fréquemment : le problème est connu sous le nom anglais de « *Split Tract Problem* » [Howenstine 93]. L'objectif de ce premier chapitre est de rapporter quels sont les outils et les modèles existants capables de rendre compte et de prendre en compte ces mutations.

Cet état de l'art évalue les divers formalismes et représentations du temps et de l'espace dans les SIG.

La première section est consacrée à la définition du temps et de l'espace et aux formalismes associés en informatique. La section suivante décrit des travaux portant sur la modélisation de l'information spatio-temporelle. Plus particulièrement nous cherchons à mettre en évidence la difficulté que peut représenter la gestion d'une information attachée à des objets géoréférencés qui ne cessent de se transformer au cours du temps. Enfin, la dernière section conclut ce chapitre par une synthèse qui rappelle quels seraient les meilleurs modèles pour répondre à notre problématique et quelles sont encore les difficultés à surmonter.

2.1 Le temps et l'espace

Commençons par présenter les différents objets qui sont à l'étude : le temps et l'espace et éudions les formalismes de représentation associés.

2.1.1 Le temps

2.1.1.1 Définitions

Le temps est un objet complexe pour l'homme comme l'exprime Saint Augustin, (Confessions, XI, 14, 17) :

> Qu'est-ce que en effet que le temps ? Qui saurait en donner avec aisance et brièveté une explication ? ... Si personne ne me pose la question, je le sais ; si quelqu'un pose la question et que je veuille expliquer, je ne sais plus.

L'homme accorde souvent des propriétés au temps qui sont celles des phénomènes qu'il observe au cours du temps [Klein 09]. Par exemple, il envisage le temps comme étant cyclique, parce qu'il observe des phénomènes qui se répètent à intervalles réguliers. En vérité, la physique moderne retient que le temps est linéaire et orienté, pour respecter le principe de causalité, énoncé par Leibniz « tout évènement est l'effet d'une cause qui l'a précédé ». Or, la seule façon de garantir ce principe est de choisir un *temps linéaire* qui protège les évènements du passé de toute modification ultérieure. Le temps s'écoule donc du passé vers le futur : il est *orienté*. Dans cette perspective, deux approches pour la représentation du temps sont utilisées :

- la conception newtonienne, avec un temps absolu qui n'a qu'une dimension, et qui donc est représenté par une courbe à une dimension. C'est un flux mesurable et quantifiable par des dates dont la précision s'exprime en unités qui varient en fonction des usages (siècles, années, mois, jours, heures, minutes, ...).
- la conception de Leibniz qui appréhende le temps par la succession des évènements afin de déterminer des séquences exprimées sur une échelle ordinale (quelque chose se passe *avant*, *pendant*, ou *après* autre chose).

Ces deux conceptions peuvent donner lieu à la production de structures quantitatives (temps mesuré et positionné avec sa coordonnée sur l'axe temporel) et/ou qualitatives (temps représenté par des évènements ordonnés selon leurs positions relatives (*i.e.* leur topologie)) [Thériault 99]. Les deux structures peuvent se combiner, si d'une part une mesure du temps est établie et que, d'autre part, une algèbre définit les relations topologiques entre ces mesures.

Lardon, Libourel et Cheylan font cependant observer que [Lardon 99] :

> L'irréversibilité du temps et « l'unicité de sa ligne » ne sont pas nécessairement vérifiées dans tous les domaines d'applications. Des hypothèses multiples sur le futur comme sur le passé peuvent conduire à des structures de temps embranché. De même, l'étude des cycles naturels, biologiques ou sociaux peut conduire à des représentations alternatives du temps, de forme cyclique.

Il convient donc de garder à l'esprit que la conception linéaire, orientée et unique du temps qui sera employée par la suite correspond à nos besoins, mais n'est pas universelle.

Raisonner sur le temps

Concernant la structure ordinale du temps, les travaux de Allen [Allen 83] fondent une algèbre temporelle permettant de définir des relations topologiques entre objets datés. Dans une vision linéaire continue et non bornée du temps, le temps est structuré en un ensemble d'*intervalles* I_i, un intervalle I étant une paire ordonnée de points (des *instants*). Il munit cet espace temporel d'un ensemble de treize relations binaires et mutuellement exclusives qu'il peut appliquer à ces intervalles (avant, rencontre, égal, chevauche, débute, pendant, termine et leur réciproque), voir figure 2.1. Ces relations permettent de répondre à des questions sur la proximité temporelle de deux phénomènes, à condition d'employer pour les intervalles la même granularité.

Soit x1 et x2 deux intervalles temporels définis par les dates a, b, c, et d

x1 := [a,b] 

x2 := [c,d] 

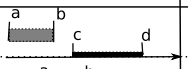
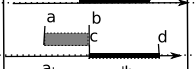
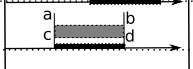
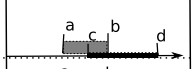
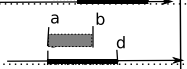
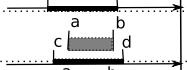
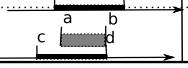
Nom de la relation (anglais / français)		contrainte	schéma	Réciproque (anglais / français)	
before	avant : x1 est avant x2	b<c		after	après
meets	rencontre : x1 rencontre x2	b=c		met-by	rencontré par
equals	égal : x1 est égal à x2	a=c ∧ b=d			
overlaps	chevauche : x1 chevauche x2	d ∧ b		overlapped-by	chevauché par
starts	débute : x1 débute x2	a=c ∧ b <d		started by	débuté par
during	pendant : x1 est pendant x2	c ∧ d		contains	englobe
finishes	termine : x1 termine x2	b=d ∧ c <a		finished-by	termine

FIGURE 2.1 – Relations entre intervalles temporels, d'après [Allen 83].

Dans [Cheylan 99], le temps est défini ainsi :

> Un milieu indéfini où paraissent se dérouler irréversiblement les existences dans leur changement, les événements et les phénomènes dans leur succession.

Donc formaliser le temps c'est aussi définir les changements et les évènements qui se produisent dans l'espace géographique étudié.

[Cheylan 99], précise que le *changement* ne se produit qu'à la condition que, pour une proposition P et des instants distincts t et t', P soit vraie à t, mais fausse à t'. Grossièrement, un changement peut être simplement défini comme une différence significative d'état pour l'objet d'étude (la propriété P de la définition) qui peut être tangible, une forêt par exemple, ou bien intangible (un bassin d'emploi) entre deux moments d'observation distincts. Pour le changement, qui est un terme générique, deux notions se distinguent : celle d'*évolution* ou bien celle de *mutation*. L'évolution, selon [Sanders 99], caractérise dans son sens premier un changement graduel dans le temps, marqué par une suite d'états différents observés. Mutation est utilisée plutôt pour qualifier un changement brusque, où la granularité temporelle de l'observation ne permet pas l'étude des états intermédiaires. Il est à noter également que le terme « mutation » possède un sens un peu différent dans le contexte de la programmation orientée-objet : il signifie alors changement de classe, [Chignoli 97]. La mutation signifie alors qu'un objet typé suivant une classe change de type au cours de son existence. Là aussi la durée de ce changement n'est pas explicitée, ni la description de la phase transitoire, si elle existe.

Dans sa définition classique, [Cheylan 99], un *évènement* est forcément bref à l'échelle temporelle considérée. De même, en UML, la notion d'évènement intervient dans les diagrammes d'Etat-Transition, et un évènement peut être la cause d'une transition qui sera toujours de durée nulle. Cependant, les travaux de Galton, [Galton 04], mettent en évidence qu'un évènement est également un objet, dont l'emprise spatio-temporelle est celle des objets impliqués dans l'événement. Il définit en effet un événement comme un épisode borné dans le temps qui produit des effets remarquables sur un objet, ou un groupe d'objets localisés, ou un endroit du monde. Un évènement est instantané (marqué par un instant) ou dure dans le temps (marqué par un intervalle temporel), mais ceci dépend de l'échelle d'observation. Par exemple, à l'échelle du siècle, une inondation est vue comme un évènement ponctuel, alors que durant les jours de l'inondation, l'inondation est un évènement qui dure. De plus, Galton observe que les évènements peuvent aussi être composés de sous-évènements, de granularité plus fine. Il cite comme exemple la seconde guerre mondiale qui peut être vue aussi comme une succession de batailles. Il introduit l'objet temporel pour généraliser la notion d'évènement (ou processus, qu'il considère comme des synonymes), et ainsi l'histoire du monde se formule alors sous la forme d'une succession d'évènements, ou *chronique*. Formaliser le temps avec des évènements nécessite une algèbre appropriée pour raisonner : Allen et ses collègues, [Allen 84, Allen 94], proposent également une algèbre pour traiter des évènements. Ils généralisent les relations proposées pour les intervalles temporels avec le postulat suivant : les évènements ont une durée non nulle, et donc ils peuvent être modélisés sous la forme d'intervalles temporels.

Mesurer le temps

La mesure du temps présuppose de définir un référentiel et des unités. En effet, le temps est un phénomène continu, mais qui ne peut être enregistré sur un support informatique qu'après échantillonnage, à cause de limitations physiques évidentes. Cette discrétisation du temps obligatoire conduit à introduire la notion d'*intervalle temporel*, qui définit la durée d'observation du phénomène étudié (par exemple, cinquante ans), ainsi que la notion de *résolution temporelle*, correspondant à la fréquence d'observation du phénomène : l'année par exemple dans le cadre des recensements de population. La résolution temporelle définit l'unité élémentaire de temps, ou *chronon*, [Jensen 98], qui va permettre d'évaluer la durée des phénomènes observés en dénombrant le nombre de chronons inclus dans l'intervalle d'observation. Le chronon pourra être la seconde, le jour, l'année, etc, en fonction des besoins de l'utilisateur du modèle. La figure 2.2 illustre, par exemple, comment un même intervalle temporel peut être mesuré suivant différents niveaux de résolution (ou granularité).

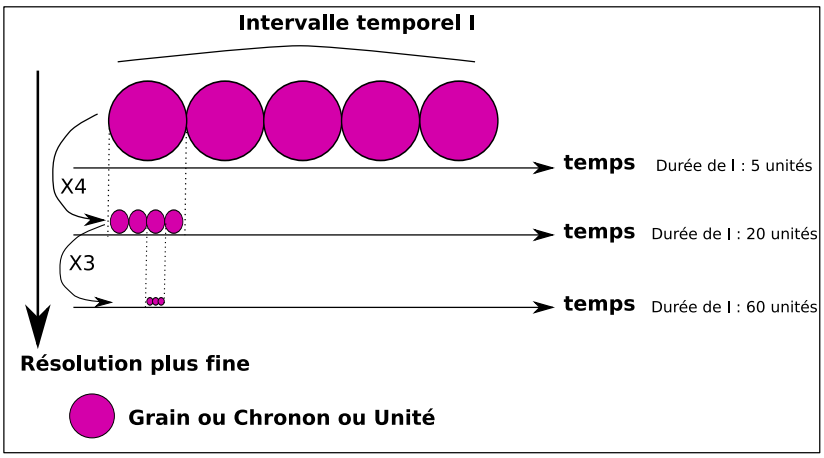


FIGURE 2.2 – Différents niveaux de résolution pour mesurer la durée d'un intervalle temporel I.

La granularité dépend des données, et du niveau de précision utilisé au moment de la capture du phénomène. Selon le contexte d'utilisation de données temporelles, il est parfois nécessaire de convertir les données au même niveau de granularité. Les travaux se rapportant à cette problématique, [Euzenat 93], proposent généralement de dater les entités au niveau le plus fin que l'usage requiert, et d'utiliser ensuite les fonctions de congruence entre les différents niveaux de granularité pour leur conversion. Par exemple, supposons que le système requiert parfois une précision des données en mois, mais que généralement l'année suffise, alors toutes les données peuvent être enregistrées avec le mois comme précision, mais traduites en années, sachant que douze mois constituent une année.

Concrètement, la nécessité d'un référentiel s'est traduit par l'établissement d'une échelle de temps, le Temps Atomique International (TAI). Le TAI se base sur la définition de la seconde, élaborée à l'aide d'horloges atomiques. Il permet de définir l'étalon de temps et l'échelle de temps de référence utilisés partout dans le monde. La seconde a été définie en 1967 lors de la treizième Conférence générale des poids et mesures comme étant la durée de 9 192 631 770 périodes de la radiation correspondant à la transition entre les deux niveaux hyperfins de l'état fondamental de l'atome de césium 133. Le TAI est établi par le Bureau international des poids et mesures et représente la moyenne de la marche de plus de 340 horloges atomiques dans le monde.

Cependant, le temps (la date, l'heure) est rarement exprimé naturellement sous ce format, et l'homme utilise plutôt le Temps Universel Coordonné ou *Coordinated Universal Time* en anglais (UTC) comme échelle de temps. Le temps UTC, adopté comme base du temps civil international par la majorité des pays du globe, se base sur le TAI, mais en diffère par un nombre fini de secondes. L'existence de ces deux échelles s'explique par le fait qu'à l'échelle humaine, le déroulement du temps est perçu depuis l'antiquité (et même avant), via les changements de position du Soleil dans le ciel provoqués à la fois par la rotation de la Terre sur elle-même, et par son parcours elliptique autour de l'astre. Et donc naturellement, l'homme est capable de repérer la succession des jours et des nuits, des saisons, et des années. Une année correspond à une révolution complète de la Terre autour du Soleil. La durée d'une journée est définie par une rotation complète de la Terre sur son axe, environ 24 heures, soit 1440 minutes soit 86400 secondes, mais la vitesse de rotation de la Terre n'étant pas constante, ce temps dit universel (*Universal Time*, UT) n'est ni stable ni exact : la durée des jours UT augmente très lentement en moyenne. Le temps UTC a été créé pour remédier à ce problème : mesuré sur le méridien de Greenwich, il correspond au temps universel, à 0,9 s près. Pour ce faire, UTC est occasionnellement incrémenté ou décrémenté d'une seconde atomique entière, pour faire en sorte que la différence entre UTC et le temps universel UT reste inférieure à 0,9 s, tout en assurant un écart d'un nombre entier de secondes atomiques par rapport au temps atomique TAI.

Enfin, il faut encore préciser que l'horaire mesuré par une horloge solaire, et perçu par l'homme, diffère suivant sa localisation. En effet, lorsque pour les positions situées sur le méridien de Greenwich il est midi, heure UTC, et le soleil est à son zénith, il est minuit, heure UTC pour les positions situés sur le méridien opposé à 180°, c'est la pleine nuit. Donc un autre décalage a été introduit en vue de décrire les activités quotidiennes en fonction d'horaires harmonisés : il fait jour à midi, pour tous, dans toutes les localisations, lorsque l'heure est ajustée sur le fuseau horaire de la localisation. Ce système a été proposé par l'ingénieur et géographe montréalais Sir Sandford Fleming en 1876, avec le méridien de Greenwich comme origine des temps, la ligne de changement de date au méridien 180° (est et ouest), et en divisant le globe en 24 fuseaux horaires de même taille. La zone couverte par un fuseau, limitée par deux méridiens distants de 15°, s'étend du pôle nord au pôle sud ; elle est centrée sur un méridien dont la longitude est multiple de 15°. Le premier fuseau est centré sur le méridien de Greenwich. En pratique, les fuseaux servent à définir l'heure légale dans un Etat, et un découpage raffiné en 43 zones horaires

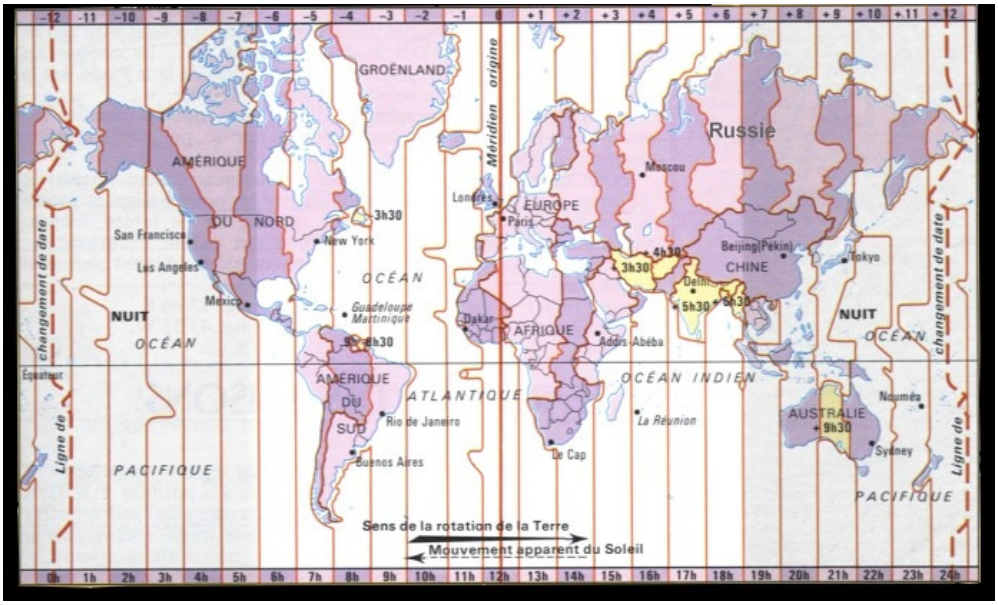


FIGURE 2.3 – Les fuseaux horaires et le temps UTC.

est utilisé. Au passage d'un fuseau à l'autre, l'heure augmente ou diminue d'une unité, (voir figure 2.3). De façon simple, un fuseau horaire peut être écrit sous la forme UTC+X ou UTC-Y, où « X » et « Y » représentent le décalage du fuseau par rapport à UTC.

Ainsi, la mesure du temps peut-être restituée sous la forme d'une date (un instant sur la ligne temporelle) qui exprime la durée écoulée depuis un moment de référence (par exemple, la naissance de Jésus-Christ, supposée il y a 2011 années en arrière, dans le calendrier Grégorien). Il existe un nombre conséquent de calendriers qui adoptent d'autres instants de référence. Par exemple, le temps dans les systèmes informatiques est lui calculé depuis le premier Janvier 1970 : l'origine de cette date vient des normes de l'IEEE qui ont standardisé l'interface de programmation de la famille de système d'exploitation UNIX. Ce compteur est universel et international : il ne compte pas les changements d'heures ni de fuseaux horaires, ce qui est pratique en informatique pour comparer des dates ou pour certains calculs. Il est appelé *timestamp*. Le timestamp est un mot composé venant de l'anglais « *time* » (heure) et « *stamp* » (marquage par un timbre ou un tampon) : le terme désigne donc la notion d' « estampille temporelle ». Ce temps système est généralement fourni avec une précision de l'ordre de la milliseconde (ms).

Les différents points de vue sur le temps

Dans les bases de données, le temps est utilisé suivant plusieurs points de vue - le temps est dit « multidimensionnel », [Jensen 96] - car il peut être :

- le temps de *validité*
- le temps de *transaction*
- le temps *utilisateur*

Le temps de validité correspond à la réalité des faits. Il peut être représenté par un intervalle de validité attaché à une entité du monde réel, dont la signification est la suivante : entre les deux instants qui bornent cet intervalle, l'entité a réellement existé. L'intervalle peut-être défini soit par la date de début de l'intervalle et sa durée, soit par les dates de début et fin.

Le temps de transaction correspond, lui, au temps d'enregistrement des valeurs dans le système informatique. Conserver ce temps est très important pour le versionnement des données. En effet, les données d'un système d'information en général peuvent être sujettes à des révisions. C'est le cas en particulier des publications des variables statistiques dont les instituts statistiques publient des révisions et des réestimations pour une même période de validité. Pour la qualité et la reproductibilité des expériences menées avec les données, il est essentiel de ne pas simplement écraser les données présentes dans le système avec les données mises à jour pour un même temps de validité. Ainsi, les systèmes d'information doivent aussi gérer le temps de transaction. Bien que la prise en compte simultanée du temps de transaction et du temps de validité soit un problème difficile, [Jensen 91], il existe plusieurs les travaux (ceux de Snodgrass, [Snodgrass 92], Claramunt, [Claramunt 95], ou Worboys, [Worboys 98]) qui intègrent la gestion des deux types de temps dans un système d'information, dit *bi-temporel*.

Le temps utilisateur concerne toute autre interprétation que le temps de validité ou des transactions : ce peut être le temps réel pour parler du moment où un phénomène survient, le temps perçu (celui où il est observé), ou le temps d'usage (celui où la donnée est utilisée pour réaliser un traitement), par exemple.

2.1.1.2 Représentation du temps dans les systèmes informatiques

La norme internationale ISO8601, [ISO 04a], développée à partir de 1971 et mise à jour en 2004, rend compte des différents concepts précédemment exposés (la linéarité et la granularité du temps), et permet de représenter le temps par des intervalles ou des instants, de préciser des durées, dans une conception newtonnienne du temps. La représentation du temps dans les systèmes informatiques qui est avancée propose six niveaux de granularité : l'année, le mois, le jour, l'heure, la minute et la seconde, et se base sur le calendrier grégorien. Dans ce calendrier instauré en 1583 après JC, les années contiennent 365 jours, sauf les années bissextiles (366 jours), et 53 semaines. Le patron utilisé pour la notation est le suivant (certains détails ou alternatives sont omis, ces notations sont données à titre indicatif) :

- Un instant s'écrit « yyyy-MM-DDTHH:mm:ss.n-Fuseau » ;
- Une durée s'écrit « **P**x**Y**y**M**z**DT**a**H**b**M**c**S** » ;
- Un intervalle s'écrit « instant/instant » ou « instant/durée » ;
- Une répétition de n occurences séparées par une période (durée) s'écrit « **R**n/durée ». L'instant de la première occurence peut également être signalé : « **R**n/instant/durée ».

Le tableau suivant explicite les symboles employés pour définir le patron :

Voici quelques exemples d'utilisation de cette notation :
- 1977-03-16T22:10:17.25-01:00 signifie 16 mars 1977 à 22 heures 10 minutes 17 secondes et 25 centièmes de secondes, heure décalée de moins une heure par rapport à UTC ou 23 heures 10 minutes 17 secondes et 25 centièmes de secondes, heure UTC.
- P33Y9M3 signifie une durée de 33 ans et 9 mois et 3 jours.
- 1977-03-16T22:10:17.25-01:00/P33Y9M3 décrit un intervalle de 33 ans et 9 mois et 3 jours ayant débutée le 16 mars 1977 à 22 heures 10 minutes 17 secondes et 25 centièmes de secondes, heure décalée de moins une heure par rapport à UTC.
- R33/1977-03-16T22:10:17.25-01:00/P1 décrit un anniversaire répété tous les ans (P1 signifie un an) depuis le 16 mars 1977 à 22 heures 10 minutes 17 secondes et 25 centièmes de secondes, heure décalée de moins une heure par rapport à UTC.

TABLE 2.1 – Signification des éléments du patron de notation défini par la norme ISO8601.

Symbole	Signification	Domaine des valeurs
yyyy	année du calendrier grégorien sur 4 chiffres	1583 à 9999
MM	mois du calendrier grégorien, sur 2 chiffres	01 à 12
DD	quantième du mois, sur 2 chiffres	01 à 31
HH	heure du jour sur 2 chiffres	00 à 23
mm	minutes sur deux chiffres	00 à 59
ss	secondes sur deux chiffres	00 à 59
n	fractions de secondes sur un à plusieurs chiffres	un entier positif ou nul
Fuseau	le fuseau horaire exprime le décalage en heures(hh) et minutes (mm) depuis l'heure UTC par +hh :mm ou -hh :mm. Z signifie sans décalage	Z ou heures entre 00 et 23 et minutes entre 00 et 59
x	nombre d'années	entier positif ou nul
y	nombre de mois	entier positif ou nul
z	nombre de jours	entier positif ou nul
a	nombre d'heures	entier positif ou nul
b	nombre de minutes	entier positif ou nul
c	nombre de secondes	entier positif ou nul

Un aspect important de cette notation est que certains éléments peuvent être omis. La date peut-être précisée sans l'heure (1977-03-16), une durée exprimée en secondes plutôt qu'en heures (P3600S pour dire une heure soit 3600 secondes). Ainsi, l'utilisateur peut choisir le niveau de granularité qui lui convient pour raisonner.

Les avantages de cette normalisation sont les suivants :
- elle est indépendante de toute langue naturelle et compréhensible par un algorithme informatique ;
- elle est facile à comparer et à trier (en gardant un format fixe dans un contexte donné) car le tri par ordre lexicographique correspond à l'ordre chronologique ;
- elle présente peu de risques de confusion avec d'autres notations ;
- elle est concise et de taille constante ;
- elle permet une compréhension intuitive des éléments de date et d'heure.

Cette norme est à la base des représentations employées dans l'ontologie OWL-Time de Pan et Hobbs, [Pan 04]. OWL-Time est définie comme une ontologie d'objets temporels, qui peut servir à dater des faits, pour leur date de validité ou bien de transaction. OWL-Time, dont nous fournissons une représentation de la structure conceptuelle sous la forme d'un diagramme de classe, figure 2.4, propose au concepteur de décrire par exemple un intervalle de durée non nulle par le type *DateTimeInterval*, composé d'au minimum deux instants, décrits par le type *DateTimeDescription*. Ce modèle offre la possibilité de moduler la granularité du temps au moment de son utilisation, puisque le niveau de granularité est contrôlé par l'attribut *unitType*, au niveau de la définition des *Instants* qui constituent l'intervalle. La date qui définit les instants peut être décrite avec une très haute résolution, mais l'utilisateur a la possibilité de choisir la granularité du temps qui l'intéresse en faisant varier l'attribut (*TemporalUnit*) par exemple.

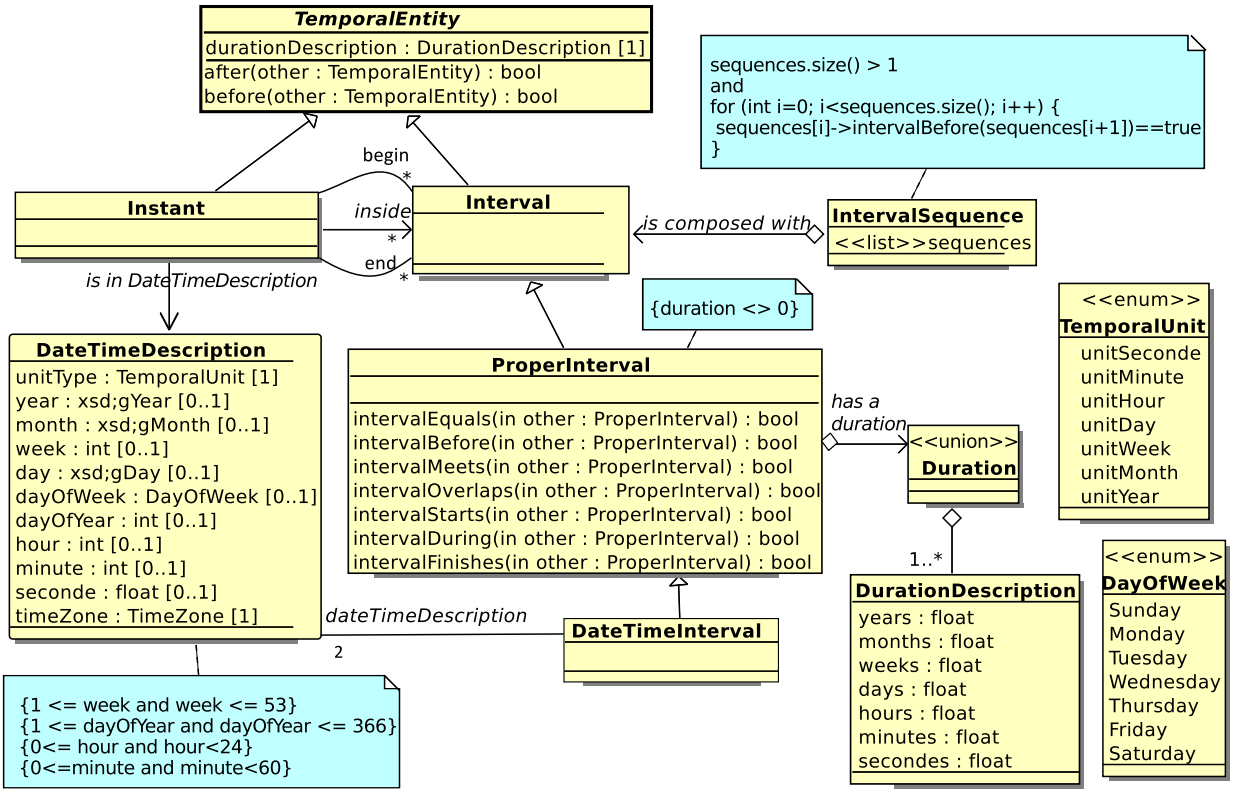


FIGURE 2.4 – Schéma des entités temporelles dérivées du modèle OWL-Time, d'après [Pan 04].

Les algèbres de Allen [Allen 83, Allen 84, Allen 94] ont déjà été évoquées pour leur capacité à raisonner sur le temps. Une revue plus complète des algèbres est dressée par [McKenzie 91], mais nous retiendrons essentiellement les travaux de Allen pour notre problématique.

2.1.2 L'espace

Cette section présente le vocabulaire relatif à l'espace, compris comme étant l'*espace géographique*, et introduit les principaux concepts et modèles informatiques permettant de produire une représentation de l'espace géographique dans un système d'information géographique.

2.1.2.1 Définitions

Dans [Pumain 97], l'*espace géographique* est défini comme l'ensemble de lieux et de relations entre les lieux, définies par les interactions entre des acteurs sociaux localisés. Il est le produit de l'organisation des sociétés et de la nature, et l'agent du maintien et du développement des sociétés sur leur territoire.

Comme pour l'espace de la physique, il y a deux manières d'envisager l'espace géographique pour le concevoir et le représenter :

– soit comme un simple contenant, repère immuable dans lequel on situe les objets et analyse leurs relations : c'est l'*espace-support*, ou espace absolu. Il est isotrope et homogène, c'est-à-dire qu'il

a les mêmes propriétés dans toutes les directions ;
– soit comme un ensemble de relations dont les propriétés, variables dans le temps et dans l'espace, sont définies par la nature et la forme des interactions entre les objets et les unités spatiales : c'est l'espace relatif ou *espace-produit*.

Dans le cadre de cette étude, les zonages qui sont utilisés pour la collecte d'information statistiques, s'ils étaient immuables, constitueraient l'espace-support de l'information. C'est là une conception commune lors d'étude de l'évolution de variables statistiques, quelque soit leur nature. Ainsi, Google, ([Google 10]), qui a mis en ligne un nouveau site (*Google Public Data Explorer*) permettant de sélectionner des données publiques issues de plusieurs organismes comme la Banque mondiale, l'OCDE, Eurostat, etc., propose quatre types de visualisation des indicateurs statistiques, dynamiques, avec une ligne de temps permet de faire évoluer les données, mais toujours basée sur un support spatial stable dans le temps.

Cependant, les zonages sont sujets à recompositions. En outre, ils reflètent une organisation humaine de l'espace en constante évolution, mais aussi *anisotrope* par nature. L'espace est marqué par des similarités, des pôles d'attractivité et des zones de discontinuités (culturelle, linguistique, économique) [Grasland 02]. Ce problème est développé par Brunet, [Brunet 97]. L'objectif de l'analyse spatiale, qui est définit dans [Pumain 97] comme « l'analyse formalisée de la configuration et des propriétés de l'espace géographique, tel qu'il est produit et vécu par les sociétés humaines » est justement d'étudier cette anisotropie. Cette anisotropie peut aussi être expliquée par les sciences sociales du comportement, qui décrivent les processus de compétition ou communication entre les humains, processus qui s'inscrivent à la fois dans le temps et l'espace. Ainsi, dans une tentative de définition d'un nouveau paradigme pour l'organisation spatiale des sociétés, [Villeneuve 72], Villeneuve rapproche la géographie du comportement de l'analyse spatiale à travers le concept de *proximité optimale*, qui s'explique comme la réponse au principe général de survie et aux contraintes que l'espace invariant définit. Les décisions individuelles ou collectives seraient un compromis entre les tensions centrifuges (les processus de compétition, au niveau biotique) et les tensions centripètes (les processus de communication, au niveau culturel) qui engendrent le principe fondamental d'organisation des formes spatiales. Lorsque l'étude des formes spatiales porte sur les zonages, il apparaît qu'en réalité leur forme est étroitement liée au peuplement, et qu'ils matérialisent les frontières et les zones de discontinuités (linguistique, culturelle, économique) que l'occupation humaine produit. Cette observation s'applique aussi dans le cas de l'étude des *maillages*, [Grasland 98], qui sont une forme particulière de zonage, définissant une partition complète et totale de l'espace. Les zonages ne sont donc en aucun cas une représentation de l'espace isotrope ou homogène. Nous envisageons donc ici l'espace comme un espace-produit.

Cet espace-produit est lui-même modélisé et analysé selon plusieurs niveaux de granularité : le terme *échelle spatiale* est employé. Contrairement au temps, où l'unité élémentaire a pu être établie (la seconde), il n'existe pas, pour l'espace, de grain élémentaire consensuel ou naturel, [Lardon 99]. L'échelle au sens premier se définit comme un ensemble ordonné de grandeurs. L'échelle spatiale est donc un ensemble ordonné de grandeurs spatiales. En parlant d'échelle, il est toujours fait référence à l'échelle de résolution spatiale. Cependant, celle-ci peut être comprise comme l'*échelle de représentation*, ou bien *le niveau d'observation* et d'analyse. Pour l'échelle de représentation, il faut utiliser un nombre qui exprime le rapport entre une distance sur la représentation et une distance réelle, dans la même unité de distance. Ainsi, dire qu'une carte est au 25 000 millième signifie qu'un centimètre sur la carte représente 25 000 centimètres dans la réalité. Par contre, l'échelle d'observation (ou niveaux d'analyse) sous-tend l'usage d'une structuration de l'espace en niveaux hiérarchiques [1] :

1. Ces niveaux sont ceux couramment compris par les géographes en employant les termes macro, méso ou micro. Mais dans l'absolu, ce n'est pas l'échelle de référence, et selon le contexte de l'étude, le canton sera par exemple au niveau macro.

– le niveau micro : la maison, la rue, le quartier
– le niveau méso : le canton, le pays, la région, l'état
– le niveau macro : cadre national, supranational (comme l'Europe) et mondial

Toute portion de l'espace terrestre peut donc être observée et étudiée à différents niveaux de détails. Claval, [Claval 68], a montré que combiner les échelles favorise notre compréhension de l'espace terrestre ou zonal. Toute valeur en un point donné est fonction de processus agissant à différents niveaux, locaux, régionaux et internationaux (par exemple, la localisation d'une entreprise industrielle). La compréhension d'une partie de l'espace terrestre passe obligatoirement par l'examen de niveaux différents et imbriqués. C'est l'objet de la théorie formelle des systèmes hiérarchiques qui postule qu'un espace géographique forme un système dont les composantes sont reliées par des relations dissymétriques donc non égalitaires, [Takahara 80]. Cette théorie est largement développée, par exemple dans le domaine des systèmes d'information, elle amène à produire une formalisation des partitions de l'espace, très utile pour modéliser les hiérarchies spatiales, [Rigaux 95]. Ce formalisme s'applique lorsque les niveaux s'imbriquent pour former une hiérarchie stricte, c'est-à-dire lorsque toute unité (exceptée la racine) possède une et une seule unité supérieure. Ce formalisme offre une formalisation des opérations d'agrégation sur les attributs attachés aux zonages. En effet, lorsque les attributs sont des effectifs dénombrables, comme par exemple le décompte de la population ou de logements, ils ont la propriété de pouvoir s'additionner pour calculer l'effectif de la maille supérieure. Ainsi, la population d'un département est, en théorie [2], la somme des populations des communes qui lui sont rattachées. Il apparaît que la NUTS (Nomenclature des Unités Territoriales Statistiques), établie par l'Union Européenne, depuis 2003, [Parlement européen 03], utilisée par les instituts statistiques nationaux ou européens (Eurostat), est effectivement une hiérarchie spatiale stricte, pour laquelle ce type de formalisme est très utile, notamment pour reconstituer par agrégation les valeurs d'attributs à des niveaux supérieurs à celui de leur mesure effective.

Cette théorie donne également lieu à des études où les niveaux ne sont pas nécessairement strictement emboîtés, ni ne forment une partition complète et totale de l'espace. Mathian et Piron, [Mathian 01], citent notamment l'exemple des secteurs de caisses d'allocations familiales, ou des chambres de commerce. Nous pourrions également citer le cas des intercommunalités, sur lesquels nous avons travaillé en partenariat avec Guillaume Vergnaud, [Plumejeaud 09a]. Par rapport à cette problématique, Mathian et Piron, [Mathian 01], proposent notamment un formalisme mathématique adéquat pour la représentation de l'appartenance d'une unité territoriale à des classes, sous forme de graphes de partition ou bien de proximité, et facilitant l'usage de méthodes statistiques multidimensionnelles.

Notre approche pour l'étude de l'espace (en tant qu'espace-produit) et de ses évolutions sera également celle de l'analyse spatiale, qui postule que les caractéristiques d'un lieu dépendent des relations de *proximité* de ce lieu par rapport à d'autres lieux, [Pumain 97]. La proximité est évaluée par la *distance*, qui est une notion géographique fondamentale. Les travaux de nombreux géographes mettent en exergue le rôle que la distance peut avoir sur la compréhension de la spatialisation des phénomènes sociaux, des échanges et des flux [Grasland 10a]. Ces études dérivent directement de la première loi de géographie édictée par Tobler, en 1970, [Tobler 70] : « *Everything is related to everything else, but near things are more related than distant things.* » qui peut être traduit en français par : « Tout interagit avec tout, mais deux objets proches (spatialement parlant) ont plus de chance de le faire que deux objets éloignés. ». Cette notion d'espacement ne respecte pas forcément les propriétés d'une *distance mathématique*.

2. En réalité, pour des niveaux de recensement différents, le nombre d'individus dénombrés sur une même aire peut diverger. Par exemple, les populations sans domicile fixe peuvent être localisées au niveau supérieur (le département ou l'état), sans avoir été décomptées dans les ménages des communes.

En effet, la distance mathématique entre deux points A et B est une mesure toujours positive, $d(A, B) > 0$, qui :

- est nulle seulement si A est confondu avec B : $d(A, B) = 0 \Leftrightarrow A = B$;
- vérifie l'inégalité triangulaire entre 3 points A, B, C : $d(A, C) \leq d(A, B) + d(B, C)$;
- est symétrique : $d(A, B) = d(B, A)$.

Or, les relations de distance entre unités géographique ne sont pas forcément symétriques, et par exemple, le temps mis par un routier pour atteindre Grenoble en partant de Gênes n'est pas identique au temps de retour. Ceci tient au fait que l'espace géographique n'est pas isotrope. En fait, en géographie, ce sont des *mesures d'éloignement* que, par commodité, on continue d'appeler distance, [Pumain 97]. La contiguïté est un cas particulier de cette mesure d'éloignement qui peut être utilisé dans le cadre d'un espace maillé : entre deux unités A et B, la mesure vaut 1 (vrai) si elles se touchent, 0 (faux) sinon.

2.1.2.2 Représentation quantitative de l'espace dans les systèmes informatiques

Concrètement, pour représenter l'espace, il faut considérer que les objets du monde, qu'ils soient tangibles (forêt, ville, rivière) ou intangibles (frontière d'un pays, centre du monde, l'Atlantide) ont une position (ou empreinte spatiale) géographique. La représentation de l'espace fournit des modèles pour l'enregistrement des positions de ces objets, qu'elles soient *qualitatives* comme dans « le restaurant est *à gauche* de la bibliothèque, *non loin* de l'arrêt de tram » ou bien *quantitatives* comme dans « Paris est situé à 48° 49'N, 2° 19'E sur la surface du globe ».

Ces modèles de représentation de l'information spatiale peuvent être classés selon trois niveaux : le *niveau géométrique*, le *niveau informatique*, et le *niveau utilisateur*, [Clementini 08].

- Au niveau géométrique, les objets spatiaux sont modélisés comme des objets mathématiques (points, courbes, surfaces) dont les relations topologiques peuvent être explicitement formalisées : les objets surfaciques se touchent ou bien ne se touchent pas par exemple, ou un point est dans une surface, sur la frontière de la surface, ou en dehors de cette surface. Le niveau géométrique peut être considéré comme le niveau le plus primitif pour l'étude des relations spatiales, puisqu'il permet de retrouver des définitions formelles, et peut être considéré comme sans erreur [Clementini 08].
- Au niveau informatique, les objets spatiaux sont représentés comme des types de données spatiaux et les relations spatiales entre les objets sont, en général, calculées par des opérateurs spatiaux. La représentation des objets géographiques à ce niveau est intrinsèquement concernée par l'approximation car les objets réels sont représentés avec un modèle simplifié, et il peut exister plusieurs descriptions du même objet selon des niveaux de précision différents [Ruas 04]. Le degré d'incertitude qui en découle est une question de recherche en cours, loin d'être résolue [Clementini 08]. Par exemple, certaines approches proposent de modéliser l'incertitude dans la frontière des objets comme une bande bidimensionnelle autour de l'intérieur des objets [Cohn 96], [Clementini 01], afin de continuer d'utiliser les mêmes modèles pour les relations topologiques définies au niveau géométrique.
- Au niveau utilisateur, les objets et les relations spatiales sont liés à un contexte spécifique d'application, et décrivent souvent l'espace via des termes flous qui varient énormément selon les différents pays et langages. Par exemple, "le restaurant est *à gauche* de la bibliothèque, *non loin* de l'arrêt de tram" est une représentation de l'espace du niveau utilisateur. Le terme « non loin » renvoie à l'appréciation que l'utilisateur a de la distance : des mètres, des dizaines de mètres, des centaines de mètres, etc. Le terme « à gauche de » réfère à un positionnement relatif dans un certain contexte, qui n'est pas toujours spécifié. C'est pourquoi il est difficile de transférer les relations spatiales définies au niveau utilisateur au niveau géométrique. Il existe des travaux visant à définir

les concepts spatiaux à l'aide d'ontologies spatiales [Spaccapietra 04], [Miron 09], mais cela reste encore un domaine ouvert de la recherche.

Du point de vue de l'analyse spatiale, c'est essentiellement les deux premiers niveaux qui sont utiles et permettent de construire des méthodes mathématiques de raisonnement quantitatif sur les processus spatiaux. C'est le domaine de la géographie dite « quantitative », qui a émergé en France à partir des années 1970, en bâtissant ses raisonnements à l'aide d'outils mathématiques [Pumain 02]. Les représentations quantitatives permettent de réaliser des calculs efficaces et robustes impliquant les géométries attachées aux objets modélisés. Elles se montrent toutefois moins adaptées lorsque les données géométriques sont incomplètes ou imprécises.

Le raisonnement *qualitatif* peut être proposé comme alternative au raisonnement quantitatif, car il permet d'opérer des déductions lorsque les connaissances sur les objets utilisés ne sont pas complètes, ou lorsque les détails sont secondaires.

Ce type de raisonnement vise à diviser l'ensemble des valeurs possibles pour une variable donnée, en classes d'équivalence (qui sont des espaces symboliques) regroupant des valeurs proches qui seront considérées comme équivalentes. Alors qu'avec un raisonnement quantitatif il faut considérer l'ensemble de toutes les valeurs possibles, le qualitatif propose un nombre de distinctions adapté à la tâche visée, simplifiant ainsi les calculs et la comparaison entre valeurs inexactes : déterminer la classe d'équivalence de la valeur inexacte ou incertaine suffit pour le raisonnement qualitatif. Freska explique que les représentations qualitatives pour les systèmes cognitifs sont indépendantes de valeurs spécifiques et de granularités de représentation ; en fonction de la situation, du contexte et de la granularité des connaissances disponibles, elles correspondent à des entités plus spécifiques ou plus généralisées [Freska 91]. Yumi Iwasaki donne ainsi l'exemple d'une pluie qui ne cesse de tomber sur les eaux d'une rivière dont le niveau monte, [Iwasaki 97]. Pour l'humain ou le système expert qui se base sur un raisonnement qualitatif, il est inutile de connaitre le niveau exact de la rivière, ni la vitesse d'élévation du niveau pour anticiper qu'il va y avoir crue bientôt, sans toutefois en connaitre le moment exact. Dans cet exemple, le niveau est mesuré approximativement, à l'aide de catégories (bas, normal, haut) tout comme sa vitesse de progression (lent, rapide), et en raisonnant à partir de ces catégories, il est alors possible de se passer des valeurs exactes du phénomène observé. Plus spécifiquement, une revue récente des problématiques, outils et formalismes consacrés au raisonnement qualitatif sur le temps et l'espace est proposée dans [Ligozat 10].

Le raisonnement qualitatif repose sur une relation d'ordre (partielle ou totale) qui est établie pour chaque espace symbolique, et dont la transitivité peut être exploitée afin de réaliser des inférences. Cependant, [Forbus 84] souligne que la transitivité ne peut pas être bien exploitée dans des espaces à plusieurs dimensions. Egalement, selon [Cohn 01], l'espace étant, par définition, multidimensionnel, il ne peut pas être modélisé de façon adaptée en utilisant une seule quantité scalaire. Même si une telle représentation est possible, l'absence de mécanismes de raisonnement spatial purement qualitatifs fait que l'exploitation de ces modèles est extrêmement difficile.

Dans cette section, la description des modèles de représentation de l'information spatiale est donc essentiellement consacrée à la description quantitative de l'espace, sur les niveaux géométriques et informatique.

2.1.2.2.1 Capturer une position Généralement, les lieux géographiques à la surface de la Terre sont repérés par un couple de nombres réels (latitude, longitude). La figure 2.5 montre comment la latitude et la longitude sont calculées, en considérant que la Terre est sphérique [3].Tous les lieux d'un même parallèle à l'équateur ont la même latitude. La latitude des lieux situés sur l'équateur est 0° (zéro degré). La latitude des pôles est 90° Nord pour le pôle Nord et 90° Sud (ou -90°) pour le pôle Sud. Tous les lieux situés sur un même méridien ont la même longitude. Le méridien de référence passe par Greenwich : sa longitude est 0°. Tous les autres méridiens sont mesurés en prenant ce méridien comme origine, avec une notation négative de l'angle en allant vers l'ouest par rapport au méridien de référence.

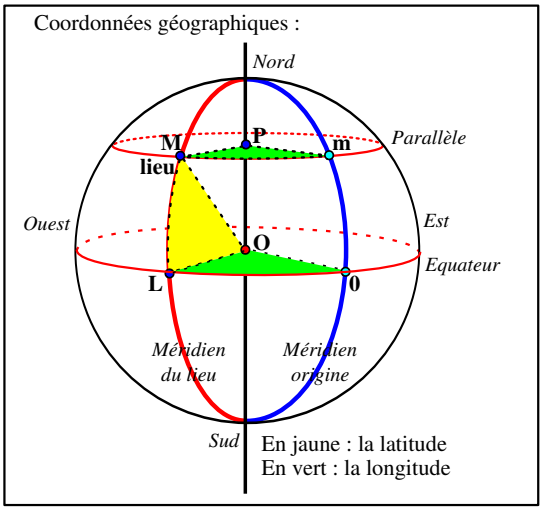


FIGURE 2.5 – Système de coordonnées géographiques pour localiser un lieu.

Les systèmes d'information géographiques peuvent stocker ces données « brutes ». Cependant, l'objectif est d'en donner une représentation planaire dans un repère cartésien orthonormé (la carte papier, ou l'écran de l'ordinateur). Aussi, ces données brutes sont le plus souvent projetées (à l'aide d'un système de référence) sur un plan, auquel est associé un repère orthonormé. La projection est un problème à la fois mathématique et de représentation, car suivant le mode de projection utilisé, certaines surfaces seront plus ou moins déformées (dans leur forme ou leur distance relative). Ainsi, en France, la projection la plus usuelle est celle de Lambert93. Il existe de nombreux modes de projection, dont certains sont référencés et manipulés aisément dans les SIG, de façon standardisée. Ainsi, par exemple, l'EPSG (European Petroleum Survey Group, créé en 1985) propose une base de données mondiale ouverte à tous des systèmes de coordonnées géoréférencés de projection, les codes EPSG, [EPSG 85], qui sont utilisés aussi bien par le groupe de producteurs de pétrole EPSG que par de nombreux logiciels de SIG. Ces codes, (environ 3750), sont notamment utilisés dans les standards de l'*Open Geospatial Consortium* (OGC) [4], et peuvent facilement être importés dans une table dédiée d'un Système de Gestion de Base de Données (SGBD) relationnel (par exemple, la table (*Spatial_ref_sys* pour postgreSQL avec sa cartouche spatiale PostGIS). De même, l'IGN propose une extension de cette base de données avec des codes spécifiques pour le territoire français, [IGN, France 10]. Via cette table, et des fonctions de conversion intégrés à des bibliothèques [5], il est aisé de réaliser les opérations de conversion d'un système de projection à l'autre.

3. Ce n'est pas tout à fait le cas : la Terre est un ellipsoïde en constante déformation sous l'effet de sa rotation sur elle-même, ainsi que de l'attraction que la Lune exerce sur elle.

4. L'OGC est un consortium créé en 1994 qui regroupe les principaux acteurs de la communauté géomatique (plus de 370 membres), en vue d'élaborer des standards favorisant l'interopérabilité entre les SIG.

5. Par exemple, la bibliothèque proj.4 disponible sur `http://trac.osgeo.org/proj/` sous licence MIT, codée en C - ou GeoTools disponible sur `http://geotools.org/` sous licence GNU *Lesser General Public License* (LGPL), codée

Il faut retenir, qu'en réalité, à toute représentation spatiale d'un objet dans un espace cartésien, il est nécessaire d'associer le système de coordonnées géoréférencés de projection utilisé.

2.1.2.2.2 Mesurer des distances L'espace est donc le plus souvent modélisé sous la forme d'un espace euclidien vectoriel, sous-ensemble de $\mathbb{R}^2$ (ou $\mathbb{R}^3$, si l'altitude est prise en compte), dans les SIG. Dans le cadre de notre problématique, nous nous limitons à un espace à 2 dimensions, dont la surface est supposée plane [6]. Dans les SIG, cet *espace euclidien planaire* est généralement muni de la distance euclidienne ayant les propriétés mathématiques précédemment décrites. L'unité étalon est le mètre (symbole m, du grec *metron*, mesure), défini par le Bureau International des Poids et Mesure, depuis 1983, [BIPM 83], comme « la distance parcourue par la lumière dans le vide en 1/299 792 458 secondes ». Dans le cadre d'une représentation des coordonnées géographiques dans un repère cartésien, la distance euclidienne est la plus évidente pour calculer la distance entre deux points à la surface de la Terre. Dans un espace euclidien orthonormé de dimension 2, la distance d_{AB} entre deux points A et B de coordonnées respectives (x_A, y_A) et (x_B, y_B) est donnée par la formule 2.1 :

$$d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \tag{2.1}$$

Cependant, il convient de noter que, du fait des déformations induites par les différents modes de projection, et de la rotondité de la Terre, ces mesures de distance sont fausses dès lors que les points mesurés sont très écartés à la surface de la Terre (par exemple, d'une extrémité de l'Europe à l'autre). La distance orthodromique, qui se base sur des coordonnées géographiques non projetées permet elle de donner une mesure de la distance « à vol d'oiseau » juste. En posant R comme étant le rayon terrestre (soit 6378 kilomètres, rayon équatorial moyen), la distance d_{AB} entre deux points A et B de coordonnées géographiques (la longitude et la latitude sont exprimées en radian) respectives $(long_A, lat_A)$ et $(long_B, lat_B)$ est donnée par la formule 2.2 :

$$d_{AB} = R * \arccos(\cos(lat_A) * \cos(lat_B) * \cos(long_A - long_B) + \sin(lat_A) * \sin(lat_B)) \tag{2.2}$$

Communément, on dira donc que Paris et Tokyo sont distantes de 9714.482 kilomètres « à vol d'oiseau » (par avion). L'emploi de la distance orthodromique implique l'usage de fonctions mathématiques coûteuses comme arcosinus, et c'est pourquoi les systèmes d'information géographiques privilégient souvent l'emploi de la distance euclidienne pour des études dont l'emprise spatiale ne couvre pas un trop grand espace géographique.

La question de la distance, lorsqu'elle est interprétée comme une mesure d'éloignement et non plus seulement comme une distance mathématique, reste cependant épineuse, car, comme il a été dit précédemment, la distance euclidienne ou orthodromique ne tiennent absolument pas compte de la réalité géographique du terrain, qui est anisotrope. Le franchissement de rivières ou de montagnes, par exemple, peut impliquer pour les humains des détours sur un chemin reliant une origine A et une destination B, dont ces distances ne tiennent pas compte dans leur mesure de la longueur du segment $[AB]$. La distance-temps, la distance perçue, l'accessibilité, sont d'autres formes de mesure de la proximité de deux lieux. Les distances mesurées en temps de parcours, ou en coûts, sont souvent bien plus expressives des situations géographiques, du point de vue de leur signification sociale [Pumain 02]. Ces nouvelles distances

en Java.

6. Le relief est pris en compte dans l'analyse à travers les différentes mesures de l'éloignement, mais la visualisation classique des données ne donne à voir qu'un espace sans relief, sans représentation de la topographie du terrain.

sont utilisées pour calculer des projections des lieux, qui, comparées aux projections topographiques usuelles, font ressortir par des « déformations », les couloirs privilégiés par la grande vitesse, en quelque sorte rétrécis, ou au contraire les zones enclavées qui apparaissent dilatées. Les transformations cartographiques, par anamorphose ou par d'autres procédés faisant appel à des géométries complexes, sont ainsi très employées en géographie des transports ou pour des cartes cognitives [Cauvin 97]. Il faut remarquer que ces mesures de proximité sont variables dans le temps : par exemple, les distance-temps entre un lieu situé en périphérie urbaine et le centre urbain d'une même ville sont soumises à des variations horaires : aux heures de transfert entre domicile et travail de la majorité des actifs, un ralentissement du traffic urbain rallonge la distance perçue entre les lieux.

Les proximités sont alors modélisées sous la forme d'un ensemble de relations entre lieux, qui peuvent être d'ordre topologique, avec la *contiguïté*, ou faire appel à des notions de mises en relation par réseau, avec la *connexité*. La contiguïté définit une proximité par l'adjacence des géométries représentant ces lieux. La contiguïté est une relation de simple proximité, dite en continuité, et qui relève de la topologie, alors que les relations de connexité sont celles qui utilisent le support d'un réseau pour joindre deux lieux qui peuvent être très éloignés. Dans la théorie des graphes, la connexité est l'intensité de la mise en relation des noeuds par les arêtes d'un réseau ; ou plus généralement, le degré de connexion interne d'un réseau [7]. La connexité peut représenter l'existence de liens commerciaux entre deux unités. Ainsi, la proximité spatiale de lieux géographiques peut être exprimée en tenant compte des réseaux de transport existants, (routes, voies ferrées, voies maritimes, voies aériennes, voies hertziennes, voies numériques), et des moyens de transport disponibles pour les entités susceptibles d'être transférées d'un endroit à l'autre. Les entités peuvent être aussi bien des êtres vivants (humains, animaux), que des biens marchands (blé, tissus), ou des biens immatériels (comme l'information par exemple).

Il apparaît toutefois que la modélisation, sur le plan informatique, de données représentant des proximités spatiales est assez complexe. En effet, à moins de disposer des graphes modélisant les différents réseaux de transport, ces distances ne peuvent être calculées par des formules, et il faut alors les conserver dans le système d'information. Les matrices de distance en sont le principal artefact. Elles sont une généralisation de la matrice de contiguïté. Bien que les relations de contiguïté puissent être calculées à la volée, le coût du calcul des relations topologiques conduit à l'emploi de matrices de contiguïté permettant d'enregistrer la contiguïté entre deux unités territoriales d'un espace donné : si une unité i partage une frontière commune avec une autre unité j, alors la cellule (i, j) contient la valeur booléenne vrai, et cette matrice est normalement symétrique. Les matrices de distance remplacent la valeur booléenne d'une cellule (i, j) par une distance (dont l'unité peut être le temps, mais pas obligatoirement) : si la cellule (i, j) vaut 25, alors que le lieu i est distant de 25 unités de j, mais réciproquement, j peut être considéré à 30 unités de i, et alors la cellule (j, i) vaudrait 30. Les matrices de distances ne sont généralement pas symétriques. En effet, le temps de transport entre deux villes diffère dans un sens ou l'autre car le réseau de transport n'est pas forcément symétrique : par route, par exemple, des péages ou des voies d'accélération peuvent avoir été installées dans un sens mais pas dans l'autre.

Outre la distance géographique physique, la proximité spatiale pourrait être calculée à partir de critères sociaux de ressemblance, comme l'argumente François, [François 02]. Il utilise l'exemple de la carte scolaire, pour montrer que la notion de ressemblance sociale est fortement corrélée à celle de proximité spatiale. Nous observons que la relation hiérarchique est aussi finalement un cas particulier de la mesure d'éloignement, puisque deux unités territoriales appartenant à la même unité territoriale

7. Source : `http://www.hypergeo.eu/spip.php?article50` et `http://www.hypergeo.eu/article.php3?id_article=400` - connexité est moins ambiguë que le terme « connectivité » qui comprend deux définitions contradictoires, l'une en théorie des graphes et l'autre en géométrie.

sont régies par des lois et des règles qui leur appliquent des contraintes de ressemblance, donc de proximité. Cependant, les outils pour définir des proximités spatiales en fonction de ressemblance sociale (ou autre) ne sont pas encore utilisés de façon systématique dans les SIG. Miller, [Miller 00], fait lui aussi remarquer qu'il existe quantités d'autres espaces dits « pré-géographiques » qui seraient plus appropriés pour l'étude de phénomènes géographiques, et il s'appuie en cela sur les travaux de Cliff, [Cliff 98] sur la mesure de l'auto-corrélation spatiale, mais il observe que finalement, la majorité des travaux en analyse spatiale ont été menés dans le cadre restrictif de l'espace euclidien planaire muni d'une distance euclidienne. Sur le plan informatique, les travaux qui s'appuient sur les méthodes développées en analyse spatiale pour faire de la fouille de données, comme ceux de Ester, [Ester 97], ou Zeitouni, [Zeitouni 00, Zeitouni 01], reprennent d'ailleurs essentiellement, soit la notion de distance euclidienne, soit la notion de connexité pour constituer des graphes dits de « voisinage » où les arcs reliant les noeuds du graphes (les lieux géographiques) peuvent être valués en fonction de différentes relations (de distance, ou de contiguïté).

2.1.2.2.3 Représenter des formes L'espace étant muni d'une métrique mesurant la distance entre les différentes entités qui le peuplent, et leur position réelle étant connue, il reste à expliquer comment elles sont représentées dans un système informatique. Deux modes de représentation de l'information spatiale se distinguent : le mode matriciel, ou *« raster »*, et le mode vectoriel. La figure 2.6 montre le type d'image produite par les deux types de représentation.

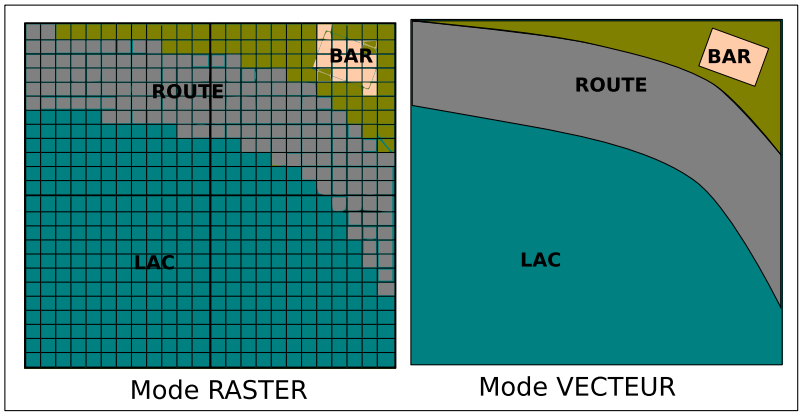


FIGURE 2.6 – Système de coordonnées géographiques pour localiser un lieu.

En mode « *raster* », une surface est représentée par un ensemble de cellules, également appelées pixels, dotées d'une couleur et une luminosité donnée. Chaque pixel de coordonnées (i, j) représente une surface, à laquelle des attributs thématiques sont rattachés (population, PIB, ou type d'occupation du sol). Le modèle matriciel sert à représenter aussi bien des variables qualitatives (utilisation du sol, type de forêt, etc.) que quantitatives (réflectance, pente, drainage, etc.). Avec ce type de représentation, la résolution dépend de la taille de la cellule et la topologie est toujours représentée implicitement. L'avantage est que les opérations sur les surfaces sont aisées (calcul d'aire, superposition de couches). Etant une définition de l'espace en extension, le modèle matriciel peut générer de gros volumes de données. La qualité graphique des images produites est dépendante de l'échelle de représentation. En particulier, en augmentant la résolution de l'image, ou bien la surface couverte par cette représentation, le volume de données augmente de façon quadratique [8].

8. mais pas le volume d'information, donc les données peuvent être compressées sans perte.

En mode vectoriel, la surface contient des objets géographiques, qui sont définis chacun par leur description géométrique : points, lignes, polygones, (formes géométriques élémentaires), ou une combinaison de ces formes élémentaires. À chaque entité représentée est attaché un ensemble d'attributs descriptifs : des variables qualitatives ou bien quantitatives, comme en mode raster. C'est en quelque sorte une définition de l'espace en compréhension, comme lorsqu'on définit en mathématiques l'ensemble des nombres entiers pairs par la formule suivante : $\{x \mid x = 2 * k,\ k \in \mathbb{N}\}$. Lorsqu'il est employé pour des zonages, les unités du zonage sont modélisées comme des polygones (ou multi-polygones, si des archipels d'îles sont à représenter, par exemple). La topologie peut alors être représentée implicitement ou explicitement. Par exemple, dans les modèles dits « spaghettis », [Voisard 01], la topologie n'est pas explicitement décrite : si deux polygones partagent la même frontière, ils sont stockés indépendamment, sans signaler qu'il partagent une suite de points commune. Par rapport au mode raster, le mode vectoriel a l'avantage d'économiser l'espace de stockage. Celui-ci ne dépend que du nombre d'entités à représenter et non plus de la résolution de l'image. Cependant, généralement, pour rendre compte plus en détail de la réalité (sur un niveau de résolution fin), comme, par exemple, pour dessiner une côte maritime, le nombre de points utilisés pour définir le contour des objets doit être augmenté.

En effet, les objets peuvent être observés à différents niveaux de granularité (ou échelles), et cette granularité affecte le type de la géométrie utilisée pour représenter la position de l'objet. En effet, de loin, une ville peut se résumer à un point, mais en zoomant, à plus grande échelle, les formes de la ville se dessinent - elle occupe alors une surface non nulle -, elle est alors mieux représentée par un polygone par exemple. Il peut donc exister plusieurs représentations d'un même objet géographique, qui dépendent à la fois de l'usage souhaité et du niveau de généralisation de l'objet, [Ruas 99], mais aussi du niveau d'analyse de l'espace considéré [Mathian 01]. Nous nous référons ici à la question des échelles spatiales, précédemment abordée sur le plan théorique en section 2.1.2.1 page 38.

Le modèle raster est particulièrement utilisé comme fond de carte car il réussit à communiquer beaucoup d'informations [Longley 05]. L'avantage du modèle vectoriel est qu'il est plus adapté à l'abstraction et au raisonnement sur les positions relatives des objets. Les principaux acteurs commerciaux du domaine des SIG proposent des systèmes permettant de représenter des formes spatiales, que ce soit en mode raster ou bien en mode vectoriel. Parmi les produits commerciaux, les suivants sont les plus largement répandus :

- la *geodatabase* d'ArcGIS [9] qui est vendue par ESRI ;
- le SIG *MapInfo* [10], qui est vendu par Pitney Bowes Business Insight ;
- *Oracle Spatial* [11] qui est une extension spatiale du SGBD *Oracle 11g* vendu par Oracle.

Cependant, les primitives géométriques qui sont utilisées pour modéliser l'information spatiale en mode vectoriel diffèrent suivant ces outils, tout comme les opérateurs proposés pour manipuler les données. Par exemple, MapInfo utilise le type region pour décrire une collection de polygones, alors que Oracle Spatial n'utilise pas de type particulier pour les collections de polygones, mais distingue en revanche le rectangle (rectangle) comme un type particulier de polygone. De même, les collections de polygones exportées dans le format propriétaire *ShapeFile* d'ArcGIS sont spécifiés comme étant des objets de type PolygoneM ; ce format propose également le type MultiPatch pour représenter des collections d'objets de différents types.

Face à ce problème d'hétérogénéité, l'OGC a proposé un modèle, le *Simple Features Specification* [OGC 99] dès 1999, qui a évolué puis s'est concrétisé dans deux normes, les normes ISO 19107 - Spatial

9. http://www.esri.com/software/arcgis/arcinfo/index.html
10. http://www.pbinsight.com/welcome/mapinfo/
11. http://www.oracle.com/fr/products/database/options/spatial/index.html

Schema, et ISO 19123 - Schema for coverage geometry and functions. Ainsi, dans la norme ISO 19107, voir figure 2.7, les types region, rectangle ou MultiPatch sont abandonnés. La classe abstraite racine de la hiérarchie (Geometry) a des sous-classes (également abstraites, à l'exception de la classe Point) pour représenter des points (Point), des courbes (Curve), des surfaces (Surface) et des collections de géométries quelconques (GeometryCollection). Chaque objet géométrique est associé à un système spatial de référence (SpatialReferenceSystem), qui décrit le système de coordonnées dans lequel l'objet géométrique est défini, et l'unité de mesure des coordonnées (MeasureReferenceSystem). Les trois primitives que sont le point, la courbe ou bien la surface servent à définir ensuite des objets plus complexe par composition. Par exemple, la modélisation d'un archipel d'îles à l'aide d'un unique objet se fait via un objet de type MultiPolygon, spécialisant l'objet GeometryCollection, qui se fabrique par composition avec des objets de type Polygones, représentant chacun une île de l'archipel. Le type MultiPolygon est l'équivalent des types region ou PolygoneM qui ont été abandonnés.

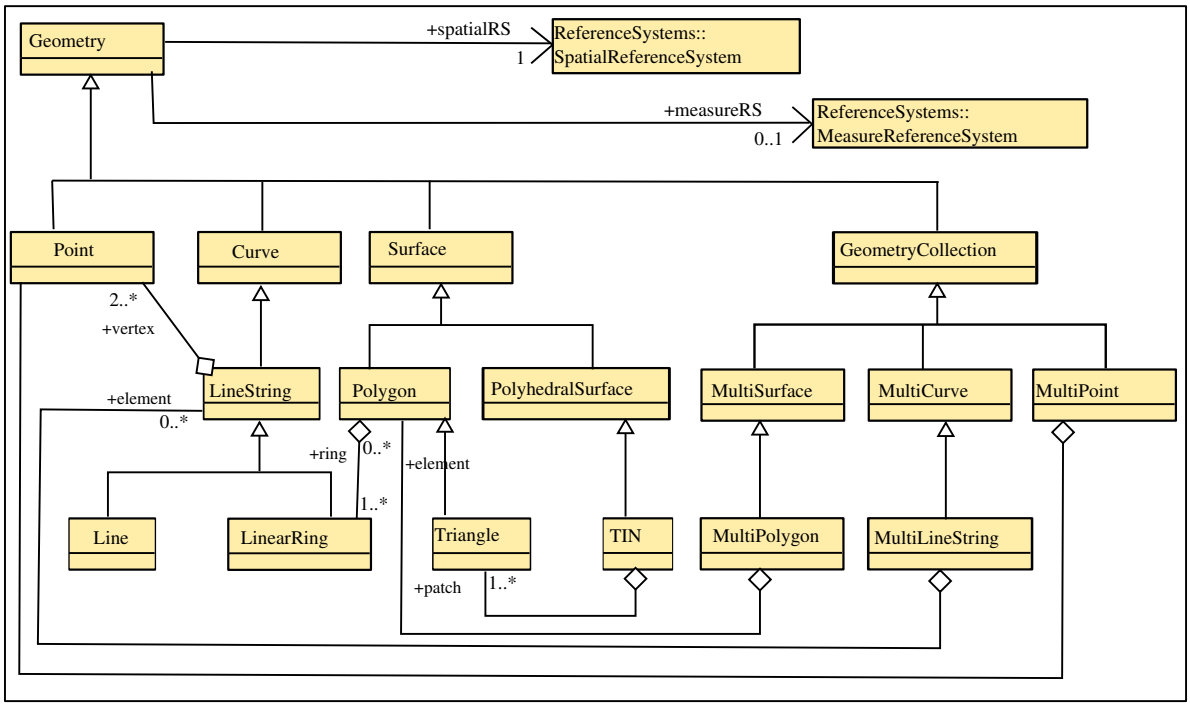


FIGURE 2.7 – Standard ISO 19107, (*Geographic information - Spatial schema*) pour représenter des données géographiques en mode vecteur.

Ainsi, les SIG et les bases de données spatiales peuvent s'appuyer aujourd'hui sur ces normes pour la représentation interne des données. Aujourd'hui, seul le SGBD open-source PostgreSQL (et sa cartouche spatiale PostGIS [12]), distribué sous licence open-source GNU *General Public Licence*, déclare être complètement conforme à ces normes. Par ailleurs, si les SIG n'utilisent pas ces normes pour la représentation interne des données, il existe une alternative pour l'interopérabilité entre SIG, celle d'exporter et d'échanger les données dans un format commun. À cette fin, le *Geography Markup Language* qui est une implémentation en XML des modèles standards de l'ISO 19107, a également été définie au sein de l'OGC, et la version 3.2.1 de cette spécification a été publiée en tant que norme ISO 19136 à la mi-2007, [ISO 07b].

12. `http://postgis.refractions.net/`

2.1.2.2.4 Raisonner dans l'espace Une fois la représentation de l'espace établie, il est nécessaire de savoir raisonner sur les positions relatives des objets qui le constituent. Il s'agit de définir des relations spatiales. Les relations spatiales ne servent pas qu'au domaine des SIG : elles sont utilisées et définies mathématiquement par les chercheurs en Intelligence Artificielle (I.A.), vision, imagerie algorithmique, cartes cognitives, représentation des connaissances, analyses spatiales, structures de données spatiales. Il existe trois grandes catégories de relations spatiales [Egenhofer 89] : les *relations métriques*, les *relations topologiques* et enfin les *relations d'ordre*. Clementini, [Clementini 08], note que les espaces topologiques, projectifs et métriques sont liés hiérarchiquement, dans le sens où :

> « en suivant l'approche axiomatique dans les espaces géométriques, un espace peut être construit à partir du précédent en ajoutant davantage d'axiomes. Si une propriété géométrique est évaluée dans l'espace topologique, la même propriété restera valable dans l'espace projectif et l'espace euclidien. En revanche, une propriété vraie dans l'espace euclidien peut ne plus rester vraie dans l'espace projectif ou topologique. »

2.1.2.2.4.1 Les relations métriques Les relations métriques donnent une mesure quantitative de l'espace. La plus connue se base sur la distance euclidienne, que nous avons présentée précédemment, en exposant les limitations pour le raisonnement en géographie qu'elle présente. L'expression la plus générale de la mesure d'une distance à partir des coordonnées x_i et y_i du point i, x_j et y_j du point j est la suivante :

$$d_p(ij) = (|x_i - x_j|^p + |y_i - y_j|^p)^{\frac{1}{p}} \tag{2.3}$$

Si $p = 2$, on définit la distance euclidienne. Si $p = 1$, on a la distance de Manhattan, dite encore distance rectilinéaire ou rectangulaire. Si $p < 1$, l'espace défini par cette distance n'est plus métrique [Pumain 97].

2.1.2.2.4.2 Les relations topologiques L'ensemble des relations topologiques est un sous ensemble des relations géométriques. Elles ont la propriété d'être préservées lors de transformations topologiques comme des translations, des rotations ou des facteurs d'échelle. L'information topologique est par ailleurs purement qualitative et exclut toute mesure quantitative. Il existe plusieurs théories topologiques pour la définition de relations topologiques spatiales. Il y a, d'une part, la géométrie élémentaire définie par Tarski, [Tarski 59], qui est fondée sur la notion de points comme uniques primitives décrivant l'espace avec l'usage d'un ensemble d'axiomes relatifs aux notions de distance et de voisinage, et, d'autre part, les théories topologiques basées sur des notions de régions et d'axiomes relatifs aux relations entre régions ([Clarke 85]). Ces théories donnent lieu à trois modèles, le modèle des 9-intersections (dit 9-i) [Egenhofer 91], le modèle CBM [Clementini 93] et le modèle des RCC [Randell 92]. Une étude comparative de ces trois modèles montre que ces modèles ont des pouvoirs d'expression similaire, et qu'il est en fait impossible d'en distinguer un meilleur que les autres. Le modèle des 9-intersections est celui ayant rencontré le plus de succès dans ses applications. Dans le SGBD PostgreSQL et sa cartouche spatial PostGIS, les opérations topologiques proposées intègrent les modèles CBM et 9-i.

Le ***modèle des 9-intersections*** proposé par Egenhofer, [Egenhofer 91], permet de caractériser les différentes manières suivant lesquelles deux objets peuvent entrer en relation topologique, et ceci pour des dimensions quelconques. Ce modèle est basé sur la théorie des ensembles de points de la topologie algébrique, et repose sur l'usage de deux primitives appliquée à un ensemble de points A placé dans un espace X, que sont la *frontière* ∂A, et l'*intérieur* A° et sur la distinction entre *ouvert* et *fermé* : un ouvert est un sous-ensemble topologique qui ne contient aucun point de sa frontière, tandis qu'un fermé est le complémentaire d'un ouvert.

– l'intérieur A° est l'union de tous les ouverts inclus dans A, et A° est le plus grand ouvert inclus dans A ;
– l'extérieur (se note aussi A^-) $\neg A$ est l'intérieur du complémentaire de A ;
– l'adhérence $\bar{A}$ (ou la fermeture de A) est le plus petit ensemble fermé contenu dans A ;
– la frontière ∂A est l'adhérence de A sans l'intérieur de A ($\partial A = \bar{A} \backslash A^\circ$) ;

De ces définitions, il découle que la frontière ∂A, l'intérieur A°, et l'extérieur $\neg A$ d'un objet A sont mutuellement exclusives, et que leur union forme une couverture complète de l'espace. Dans le cas particulier d'un espace métrique (E, d), on peut définir simplement la frontière comme l'ensemble des points y pour lesquels toute boule de centre x et de rayon strictement positif ε possède une intersection non vide avec A ainsi qu'avec son complémentaire $\neg A$. De même, une partie A de cet espace est ouverte si et seulement si pour tout point x de A, il existe une boule centrée sur x et et de rayon strictement positif ε incluse dans A. La figure 2.8 illustre ces notions.

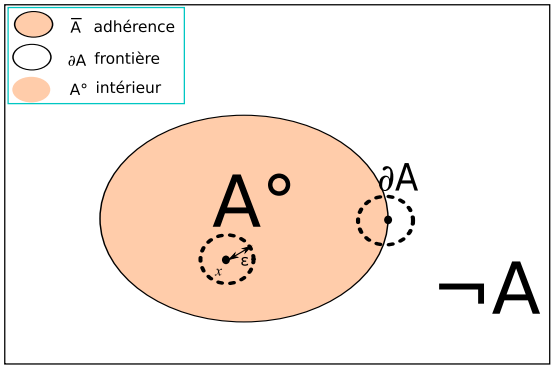


FIGURE 2.8 – Intérieur, frontière et extérieur d'une partie A.

À partir de ces définitions, la matrice des 9 intersections entre une région A et B peut-être définie, quelque soit la dimension de A et B (0 pour des points, 1 pour des lignes, 2 pour des régions). Les valeurs de la matrice sont le vide, noté $\emptyset$ ou bien le non vide, noté $\neg\emptyset$, suivant que la frontière, l'intérieur ou le complémentaire d'une partie intersecte ou non la frontière, l'intérieur ou bien le complémentaire de l'autre partie, voir equation 2.4

$$\begin{pmatrix} A^\circ \cap B^\circ & A^\circ \cap \partial B & A^\circ \cap \neg B \\ \partial A \cap B^\circ & \partial A \cap \partial B & \partial A \cap \neg B \\ \neg A \cap B^\circ & \neg A \cap \partial B & \neg A \cap \neg B \end{pmatrix} \tag{2.4}$$

Le modèle des 9-i est valide dans tout espace topologique connexe [13] ou non. Cependant, pour la partie qui définit les relations binaires entre régions du plan, toute région est considérée comme connexe. Autrement dit, les régions considérées sont sans trous. Sur la base de la matrice d'intersections, on déduit qu'il existe 8 types de relations possibles entre deux régions connexes de dimension 2. Les relations proposés sont la déconnexion, connexion extérieure, intersection, connexion intérieure, inclusion, égalité entre deux formes représentées en 2 dimensions, voir figure 2.9. Par exemple, l'équation 2.5 donne la valeur de la matrice associée à la connexion extérieure, et signifie que A et B s'intersectent seulement par leurs frontières ($\partial A \cap \partial B = \neg\emptyset$).

13. Un espace X est dit connexe si les seules parties ouvertes et fermées de X sont X et l'ensemble vide.

$$\begin{pmatrix} \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix} \tag{2.5}$$

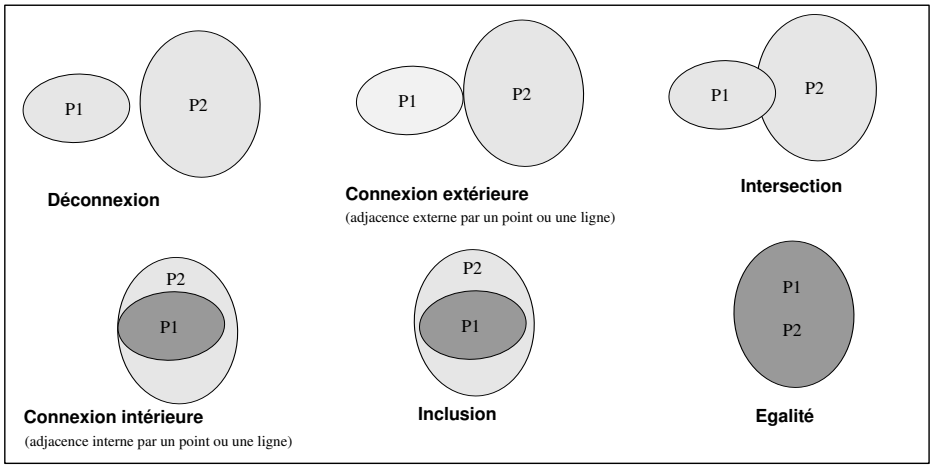


FIGURE 2.9 – Les relations spatiales entre deux ensembles de points, (d'après [Egenhofer 91]).

Ce modèle de relation s'applique, par exemple, aux formes géométriques en deux dimensions des entités territoriales d'un maillage, et permet de déduire les relations de recouvrement, disjonction, inclusion, etc., qu'entretiennent ces unités entre elles. En particulier, cette grammaire sert de support aux spécifications de l'OGC et du DGIWG [14], qui elles-mêmes sont intégrées dans les SGBD spatiaux qui proposent des fonctions prédéfinies pour calculer la topologie d'entités spatiales.

La ***méthode dimensionnelle étendue*** (*dimension extended method - DEM*), également adoptée par l'OGC, est une extension du modèle des 9-i présentée par [Clementini 93]. Elle consiste en la prise en compte dans le modèle des 9-i de la dimension de l'intersection entre primitives topologiques, puis regroupe les relations topologiques en 5 groupes mutuellement exclusifs. *Via* la ***Calcululs Based Method - CBM***, il est possible de décrire toutes les relations du modèle étendu en combinant ces groupes. Mais ce modèle est plus restrictif que le modèle des 9-i car on se place directement dans un plan (dimension 2) et tous les objets sont fermés et connectés. On exclut donc les unions d'objets séparés, et les objets n'ont pas de trous [Billen 02]. Par ailleurs, la dimension de tout objet est calculée par une fonction *dim* (0 pour un point, 1 pour une ligne, et 2 pour une région). L'idée sous-jacente du modèle CBM est de faciliter la compréhension et l'organisation des relations topologiques pour les opérateurs humains. Cinq relations sont ainsi définies entre A et B, en considérant leurs intérieurs respectifs A° et B° (voir l'équation 2.6) : touche (*touch*), est dans (*in*), croise (*cross*), recouvre (*overlap*), est disjoint (*disjoint*).

14. Groupe de Travail Militaire sur l'Information Géospatiale - en anglais, le *Defence Geospatial Information Working Group.*

$$
\begin{aligned}
&\langle A, touch, B\rangle \Leftrightarrow (A^\circ \cap B^\circ = \emptyset) \wedge (A \cap B \neq \emptyset)\\
&\langle A, in, B\rangle \Leftrightarrow (A \cap B = A) \wedge (A^\circ \cap B^\circ \neq \emptyset)\\
&\langle A, cross, B\rangle \Leftrightarrow (dim(A^\circ \cap B^\circ) = (max(dim(A^\circ), dim(B^\circ)) - 1) \wedge (A \cap B \neq A) \wedge (A \cap B \neq B)\\
&\langle A, overlap, B\rangle \Leftrightarrow (dim(A^\circ) = dim(B^\circ) = dim(A^\circ \cap B^\circ)) \wedge (A \cap B \neq A) \wedge (A \cap B \neq B)\\
&\langle A, disjoint, B\rangle \Leftrightarrow (A \cap B \neq \emptyset)
\end{aligned}
\tag{2.6}
$$

La définition d'opérateurs de frontière peut se combiner avec les relations précédentes permet de distinguer un plus grand nombre de configurations. Pour une région, le modèle propose l'opérateur (A, b) qui extrait la ligne fermée définissant la frontière de A. Entre deux régions simples, ce modèle permet donc d'identifier deux relations de plus que le modèle des 9-i, voir figure 2.10. Par exemple, des régions peuvent se toucher par une ligne ($dim(A \cap B) = 1$) ou bien par un point ($dim(A \cap B) = 0$), comme dans les cas 1) *versus* 2) et 4) *versus* 5).

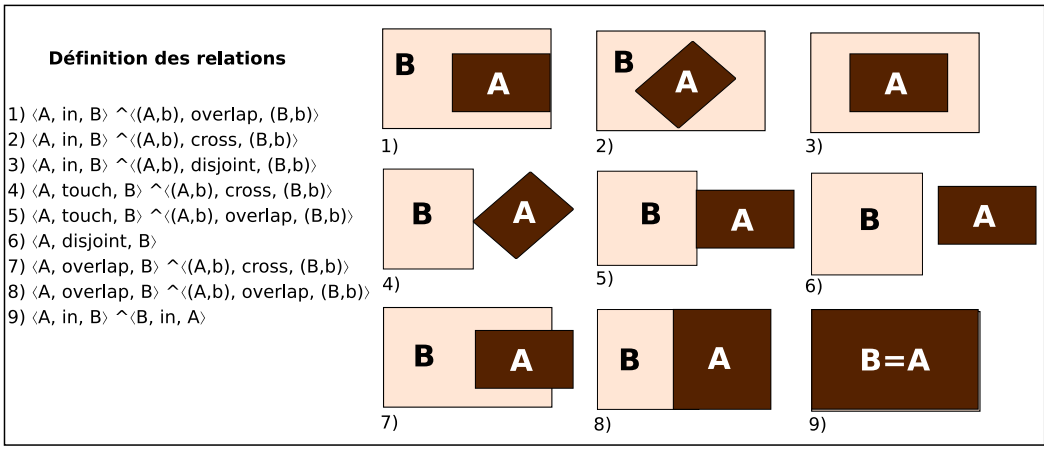


FIGURE 2.10 – Les 9 relations spatiales entre deux régions simples connexes suivant le DEM et le CBM, (d'après [Ubeda 97]).

Les théories mathématiques de l'espace sont traditionnellement basées sur l'utilisation du point comme entité spatiale primitive, et définissent des entités étendues (telles que les régions) comme des ensembles de points. La méréologie ou le calcul des individus rejette cependant le choix du point comme entité primitive et le classifie comme erreur philosophique [Cohn 01]. Ainsi, la deuxième tendance dans la communauté d'intelligence artificielle est de considérer la région comme élément primitif, et d'argumenter ce choix en soulignant le fait que l'extension spatiale de tout objet du monde réel est de type région. Les points peuvent être définis, si nécessaire, en termes de régions.

Ainsi, les ***algèbres de type RCC*** (*Region Connection Calculus*), qui sont biens connues et ont donné lieu à de multiples développements et améliorations, sont fondées sur une approche par régions comme structures de base et s'inspirent des travaux de Clarke [Clarke 81]. La théorie de Clarke définit comme notion de base la *connexion* entre deux régions de l'espace, x et y, et décrit cette connexion en utilisant la relation symétrique et réflexive $C(x, y)$. Dans les algèbres RCC, l'interprétation de la relation $C(x, y)$ est légèrement changée pour exprimer le fait que la fermeture de deux régions partage un point [Randell 92]. La primitive $C(x, y)$ est très puissante, car elle peut être utilisée pour définir formellement tout un ensemble de fonctions et de prédicats spatiaux qui capturent des relations topologiques intéres-

santes et utiles, caractérisant différents degrés de connexion entre deux régions. La palette de relations définie par [Randell 92], illustrées par la figure 2.11, inclut les situations suivantes :

- les régions sont *déconnectées* : $DC(x, y)$
- les frontières des régions sont *connectées* : $EC(x, y)$
- les régions sont *partiellement superposées* : $PO(x, y)$
- la région x est *partie tangentielle propre* de la région y : $TPP(x, y)$
- la région x est *partie tangentielle non propre* de la région y : $NTPP(x, y)$
- la région x est *partie propre* de la région y : $PP(x, y)$
- la région x fait *partie* de la région y : $P(x, y)$
- la région x est *égale* à la région y : $EQ(x, y)$
- les régions x et y se *superposent* : $O(x, y)$
- les régions x et y sont *connectées* : $C(x, y)$
- la région x est *discrète* par rapport à la région y : $DR(x, y)$

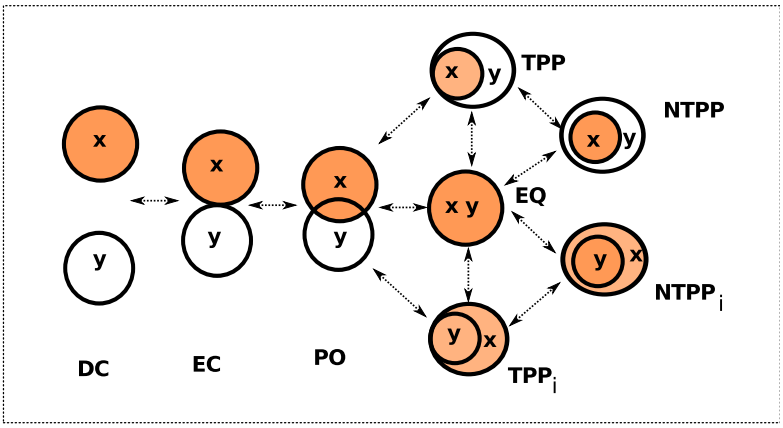


FIGURE 2.11 – Illustration des relations de l'algèbre RCC-8 et de leurs transitions continues.

Les huit relations de base $DC, EC, PO, TPP, NTPP, EQ, TPP_i, NTPP_i$ représentent un ensemble particulier de relations topologiques connu sous le nom d'algèbre RCC-8. Une autre algèbre très connue est RCC-5, qui renonce à la distinction entre la frontière et l'intérieur d'une région, et considère en conséquence égales, les relations DC et EC, ainsi que les relations TPP et $NTPP$. Il est à remarquer que ces relations sont identiques aux 8 relations distinctes définies par le modèle des 9-i, avec une correspondance un pour un entre les relations [Cui 93].

Par ailleurs, toutes les approches RCC ont en commun le fait que les relations qu'elles modélisent sont axiomatisées et définies en utilisant la logique du premier ordre qui leur fournit une sémantique formelle. Cependant, les procédures de décision basées sur cette formalisation ne sont pas très efficaces (problème, en général, NP-complet) [Renz 98]. Les travaux qui ont porté sur les propriétés formelles de ces théories [Grzegorczyk 51], [Selivanov 09], montrent que les raisonnements qu'elles offrent sont incomplets et ne sont pas décidables.

2.1.2.2.4.3 Les relations d'ordre Les relations ordinales (ou projectives, selon [Clementini 08]) peuvent être décrites par des propriétés projectives de l'espace sans avoir recours aux propriétés métriques de l'espace [Billen 04]. Comme les relations topologiques, les relations projectives sont de nature qualitative, car elles n'ont pas besoin de mesures précises pour être expliquées [Egenhofer 95]. Aussi, les relations projectives sont plus précises que les relations topologiques et peuvent servir de base pour

décrire les relations qui ne sont pas décrites par la topologie. Etant une étape intermédiaire entre la métrique et la topologie, les relations projectives sont aussi variées que « à droite de », « en avant de», « entre », « le long de », « entouré par », « devant », « arrière », « au nord de », « à l'est de », et ainsi de suite. Ce type de relation regroupe les relations cardinales, [Frank 92], d'orientation [Hernàndez 93], ou de direction cardinales [Goyal 00].

Concernant les relations cardinales par exemple, Frank définit une représentation basée sur un ensemble de symboles définissant les valeurs que peuvent prendre les relations, un ensemble d'opérations applicables à ces relations et enfin un ensemble d'axiomes définissant les résultats des opérations [Frank 92]. Il propose également plusieurs découpages de l'espace pour établir ses relations comme dans la figure 2.12.

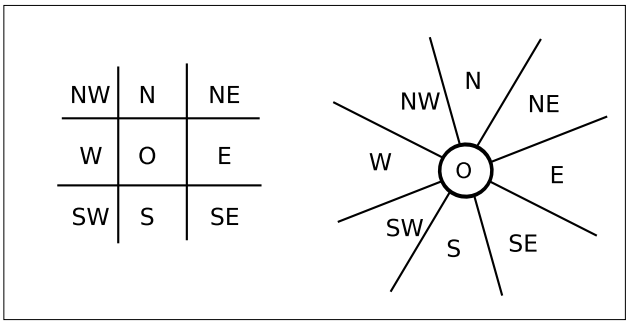


FIGURE 2.12 – Division de l'espace suivant une projection et suivant des secteurs angulaires, d'après [Frank 92].

Le travail de Clementini [Clementini 08] propose pour toutes ces relations d'ordre (qu'il nomme « projectives ») un modèle fédérateur capable de représenter toutes les variantes des relations projectives. Ce modèle est basé sur le modèle des cinq intersections [Clementini 06], qui fournit une classification des relations projectives entre trois objets géométriques dans le plan, en distinguant 34 relations qui forment un ensemble JEPD (*Jointly-Exhaustive and Pair-wise Disjoint*). Le calcul de relations projectives repose sur la colinéarité des points dans l'espace de travail.

2.2 Mettre en correspondance le temps et l'espace

Un nombre important d'études ont été menées pour modéliser et représenter l'information géographique dans toutes ses dimensions, spatiales et temporelles, montrant que les aspects cognitifs liés à la compréhension humaine de l'espace géographique sont difficiles à maîtriser [Peuquet 02]. En effet, l'espace dans lequel nous vivons est un continuum, mais la mesure de cet espace rapporte des faits ponctuels, localisés dans le temps et l'espace, pour lesquels il n'existe pas de référentiel universel, puisque les référentiels existants (échelles de temps, projections spatiales) varient en fonction des cultures et du savoir scientifique. La simulation de cet espace en tant que continuum est une approche quantitative basée sur l'emploi d'outils analogiques et de méthodes numériques, que nous écartons au profit d'une approche qualitative telle que proposée par Lardon, [Lardon 99], et Cheylan, [Cheylan 97]. Celle-ci nous semble plus appropriée dans le cadre d'une analyse exploratoire d'un espace d'étude. En effet, l'analyste est intéressé par la situation relative d'un territoire par rapport à un autre et par les relations que ce territoire entretient avec le reste de l'univers d'étude plutôt que par la visualisation d'un continuum géographique.

Les travaux de Peuquet [Peuquet 94, Peuquet 02], ainsi que ceux de Yuan [Yuan 99] mettent en exergue l'importance que revêt une gestion appropriée du temps, non pas comme un simple attribut de l'espace géographique, mais comme une dimension à part entière dans un système d'information spatio-temporel. Le temps et l'espace sont fondamentalement différents : dans l'espace, ce qui se passe en un lieu A interagit avec ce qui se passe en un autre lieu B, même pour un effet infime, et cette interaction constitue une influence réciproque. *A contrario*, le temps étant orienté, les événements qui se produisent après d'autres n'ont pas d'impacts sur ceux qui les précèdent sur la ligne temporelle. Ainsi, une simple représentation 3D+1 ne suffirait pas pour la production de requêtes adaptées aux phénomènes spatio-temporels. Ces travaux soulignent qu'un tel système doit supporter la représentation simulanée de trois domaines (ou dimensions) : la dimension spatiale (les cellules du zonage ou unités géographiques, le où), la dimension temporelle (les événements, le quand) et la dimension thématique (les valeurs des variables statistiques, le quoi). Ce système idéal doit permettre de pivoter de la dimension spatiale à la dimension temporelle ou thématique, en répondant aux questions suivantes :

- **Où** se trouvait un objet à une certain moment ?
- **Quand** se trouvait cet objet à cet endroit ?
- **Quel** objet se trouvait à cet endroit à ce moment là ?

Ceci définit la triade de Peuquet, voir figure 2.13.

A ces questions, Thériault et Claramunt ajoutent la suivante, [Thériault 99] : « **Comment** cet objet s'est trouvé à cet endroit à ce moment là ? ». Répondre à cette question, c'est décrire le caractère dynamique des évènements qui occasionnent les changements observés, et produire une description des processus responsables des transformations. La triade peut donc être complétée par un cercle qui relie les trois éléments <objet, lieu, temps> répondant aux questions du Quoi ? Où ? Quand ? pour leur donner sens comme dans la figure 2.13.

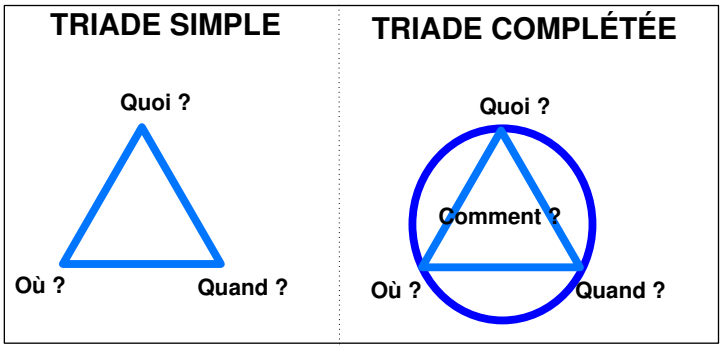


FIGURE 2.13 – Triade de spatio-temporelle simple [Peuquet 94], puis complétée [Thériault 99].

Dans le cadre de l'analyse de données statistiques associées à des zonages changeants, ces questions peuvent s'interpréter comme suit :

- pour une unité géographique donnée, quelle est la courbe d'évolution d'une variable thématique donnée ?
- pour une date donnée, quel est le découpage territorial de l'espace et les valeurs des attributs thématiques associés ?
- pour une date, et une unité géographique donnée, de quelles recompositions territoriales est-elle issue et comment se transforme-t-elle par la suite ?

Diverses tentatives de modélisation de l'information spatio-temporelle, concrétisées parfois dans des prototypes applicatifs ont vu le jour depuis les années 1990, [Albrecht 07]. Une étude fondamentale de Langran [Langran 92] détaille les besoins techniques que ces modèles doivent satisfaire au mieux. Ces besoins sont :

- la modélisation des changements du support spatial des données
- le stockage et le traitement de l'évolution des attributs non spatiaux : la thématique (ensemble des indicateurs) ou bien la sémantique (nom des localisations).
- l'interrogation des données en fonction des dimensions spatiale et/ou temporelle.
- la logistique pour le traitement des données, avec des algorithmes efficaces pour intégrer, mettre à jour, supprimer, interroger les données.

Trois tendances distinctes de modélisation se sont dessinées : la première tendance vise à décrire les états successifs de l'espace géographique au mieux, avec une conception du temps linéaire et orthogonale au plan spatial. La seconde tendance vise à répondre de façon plus satisfaisante aux questions posées par la triade de Peuquet (où, quand, et quoi ?). Enfin, la dernière tendance s'intéresse aux évènements et processus de transformation à l'oeuvre dans l'espace géographique, pour modéliser le temps sous la forme de ses effets, et répondre à la question du « comment ? ». Nous décrivons ici ces modèles, et nous les mettons en rapport avec les besoins qui ont été définis ci-dessus.

2.2.1 Des modèles pour enregistrer les changements

2.2.1.1 Datation du support de collecte

Nous présentons ici les premières approches, de type *ad-hoc*, qui sont très proches du modèle de collecte des données. Ainsi, le modèle de superposition des couches géographiques datées (*snapshot layers* en anglais) se fixe comme objet d'étude une couche géographique (*layer* en anglais) définie par l'association d'une donnée et de son support spatial de recensement ou de mesure. Le support spatial d'une couche est constitué d'un ensemble d'objets de même type : points, cellules, lignes ou polygones. La donnée est, par exemple, la mesure de la température en quelques points de l'espace, ou bien, par exemple, c'est le type d'occupation du sol majoritaire sur chaque cellule d'une grille régulière qui constitue le zonage de l'espace, comme illustré dans la figure 2.14. Ce type de modèle est très largement utilisé dans le cadre de collecte de données sur les différents types d'occupation biophysiques du sol via le traitement d'images satellitaires, avec, par exemple, le programme du Corine Land Cover mené par l'Agence Européenne de l'Environnement (*European Environmental Agency*, EEA). En supposant que les supports spatiaux ne varient pas dans le temps (comme c'est le cas pour le Corine Land Cover, puisque le georéférencement, la taille et la forme de la grille ont été établis par convention dès le début de cette collecte, en 1990), alors la dimension temporelle est introduit simplement par l'estampillage des couches géographiques avec la date de validité de la donnée. Le postulat sur l'homogénéité temporelle du support spatial des données dans le temps permet une étude de la variabilité temporelle des indicateurs thématiques dans le temps. En utilisant ce modèle, il est possible d'interpoler temporellement les valeurs des variables thématiques, comme le propose Beller [Beller 91]. Beller utilise ce type d'approche pour un prototype pour la gestion des évolutions environnementales planétaires, *Global Change Research* en anglais, cherchant à mettre en relation les données climatiques, d'occupation du sol, et de pollution. Les intervalles de temps entre chaque instantané (*snapshot*) peuvent varier, ils dépendent de la fréquence de production des grilles de données.

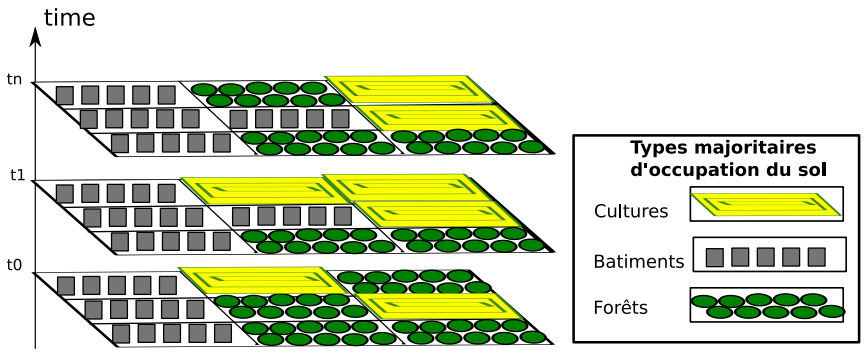


FIGURE 2.14 – Superposition de couches géographiques dans le temps - Illustration avec des données d'occupation du sol.

Ce modèle de gestion du temps s'intègre bien avec le mode d'approvisionnement des données dans des SIG, qui sont encore atemporels et ne fournissent aucune opération topologique sur la dimension temporelle. En général, chaque année, de nouvelles séries de données sont produites par des fournisseurs de données (institut statistiques nationaux ou internationaux) qui se basent autant que possible sur les supports spatiaux de données existants.

Cependant, il arrive que ces supports (localisation des mesures, ou bien la forme et le nombre d'unités géographiques) soient modifiés, et dans ce cas, les hypothèses simples de ce modèle ne sont plus vérifiées et perturbent l'analyse. De plus, certains auteurs, [Armstrong 88], [Langran 88], observent que l'approche *snapshot layers* est très consommatrice d'espace-mémoire, car, avec la multiplication des couches géographiques et des données, la quantité de données enregistrée est également multipliée, et de façon parfois inutile. En effet, certains enregistrements sont redondants, puisque les parties du support spatial inchangées d'une version à l'autre sont tout de même enregistrées en doublon.

2.2.1.2 Définition d'un support stable dans le temps

D'autres approches, visant à prendre en compte les changements de forme du support, ont vu le jour. Le point commun de ces approches est de définir un référentiel spatial fixe dans le temps, auquel les formes qui évoluent dans le temps sont rattachées par un lien établi lors de la saisie des données. Ce lien, ou table de passage, est utilisé pour reconstituer la forme d'une version du support, qui constitue alors une composition des formes du référentiel fixe dans le temps. Ceci nécessite en général que ce référentiel fixe dans le temps soit le plus fin possible, et nous appelons ce référentiel spatial fixe le "Plus Petit Commun Dénominateur Spatial" (PPCD-Spatial) par analogie avec la manipulation de nombre en arithmétique, où 6, 15 et 27 partagent en commun 3 comme dénominateur, et peuvent donc être décrits suivant un multiple de 3 : 3x2, 3x5 et 3x9. Ce terme est emprunté à d'autres auteurs, [Ott 01], qui utilisent le terme anglais de *Least Common Geometry*, pour implémenter cette approche dans un modèle vectoriel pour l'analyse de la répartition des sites de peuplements juifs en Palestine entre le XIXème et le XXeme siècle. Ce type d'approche se décline différemment en fonction de la modélisation du support employée pour le PPCD-Spatial, qui peut être vectorielle ou matricielle.

2.2.1.2.1 Manipulation d'un format vectoriel Le modèle proposé par Langran et Chrisman, [Langran 98], nommé *Space-Time Composite Model* en anglais, ou "Modèle à composition d'entités Spatio-Temporelles" en français, est un précurseur de cette famille de modèles qui nécessitent l'introduction d'un PPCD-Spatial, et qui le définissent sous un format vectoriel.

Ce modèle représente le monde réel comme une collection d'entités spatio-temporelles homogènes, qui évoluent individuellement à des rythmes différents dans le temps. L'objectif est de pouvoir restituer l'histoire des changements survenus sur une parcelle de terrain donnée. Ce modèle repose sur une couche géographique initiale (qui sera donc considérée comme le PPCD-Spatial), pour laquelle les caractéristiques spatiales de toutes les entités présentes sont enregistrées. Un identifiant unique est attribué à chaque entité. On introduit successivement dans le modèle les couches dans l'ordre chronologique, et pour chaque série, n'est enregistré que ce qui a été modifié par rapport à la couche précédente. Ce modèle réduit donc l'espace de stockage puisque ne sont enregistrés que les éléments nouveaux ou modifiés à chaque version du support.

La figure 2.15 présente le modèle composite simple (*Space-Time Composite* de Langran), qui modélise les changements d'occupation du sol (entre rural et urbain) à chaque instant t_i, et ne stocke que les changements entre chaque instant.

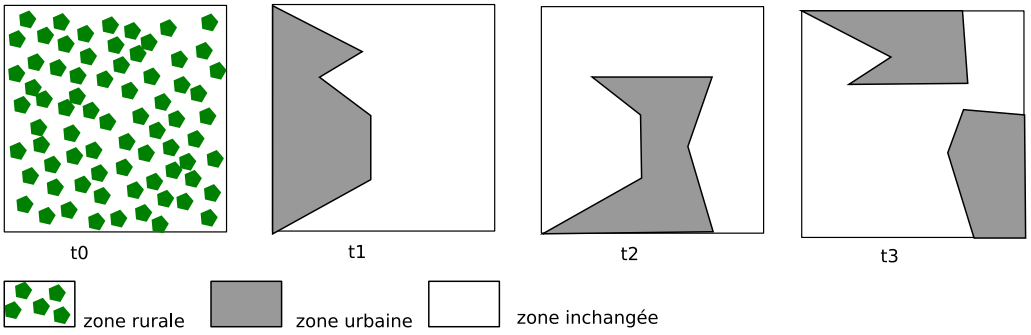
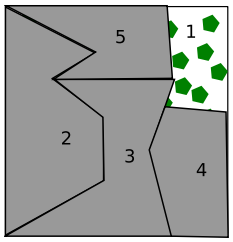


FIGURE 2.15 – Utilisation du modèle *Space-Time Composite* pour des changements d'occupation du sol.

Ce modèle a été perfectionné par Belussi et al., [Belussi 99], qui, au lieu de conserver les changements dans les versions successives, constituent une couche historique pour intégrer toutes les mises à jour historiques dans une même couche fusionnée. Ainsi l'enregistrement successif des couches de changement est économisé : seules restent la couche initiale de référence, et la couche historique, résultat de la fusion par intersection de toutes les couches (le PPCD-Spatial). Lors de l'introduction d'une couche, les supports spatiaux sont intersectés de manière à créer le support le plus fin, et chaque enregistrement dans la base concerne les parcelles ayant subi des modifications de forme et de taille. Le résultat des multiples fusions par croisement de support ne ressemble pas au terrain réel à aucune date. Cependant, grâce aux enregistrements datés des changements, il est possible de reconstruire l'état du zonage à un instant t_q à partir du support initial en y ajoutant les éléments présents à l'instant considéré.

La couche historique correspondant à l'exemple des changements d'usage du sol est décrite dans la figure 2.16 : le terrain initial a été divisé en l'ensemble des parcelles numérotées où se sont produits des changements d'occupation du sol. La relation (parcelle, temps, occupation du sol) permet de retrouver

quelle était la forme du zonage à un instant précis.



Parcelles	t0	t1	t2	t3
1	Rural	Rural	Rural	Rural
2	Rural	**Urbain**	Urbain	Urbain
3	Rural	Rural	**Urbain**	Urbain
4	Rural	Rural	Rural	**Urbain**
5	Rural	Rural	Rural	**Urbain**

Relation (Parcelle, temps, occupation du sol)

FIGURE 2.16 – Adaptation du modèle *Space-Time Composite* intégrant la couche historique pour des changements d'occupation du sol.

Dans cet exemple, les formes des unités composant une version de zonage sont définies par le type d'occupation du sol. Mais, en général, dans le cadre de la collecte de données statistiques, la forme du zonage est établie à l'avance, et reflète le plus souvent des territoires administratifs, et les données thématiques collectées ne définissent pas la forme du zonage. Gregory, [Gregory 02], fait alors remarquer que l'usage du modèle *Space-Time Composite* avec couche historique pré-suppose donc un choix de plus petites unités stables dans le temps, qui peuvent soit déjà exister comme les paroisses (*parish*) en Suède dans l'étude de [Kristiansson 00], soit être fabriquées à partir de l'intersection de toutes les versions de zonage, comme dans l'étude de [Ott 01] précédemment citée.

Avec le modèle *Space-Time Composite*, il est aisé de répondre à la question suivante « Quels lieux ont vécu un changement d'occupation du sol entre t_i et t_j ? » par une simple requête qui sélectionne toutes parcelles dont l'occupation du sol a changé entre t_i et t_j. Egalement, la question « Quels lieux ont changé de forme entre t_i et t_j ? » trouve une réponse immédiate dans la lecture du modèle. En revanche, la question «Comment a évolué tel lieu sur la période bornée par les instants t_i et t_j ? » ne trouve pas de réponse dans le modèle. Ceci signifie que ce modèle ne permet pas de saisir les mouvements ou les transformations des entités spatiales. Par exemple, il est impossible de vérifier si deux parcelles ont fusionné pour former une parcelle unique entre deux versions de zonage.

L'autre faiblesse d'un tel modèle vient de sa complexité, comme l'observent [Yuan 96] et [Renolen 96]. D'une part, il est difficile de reconstituer la situation et les relations spatiales entretenues par les parcelles à un instant quelconque, et d'autre part, les mises à jour impliquent toujours une reconstruction des plus petites parcelles, et par conséquent une mise à jour des relations topologiques et des géométries de chaque parcelle, ainsi que des attributs temporels et thématiques.

2.2.1.2.2 Manipulation d'un format matriciel La question du changement spatial est aussi abordée dans l'informatique décisionnelle, avec la constitution d'entrepôts de données à références temporelles et spatiales. Ces systèmes reposent sur un concept, connu sous le nom anglais de *On-Line Analytical Processing* (OLAP), qui vise à organiser les données (ou mesures) sur des axes orthogonaux, (les dimensions), de façon à offrir des résumés des données sur chacun de ces axes à la demande. En effet, sur chaque dimension, les mesures peuvent être agrégées suivant une échelle croissante définie. En particulier, les systèmes *Spatial OLAP* (SOLAP), [Bédard 97], [Bédard 01], [Bimonte 07], qui sont destinés à produire des analyses en ligne sur des questions géolocalisées, définissent l'agrégation spatiale sur des échelles géographiques. Ces systèmes sont donc sensibles aux changements de forme du support.

La question de l'évolution des supports géographiques est abordée dans ce cadre par [Tchounikine 05]. Le modèle M3 proposé permet de définir des dimensions dont les liens et les membres sont associés à des intervalles de temps de validité, comme déjà exposé par [Body 03], mais en prenant en compte le changement spatial. Le modèle M3 repose en fait sur la même logique que celle exposée dans les modèles de type PPCD-Spatial vectoriel : il s'agit d'utiliser un PPCD-Spatial, et d'effectuer un mapping entre les différentes versions du support géographique via le PPCD-Spatial. Cependant, ce PPCD-Spatial est en fait une grille régulière qui produit une représentation matricielle de l'espace géographique, et chaque objet de l'espace est associé à un ensemble de cellules de cette grille, comme le présente la figure 2.17.

Cette grille est utilisée comme pivot, en conjonction avec une fonction de passage f entre les parcelles des différentes versions du support géographique, comme dans l'exemple de la figure 2.17. Le passage d'une version à l'autre est décrit grâce à une fonction f indiquant dans quels nouveaux polygones (et dans quelle proportion) chaque polygone d'une version antérieure se transforme. Par exemple, si $f(P1) = 1/2\ PF1 + 1/3\ PF2 + 1/6\ PF3$, alors PF1 intersecte la moitié de P1, PF2 intersecte le tiers de P1 et PF3 intersecte le sixième de P1. La fonction inverse $f^{-}1$ est aussi définie et enregistrée. Par exemple, si $f^{-}1(PF1) = P1$, alors P1 couvre complètement PF1. Cette fonction est utilisée pour transférer directement les variables associées (comme la population, ou un compte de ventes sur un territoire). Soit x'_1 la variable à estimer sur la parcelle PF1 et connue sur P1 avec la valeur x_1, alors $x'_1 = 1/2\ x_1$.

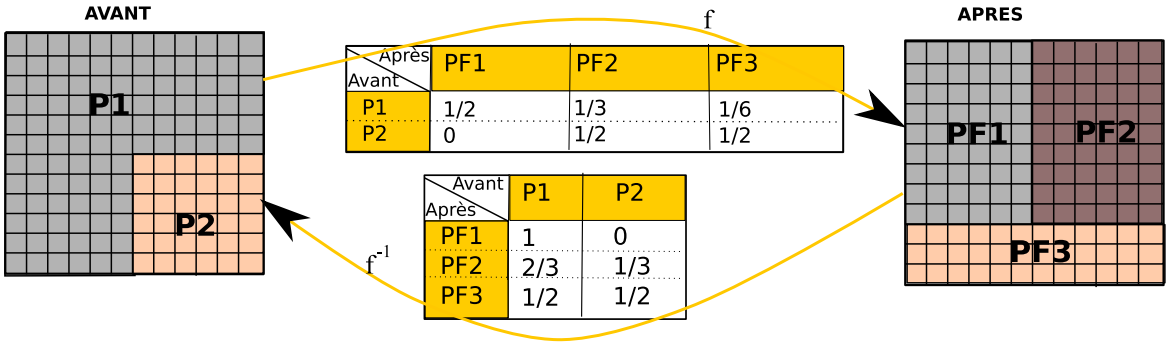


FIGURE 2.17 – Exemple d'utilisation de la grille et des fonctions de passage dans le modèle M3.

Cependant, cette approche présente les inconvénients suivants :

1. La fonction de passage est utilisé avec l'hypothèse implicite forte que la répartition des mesures est uniforme spatialement, ce qui est en général faux et relève d'une simplification extrême de la réalité géographique que nous voudrions éviter.
2. Aucun lien sémantique n'est établi entre les zones du support qui forment chacune un territoire : elles restent ici totalement anonymes. Ainsi, il est impossible de vérifier si une certaine composition de cellules est nouvelle ou bien identique à une précédente. Ce modèle ne réussit donc pas à rendre compte des mouvements ou des transformations des entités spatiales.
3. L'emploi de ce modèle présuppose une approximation des polygones par des cellules régulières : cette approximation peut ne pas être suffisamment précise si la grille est large, mais augmenter la résolution de la grille implique également d'augmenter l'espace mémoire consommé pour conserver le lien entre chaque zone et les cellules de la grille.
4. La gestion des évolutions spatiales est limitée à celle d'une couche géographique, de niveau élémentaire. La gestion des changements de l'organisation de la hiérarchie spatiale des unités n'est pas mentionnée explicitement.

2.2.2 Des modèles pour répondre aux questions du Quoi, Où, Quand ?

L'étude des précédents travaux montre qu'il manque un moyen pour donner un sens aux changements et qu'il est difficile de répondre de façon satisfaisante à la question suivante : "comment telle zone du support a-t-elle évolué au cours du temps ? ". Cette question peut être résolue par l'introduction d'un nouveau paradigme, comme l'argumente [Cheylan 93] : le paradigme identitaire, qui accorde une identité propre à chaque zone du support de l'information thématique.

2.2.2.1 La dynamique des entités spatio-temporelles

Les travaux de [Cheylan 93] et [Cheylan 97] mettent en avant, comme concept central de la modélisation, l'identité des entités spatio-temporelles qui constituent le support de chaque version de zonage. En effet, ces travaux s'inscrivent dans une approche qualitative, qui s'attache essentiellement aux structures d'ordre des espaces étudiés : un événement se produit avant, après ou simultanément par rapport à un autre, un lieu est localisé à l'intérieur, ou à côté d'un autre, etc. Toute information devient alors relative à la connaissance d'un autre phénomène, et la connaissance est organisée en fonction des relations qu'entretiennent les objets à la fois dans la dimension spatiale [Egenhofer 91] et dans la dimension temporelle [Allen 83]. Cette approche est donc basée sur la caractérisation d'objets particuliers dans l'espace, auxquels on accorde une identité propre. En résumé, cette conceptualisation se focalise sur l'entité géographique (ou unité géographique), qui est définie comme étant constituée de trois parties qui évoluent indépendamment au cours du temps : l'extension spatiale, l'identité, et la partie attributaire (dite aussi thématique), comme le montre la figure 2.18.

FIGURE 2.18 – Structuration d'une entité géographique, d'après [Cheylan 97].

Les données thématiques (la population, le niveau de richesse, de pollution...) décrivent les états d'objets géographiques (des villes, des régions, des fleuves...), et à ces objets se rattachent une extension spatiale qui renseigne sur leur localisation. L'extension spatiale est décrite par une géométrie euclidienne en deux dimensions, parfois trois dimensions, et elle se décline fondamentalement en trois types d'objets : ponctuels, linéaires et surfaciques. Ce sont des objets géométriques auxquels on peut appliquer des opérations topologiques [Egenhofer 91]. Ces objets ont une identité intrinsèque, indépendante de leurs attributs, qui eux, évoluent, changent au cours du temps. Ces attributs sont, par exemple, le nom, ou le code dans une nomenclature. La plupart des outils et formalismes de conception de Système Spatio-Temporels tels que MADS ([Vangenot 98], [Parent 06]), Perceptory ([Bédard 04]), ou POLLEN ([Gayte 97]) reposent sur ce paradigme.

L'étude du changement fait apparaître que les trois aspects distingués (espace, temps, thématique) d'une entité géographique sont susceptibles d'évoluer indépendamment les uns des autres au cours du temps. Premièrement, l'extension spatiale présente un mouvement dans l'espace : sa localisation, son tracé, ou sa forme (suivant le cas) change au cours du temps. Deuxièmement, les caractéristiques thématiques ont une existence limitée (une 'vie'), c'est-à-dire que les valeurs des données associées à l'entité

géographique changent (la population augmente, le débit de la rivière augmente, etc.). De même, les attributs de l'entité changent : une ville peut changer de nom, le code identifiant un secteur de recensement peut-être remplacé par un autre, ou bien deux communes fusionnent pour créer une nouvelle commune. Le processus de suivi de l'identité d'une entité au cours du temps est appelé *généalogie*. La figure 2.19 représente une extension de la représentation statique d'une entité géographique, à laquelle l'aspect dynamique et évolutif sous-tendu par la dimension temporelle a été ajouté.

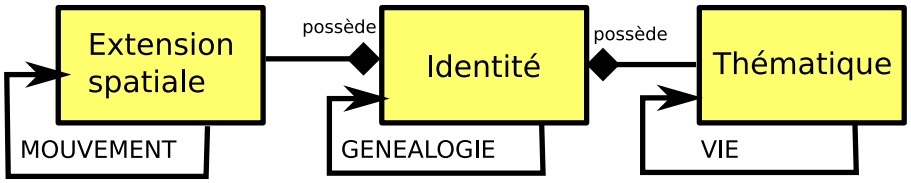


FIGURE 2.19 – Dynamisme d'une entité géographique, d'après [Cheylan 97].

2.2.2.2 Implémentations du paradigme identitaire

L'approche identitaire est également défendue par Worboys, [Worboys 98], et Wachowicz [Wachowicz 99] qui proposent de l'implémenter dans des modèles orientés-objet. Un des exemples d'implémentation que nous pouvons citer est celui de Dell'Erba et Libourel ([Dell'Erba 97]), qui s'applique aux évolutions du parcellaire cadastral, et repose sur une implémentation complètement orientée-objet, basée sur l'usage d'une base de données objet (O2), [Bancillon 91].

En effet, le paradigme identitaire se marie naturellement avec le concept de programmation orientée-objet puisque, par essence, les objets qui servent à modéliser la réalité dans des représentations informatiques, ont une identité. L'identité d'un objet reflète la continuation (ou l'endurance, selon Raper [Raper 95]) de la structure de l'objet modélisé à travers l'espace-temps. L'objet a une structure identique à celle de la famille d'objets à laquelle il appartient (sa classe), des méthodes, qui caractérisent le comportement des objets de cette famille, mais aussi des attributs qui le décrivent. Chaque instance de la classe est un objet, dont les valeurs d'attributs peuvent être identiques à celles d'autres objets de la même famille. Un objet possède toujours un identifiant le distinguant d'une autre instance qui aurait les mêmes attributs à un instant donné. Cet identifiant permet également de reconnaître un objet ayant évolué au cours du temps, et dont les valeurs d'attributs ont changé, mais qui reste pourtant toujours le même.

Le modèle orienté objet de Worboys, temporel [Worboys 92], ou bi-temporel ([Worboys 98]), identifie les attributs thématiques, spatiaux et sémantiques en tant qu'entités indépendantes évolutives (des atomes) qui composent une entité géographique globale (un objet). Dans ce modèle, le monde est perçu comme un ensemble d'atomes spatio-temporels en 3D, chaque atome couvrant une portion d'espace 2D et un intervalle de temps fixé, le temps étant orthogonal au plan spatial. Les propriétés de chaque atome sont stables (valeur des attributs thématiques, et couverture spatiale). Les atomes s'assemblent pour former des objets spatio-temporels qui représentent les changements des entités du monde réel.

Par exemple, le schéma de la figure 2.20 représente trois objets U, I et A. Dans cet exemple, l'identité de chaque objet est définie par le type d'usage du sol (agriculture, urbain, ou industrie). La zone urbaine U n'est composée que d'un atome U_1 qui apparaît au temps t_3 : sa validité est $[t3, now]$. L'objet A qui représente l'usage agricole du sol se réduit avec le temps : il est composé des atomes A_1 sur l'intervalle de temps $[t_1, t_2]$ et A_2 sur $[t_2, t_3]$ et la surface de A_2 est inférieure à la surface de A_1. Le troisième objet

du schéma est I qui porte comme information la surface occupée par des industries : I est composé des atomes I_1, I_2 et I_3 sur les intervalles de temps respectifs $[t_1, t_2]$, $[t_2, t_3]$, $[t_3, now]$. Pour reconstituer l'usage du sol à un temps t inclus dans $[t_2, t_3]$, il faut associer les atomes ayant l'intervalle de validité $[t_2, t_3]$, c'est-à-dire les atomes A_2 et I_2.

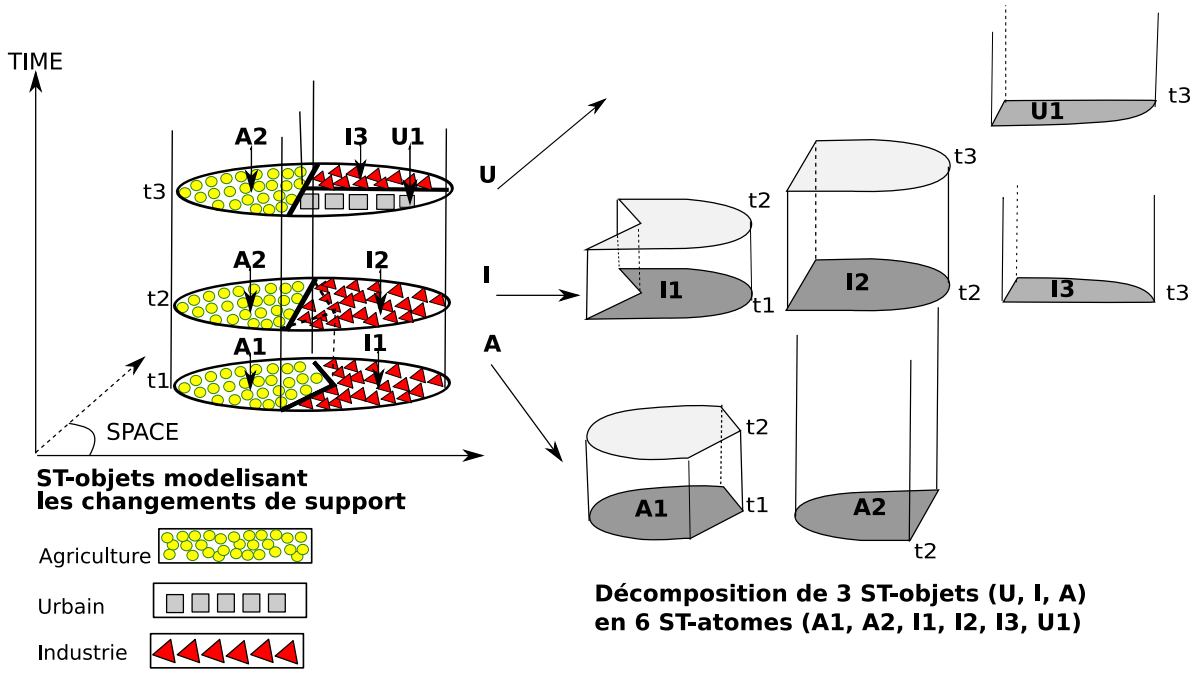


FIGURE 2.20 – Décomposition des objets en atomes, d'après [Worboys 98].

La reconstitution d'une couche géographique se fait par assemblage des atomes ayant un intervalle de validité temporelle commun (c'est une projection des atomes sur le plan spatial). La vie d'un objet est reconstituée par la projection sur l'axe temporel de la suite des atomes appartenant à cet objet. À cet effet, le système conserve le lien de composition entre les objets et les atomes via des identifiants uniques attribués aux objets comme aux atomes. Ce modèle répond donc bien aux besoins formulés par Langran. Ce modèle semble être une solution viable pour le traitement de l'évolutivité de l'information statistique.

Il manque cependant dans ce formalisme une conceptualisation des relations qu'entretiennent les unités territoriales et qui sont au cœur de notre problématique : les relations hiérarchiques, de proximité et de généalogie. De plus, la maintenance d'un tel modèle, orienté-objet et bâti sur le paradigme identitaire, pose problème : elle nécessite en effet une définition de l'identité selon des critères objectifs, pour pouvoir décider du rattachement ou non d'un nouvel atome à un objet déjà existant. Ce problème est d'ailleurs largement discuté par Worboys ultérieurement, [Worboys 05]. La question que soulève la définition de l'identité n'est pas complètement élucidée, est reste le défi à relever dans le cadre d'une conception orientée-objet de formes vectorielles changeant au cours du temps, comme l'explique Yuan, [Yuan 96] :

> « Cependant, le plus grand défi dans la modélisation spatio-temporelle est de maintenir l'identité des objets spatiaux durant leurs évolutions (affectant leurs propriétés géométriques, comme leurs relations topologiques). » [15]

15. Traduction libre de : « However, the real challenge in GIS temporal modeling is to maintain spatial objects' identities throughout the evolution in geometrical properties and topological relationships. »

La question de l'identité ralentit la diffusion d'un modèle basé sur le paradigme identitaire, comme le montre les revues de l'ensemble des modèles existant à ce jour, [Gregory 02], [Albrecht 07].

2.2.2.3 La question de l'identité

Définir la nature d'un objet, son identité intrinsèque, est un sujet délicat, qui relève de la philosophie. Dans le cadre d'une approche ontologique de la capture de la réalité du monde, Smith, [Smith 96], argumente que la définition d'un objet doit être suffisamment rigide pour pouvoir l'enregistrer (c'est-à-dire les décrire de la façon la plus objective possible, indépendamment de la perspective depuis laquelle l'objet est traité), mais aussi suffisamment souple pour permettre aux objets d'être traités suivant plusieurs perspectives, et autoriser des transformations des objets au cours du temps.

Dans le cadre de notre problématique, celui de l'identification des parcelles d'un zonage, nous serions tentés de fournir une réponse simple. Une hypothèse fréquente est de placer le traceur de l'identité sur l'un des attributs, choisi en fonction du contexte de l'étude. Par exemple, Worboys ([Worboys 92]) place ce traceur sur le type d'usage du sol. Les travaux de Chareille, Rodier et Zadora-Rio [Chareille 04], qui visent à reconstruire l'histoire des communes de Touraine sur la période [1791-1999], veulent démontrer que le centre des paroisses (l'église) serait le meilleur identifiant des objets étudiés, pour des raisons culturelles. Très souvent aussi, l'empreinte spatiale est prise comme marqueur de l'identité et parfois de façon implicite. Ainsi, sans faire référence explicitement au concept d'identité, Kauppinen, [Kauppinen 07], dans son travail sur la description de l'histoire des territoires à l'aide d'ontologies historiques, définit un territoire via son empreinte spatiale, bien qu'il ne dispose pas forcément des délimitations précises des communes qu'il étudie. Cependant, suivant les types de zonage considérés, le point de vue sur l'identité varie et il arrive que la continuité de la structure ne puisse être définie par aucune règle. Par exemple, dans les découpages administratifs, comme ceux qu'étudie [Ben Rebah 08] en Tunisie, ces découpages sont provoqués par une organisation politique qui décide arbitrairement des changements, décrits dans des textes juridiques. Ces changements, où la vie ou la mort d'une entité est prononcée arbitrairement, peuvent donner lieu à des transformations, des apparitions ou des disparitions d'unités géographiques, sans qu'aucune règle puisse permettre de déduire systématiquement l'évolution d'une unité à partir de l'étude de ses attributs.

ll apparaît que l'usage d'un unique attribut de l'unité géographique comme marqueur de l'identité n'est pas satisfaisant. Supposons, par exemple, que le nom de l'unité géographique soit son traceur d'identité. Dans ce cas, lorsque le département français numéro 22 nommé 'Côtes du Nord' change d'appellation en 'Côtes d'Armor', il faudrait considérer que le département 22 disparaît au profit d'un nouveau département 22 ayant les mêmes attributs, sauf pour le nom. Cette solution est peu satisfaisante, car en fait le département 22 a seulement changé de nom. L'empreinte spatiale peut sinon être définie comme traceur de l'identité. Le problème est aggravé pour deux raisons :

- (i) un département peut s'étendre, tout en restant le même département (nom et statut juridique) et c'est certainement l'évolution de cet objet que le système doit permettre d'étudier. Par exemple, lorsque Saint Priest, commune de l'Isère avant 1967, change de département d'appartenance pour être incluse dans le Rhône, les frontières de l'Isère et du Rhône changent sans que leur identité ne soit impactée.
- (ii) les niveaux de généralisation diffèrent souvent légèrement entre deux versions d'un même zonage, et amènent à observer des empreintes spatiales différentes pour une unité qui n'a pas changé.

La seconde raison conduit à étudier les solutions proposées dans le cadre de l'intégration de données géographiques représentées à plusieurs niveaux de généralisation qui sont en marge des travaux concernant la modélisation d'entités spatio-temporelles.

En effet, les recherches portant sur l'intégration de bases de données spatiales hétérogènes ([Spaccapietra 02], [Devogele 98], [Sheeren 04]), le traitement de multiples représentations d'un même objet géographique, [Ruas 99]), et la mise à jour de ces objets, [Badard 00]), ont produit des résultats intéressants qui adressent le problème de la reconnaissance d'objets identiques à partir de l'empreinte spatiale. En effet, le processus d'intégration nécessite à la fois :

– d'unifier la sémantique des schémas et des métadonnées,
– mais aussi d'éliminer les objets redondants et de regrouper les parties complémentaires.

Le premier problème est généralement traité par l'introduction d'une ontologie pour faciliter la comparaison et l'alignement des schémas respectifs des données ([Fonseca 03], [Abadie 08]).

Pour résoudre le second problème, il faut passer par une opération d'appariement d'objets en vue de produire des liens explicites entre les objets homologues [Walter 99]. L'*appariement* est défini comme une opération automatisée ou interactive, qui vise à figer les paires (objet de référence, objet à évaluer) afin d'y appliquer les mesures des écarts [Devogele 97]. Afin de réaliser l'appariement de bases de données géographiques, des méthodes innovantes de comparaison des attributs ont été proposées, en particulier pour l'attribut spatial. En effet, l'attribut spatial est sujet à des variations de forme suivant le niveau de détail, (qui correspond aussi au niveau de généralisation de l'objet), et l'objectif de la représentation, [Ruas 99]. Dans ces conditions, il semble indispensable de produire une description (ou mesure) des formes en vue de les comparer. Ceci reste une tâche complexe, de nombreux types de mesure ont été élaborés [Cauvin 76], [Mustière 01] : l'aire, la périmètre, l'élongation, le niveau de convexité, etc. Les travaux de Devogèle, [Devogèle 02], situés dans le domaine de la géométrie algorithmique, proposent par exemple une comparaison des empreintes spatiales basée sur la mesure de la distance entre deux objets. Devogèle utilise une distance qui possède des propriétés intéressantes pour comparer la forme d'objets entre eux : la distance de Fréchet. Cependant, le coût du calcul de l'algorithme du calcul de distance de Fréchet sont des inconvénients majeurs à leur emploi. Les travaux de Bel Adj Ali et de Vauglin ([Bel Adj Ali 01], [Bel Adj Ali 99], [Vauglin 98]) peuvent également être cités, car ils établissent une liste exhaustive de mesure de distance entre polygones. Ils font par ailleurs remarquer que la distance surfacique est une distance qui se révèle suffisamment robuste dans un nombre important de cas : elle consiste à mesurer le rapport de l'aire non commune de deux surfaces à l'aire de leur union. Si ce rapport est suffisamment petit, alors les deux surfaces sont considérées égales.

Les travaux d'Ana-Maria Olteanu-Raimond, [Raimond 07], [Olteanu-Raimond 09], proposent également une technique pour l'appariement sur comparaison d'attributs qui a fait ses preuves pour l'appariement d'entités par comparaison d'empreintes spatiales dans le cadre de cadastres urbains. Cette technique est basée sur la théorie des fonctions de croyance de Dempster-Schafer, [Dempster 67], [Shafer 76] et permet de prendre en compte l'incertitude sur les valeurs des données : il s'agit de modéliser le degré d'appariement entre deux objets en fonction de différents critères. Chaque critère comparé produit une probabilité de ressemblance, et l'appariement est une composition de ces probabilités effectuée par une fonction de croyance. Les critères proposés ici portent essentiellement sur la géométrie d'un bâtiment (forme, taille, et différents indices de la mesure de la géométrie), mais la catégorie du bâtiment considéré est également utilisée : un niveau sémantique est donc incorporé dans les critères de comparaison. Ces travaux pourraient être réutilisés dans le cadre de la gestion de l'identité.

2.2.3 Des modèles pour répondre à la question du Comment ?

Langran, [Langran 92], fait remarquer qu'une description précise de la nature des changements, de leur localisation et de leur datation dans un espace d'étude géographique est au coeur d'un SIG temporel. Elle suggère ainsi que les *états*, les *évènements*, et les *indices* sont les trois principales entités d'un SIG temporel. Les indices (ou preuves) permettent de découvrir les changements et d'en donner une mesure. Les états décrivent la distribution spatiale des phénomènes géographiques. Les évènements sont les causes de transformations de l'espace géographique (comme les inondations, ou les feux de forêt). Ainsi, la modélisation des évènements est nécessaire mais également complémentaire de la modélisation des états de l'espace géographique.

Les travaux de Renolen, [Renolen 96], s'inscrivent dans cette perspective : il modélise aussi bien les états que les processus, qui représentent le changement, au niveau des attributs, de l'emprise spatiale ou bien de la topologie des objets. Renolen s'appuie sur une typologie des processus de changement pour construire l'histoire de tous les objets du système d'information. Il combine les états et les processus dans un même graphe, le graphe historique. L'originalité de cette proposition tient au fait qu'elle peut rendre compte de changements graduels. Cependant, les zonages territoriaux qui servent à la collecte de statistiques sont des objets fictifs, (des *fiat objects* comme l'explique [Smith 94]), qui sont transformés par décret, à un instant donné. Aucun changement graduel n'est à modéliser ici : ce sont des processus de durée nulle, qui provoquent une discontinuité dans la forme de l'espace.

2.2.3.1 Modélisation des processus de changements territoriaux

Les travaux conceptuels de Hornsby, [Hornsby 98], sur les changements des entités géographiques, basés sur l'identification des entités et de leurs composants, concrétisent cette modélisation du changement à l'aide d'un langage graphique, le Change Description Language (CDL), qui produit des « graphes historiques ». L'idée directrice est que les objets géographiques possèdent une identité, indépendante de la valeur des attributs thématiques ou spatiaux, et que ces objets sont des compositions ou agrégations d'autres objets géographiques. Un suivi historique de la vie de d'un objet est rendu possible par l'usage du CDL qui exprime les processus de création, suppression ou bien de composition des objets qui le composent. La figure 2.21 illustre les primitives du langage, qui définissent trois états pour un objet : un objet inexistant sans histoire, un objet existant avec histoire, et un objet existant sans histoire. Par le terme histoire, les auteurs entendent 'en vie'. Par exemple, les unités géographiques qui ont existé puis ont disparu, comme la Tchécoslovaquie entre 1918 et 1992, sont considérées dans ce formalisme comme des unités existant sans histoire [16]. Alors que la France est une unité qui existe avec histoire. Les modifications temporelles sont représentées qualitativement, suivant l'ordre temporel du déroulement des évènements : le passage d'un état à un autre (ou transition) est représenté par une flèche, et la transition est supposé directe, sans hypothèse sur sa durée.

Ce travail propose une typologie des transitions très exhaustive, qui peut sembler un peu superflue, et qui cependant reflète la réalité des textes administratifs qui, parfois, décrivent avec subtilité les changements (se référer par exemple aux documents publiés par le ministère des impôts de l'Oregon, aux USA [Oregon 09]). Le langage propose également une modélisation des relations de composition territoriale dans ce cadre. Lorsqu'un objet C est composé de plusieurs parties A et B, sa représentation englobe celle

16. La notation choisie par les auteurs est un peu contradictoire avec la notion de vie ou activité d'un objet spatio-temporel au sens qui a été présenté. Ainsi, de façon intuitive, la Tchécoslovaquie aurait été considérée comme un objet n'existant plus avec histoire.

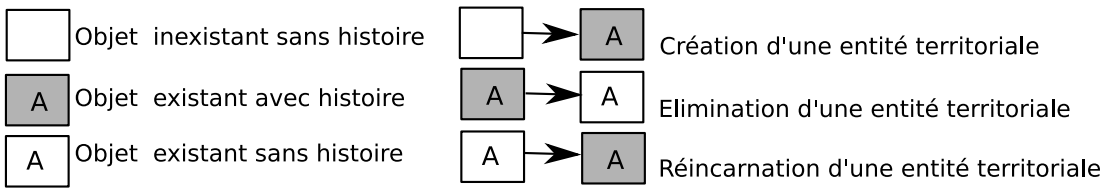


FIGURE 2.21 – Les primitives du langage de description du changement, et trois exemples de transition possible, d'après [Hornsby 98].

de A et B dans un cadre. Dans le cas de le composition, si C disparaît, alors A et B aussi. Sinon, dans le cas de l'agrégation, A et B peuvent continuer d'exister, tandis que C disparaît, comme le montre la figure 2.22.

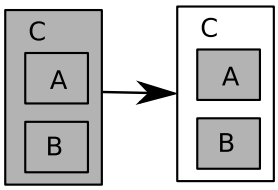


FIGURE 2.22 – Exemple d'objet composite qui est éliminé, tandis que ses parties continuent leur histoire, d'après [Hornsby 98].

Cette modélisation du temps par des changements discrets et ordonnés, ainsi que des relations de composition entre objets géographiques, est intéressante pour la problématique des modifications de zonages organisés dans des hiérarchies. Par exemple, le cas de l'éclatement de l'URSS en Décembre 1991, et de la survie en tant qu'Etats de ses parties (Russie, Lituanie, Géorgies, etc.) est illustré avec la figure 2.22, où l'URSS joue le rôle de C, et ses Etats membres ceux de A et B (plus ceux non dessinés mais faisant partie de C).

Il apparaît que les concepts sur lesquels repose ce langage visuel sont tout à fait pertinents pour notre problématique. Ce formalisme reste cependant à un niveau très conceptuel, et il manque une modélisation plus concrète des attributs thématiques et spatiaux des entités géographiques. Il manque également un formalisme permettant de raisonner avec cette approche. D'autre part, la notion d'identité d'un objet géographique est là aussi assez confuse : est-ce qu'elle tient au nom, à l'emprise spatiale, à d'autres critères, ou n'est définie par aucune règle ?

Thériault et Claramunt [Claramunt 95], [Thériault 99], proposent également une typologie des processus d'évolution qui s'inscrit dans une approche identitaire, et qui est d'un grand intérêt pour notre problématique. Elle procure une règle pour établir une démarcation opérationnelle entre chaque domaine, thématique, temporel ou spatial. Trois classes fondamentales de changement sont déterminées :

- I) l'évolution d'une entité indépendante ;
- II) le changement impliquant des relations fonctionnelles entre plusieurs entités interdépendantes ;
- III) les restructurations territoriales impliquant des unités géographiques.

Les changements liés à la catégorie III sont plus expressifs que ceux des catégories I et II par rapport à la problématique de l'évolution d'un découpage territorial, parce qu'ils associent les entités impliquées dans un même changement, et procurent une contrainte permettant de maintenir l'intégrité d'une base d'unités géographiques dans l'espace et le temps. En effet, cette classe de changement comporte trois opérations : la scission, la fusion, et la redistribution, (voir figure 2.23) qui, en dehors de la stabilité de

forme de l'entité (qui appartient à la catégorie I), sont les seuls changements qui sont étudiés dans le cadre de zonages évolutifs.

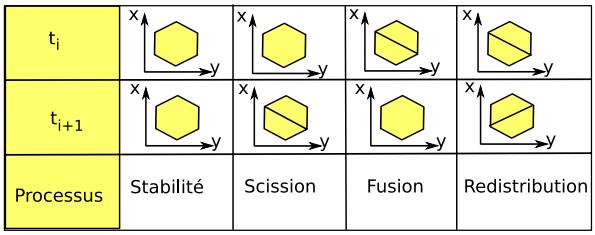


FIGURE 2.23 – Les processus de restructuration territoriale, d'après [Thériault 99].

Spéry, [Spery 01], propose un raffinement de cette classification dans le cadre de la gestion des changements cadastraux en France : il propose de créer un lignage des parcelles en les associant aux évènements de transformation, chaque évènement étant associé au document officiel qui le décrit. Il modélise un espace d'étude qui comprend seulement les parcelles du domaine privé (et donc aucun terrain du domaine public) et distingue les opérations suivantes (voir figure 2.24) :

- fusion : plusieurs parcelles adjacentes sont groupées pour former une nouvelle parcelle,
- division : une parcelle est divisée en plusieurs nouvelles parcelles adjacentes,
- extraction : une parcelle du domaine public est privatisée,
- intégration : une parcelle du domaine privé est intégrée dans le domaine public,
- rectification : un changement de frontière commune intervient entre au moins deux parcelles,
- réallocation : des parcelles sont détruites pour en créer d'autres, et l'union des géométries des parcelles détruites égale celle des géométries créées,
- expropriation : contraction de parcelles adjacentes, au profit du domaine public.

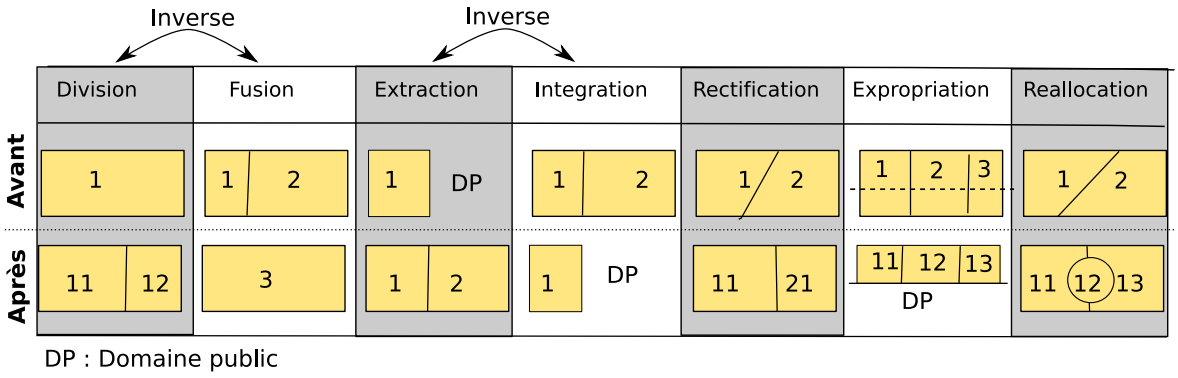


FIGURE 2.24 – Opérations de remembrement, adapté d'après [Spery 01].

Ce travail présente deux points forts : d'abord l'introduction d'une typologie adaptée aux changements territoriaux, à toutes les échelles géographiques, moins riche mais plus opérationnelle que la typologie de Hornsby. En comparant ce travail à celui des travaux de Kauppinen, [Kauppinen 07], dont l'objectif est de décrire une généalogie des parcelles en utilisant le formalisme déclaratif d'un langage sémantique comme OWL, il apparaît que l'utilisation d'évènements adaptés à la description des modifications territoriale est un avantage. En effet, Kauppinen est très embarrassé par le cas des redistributions territoriales où par exemple trois unités peuvent produire deux unités lors d'une réallocation. Sans un évènement qui permet d'associer les trois unités qui précèdent l'évènement avec les deux unités qui lui succèdent, il est difficile de construire des relations significatives dans la graphe historique des unités.

Le second point fort est l'introduction de métadonnées pour les évènements : l'évènement n'est plus simplement un marqueur temporel, il est un objet avec ses propres attributs descriptifs.

La modélisation des processus de changements en tant que cause des transformation de l'espace géographique est activement défendue ([Claramunt 95], [Wachowicz 99], [Worboys 05], [Galton 04]). Les deux principaux avantages à cette approche sont d'une part la construction d'un lignage (ou généalogie, ou graphe historique) des unités qui constituent le support ([Spery 01], [Renolen 96]), et d'autre part la modélisation du temps comme une dimension à part entière, avec une topologie (un évènement suit ou précède un autre), qui autorise une vision à plusieurs niveaux de granularité de cette dimension. C'est pourquoi cette dernière tendance correspondrait à une vraie intégration du temps dans un espace 4-D (3D pour l'espace et 1D pour un temps linéaire orienté), et est qualifiée de *"space-time geography"* en anglais.

2.2.3.2 Exemples de modèles intégrant des évènements

2.2.3.2.1 ESTDM, [Peuquet 95] Peuquet et Duan, [Peuquet 95] proposent un modèle évènementiel qui attire l'attention de l'observateur sur les évènements qui se produisent dans l'espace géographique et qui l'invite à analyser comment ces évènements modifient l'état de l'espace géographique. Cette démarche ne se contente pas de décrire la succession des états, et vise à expliquer et relier ces changements entre eux. Ce modèle, baptisé *Event-based Spatio Temporal Data Model* (ESTDM), considère des données matricielles (des images) et décrit la succession d'évènements qui se sont produits dans l'espace géographique, et les rattache ensuite aux modifications induites sur les cellules de l'image. Il utilise une ligne de temps sur laquelle sont reportés dans l'ordre chronologique tous les évènements que le territoire d'étude a connu, depuis une époque de référence (temps initial t_0) jusqu'à l'époque actuelle t_{now}. Les évènements ont une époque d'apparition t_i, et on leur associe une liste de n triplets (l, c, v) qui décrit l'ensemble des cellules de l'image impactées : l et c donnent la ligne et la colonne de la cellule, v sa nouvelle valeur. Comme dans le modèle du *PPCD-spatial*, chaque cellule de l'image est la plus petite entité non-modifiable (dans sa forme et sa localisation) de l'espace. À chaque image correspond une thématique particulière (mesure de température, ou nombre d'habitants par exemple). La matrice initiale est construite avec l'image de référence, à laquelle est associé un fichier d'en-tête qui indique le nom d'indicateur étudié, et qui pointe sur le début et la fin de la liste des évènements (cette liste se parcourt dans les deux sens). Chaque évènement est décrit par la liste des cellules modifiées (l, c, v), $j = 1..n$, avec un pointeur sur l'événement précédent et sur l'événement suivant, comme décrit dans la figure 2.25.

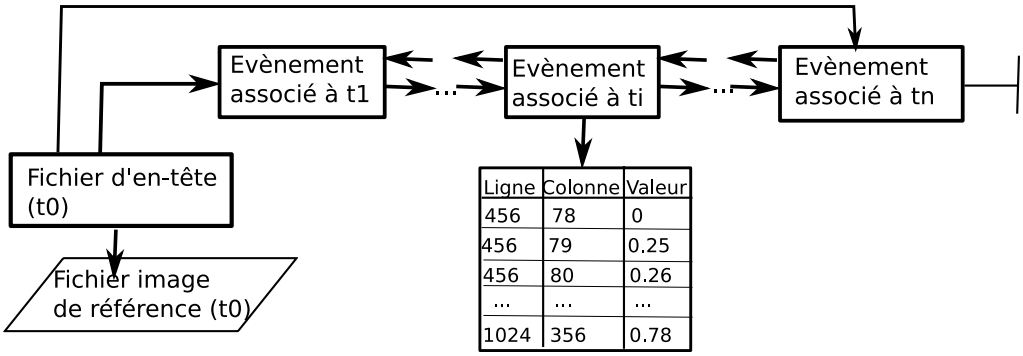


FIGURE 2.25 – Implémentation du modèle ESTDM, d'après [Peuquet 95].

Ce modèle permet de ne conserver que les modifications de l'espace géographique, pour un indicateur donné. Il est efficace pour l'interrogation spatiale et/ou temporelle des données. Mais il présuppose d'avoir construit un maillage de l'espace stable dans le temps. Son adaptation à une représentation vectorielle de l'espace est possible, à condition d'associer un identifiant unique à chaque forme géométrique composant le support, identifiant qui remplacera les index de ligne et de colonne de la matrice image. Lorsque le support spatial des données est instable, ce modèle n'est pas efficace. En effet, une des solutions à l'instabilité du support est de croiser et fusionner toutes les couches vectorielles ou matricielles successives pour composer le PPCD-Spatial. Cependant, l'intégration d'une nouvelle série temporelle pose alors des problèmes de mise à jour des cellules, parce qu'il faut re-indexer toutes les cellules et par conséquent revisiter la liste de tous les triplets associés aux évènements. En ce qui concerne l'espace de stockage que ce modèle consomme, il est idéal pour un seul indicateur, mais lorsque un nombre k important d'indicateurs sont mesurés sur une même grille, la grille est dupliquée k fois. Il serait intéressant d'optimiser l'espace mémoire en reliant différentes chaînes d'évènements distingués par leur thématique à la même grille de référence.

2.2.3.2.2 OOgeomorph, [Raper 95] Raper et Livingston, [Raper 95] proposent un système, OOgeomorph, dont la conception tente d'intégrer les évènements geomorphologiques et les théories attenantes à l'aide de classes dans une représentation orientée-objet. La conception tient en deux modules : le premier module, *geomorph_info*, extrait les données d'une base de données spatiale pour leur représentation et le second module, *geomorph_system*, décrit la dynamique et les théories attenantes à la géomorphologie d'un système (fluvial ou côtier). Les données extraites par le premier module sont transformées en objets géomorphologiques (définis par leur classe, leurs propriétés, leurs méthodes) sur lesquels sont appliqués les traitements définis dans le second module. La localisation (exprimée en trois dimensions, x, y et z) ainsi que la date (t) sont des attributs des objets. Cette approche ressemble beaucoup au modèle orienté-objet de Worboy, [Worboys 92], avec l'avantage de mettre en valeur l'importance des évènements et de règles physiques dans le système d'information. Cependant OOgeomorph manipule essentiellement une information dont la localisation est ponctuelle, il ne supporte pas très bien les données zonales et leurs relations topologiques.

2.2.3.2.3 TGIS [Claramunt 95] Claramunt et Thériault, [Claramunt 95], proposent également un modèle pour la gestion d'évènements sur une base de données spatio-temporelles, le SIG-temporel ou *TGIS* en anglais, adéquat pour des entités de nature surfacique. Ce modèle propose, comme dans les travaux de [Yuan 99], de séparer le temps, l'espace et les attributs, qui évoluent indépendamment les uns des autres. Dans la version basique du modèle, la gestion de la temporalité se fait à l'aide de versions du système, dont les successions représentent le domaine temporel et qui associent les attributs du domaine thématique avec les attributs du domaine spatial. Une version représente, en effet, un instantané, (un état présent, ou passé ou futur du territoire), et les versions sont ordonnées dans le système sur une ligne temporelle. Il n'y a pas dans ce type de modèle de duplication d'information, car si une entité spatiale reste stable, mais la valeur de ses attributs change, les versions successives permettent d'associer la même localisation aux attributs qui changent. La version étendue du modèle intègre la gestion des évènements, qui servent à relier les entités (attributs ou localisations) des instantanés entre eux par des liens de causalité, comme le montre la figure 2.26

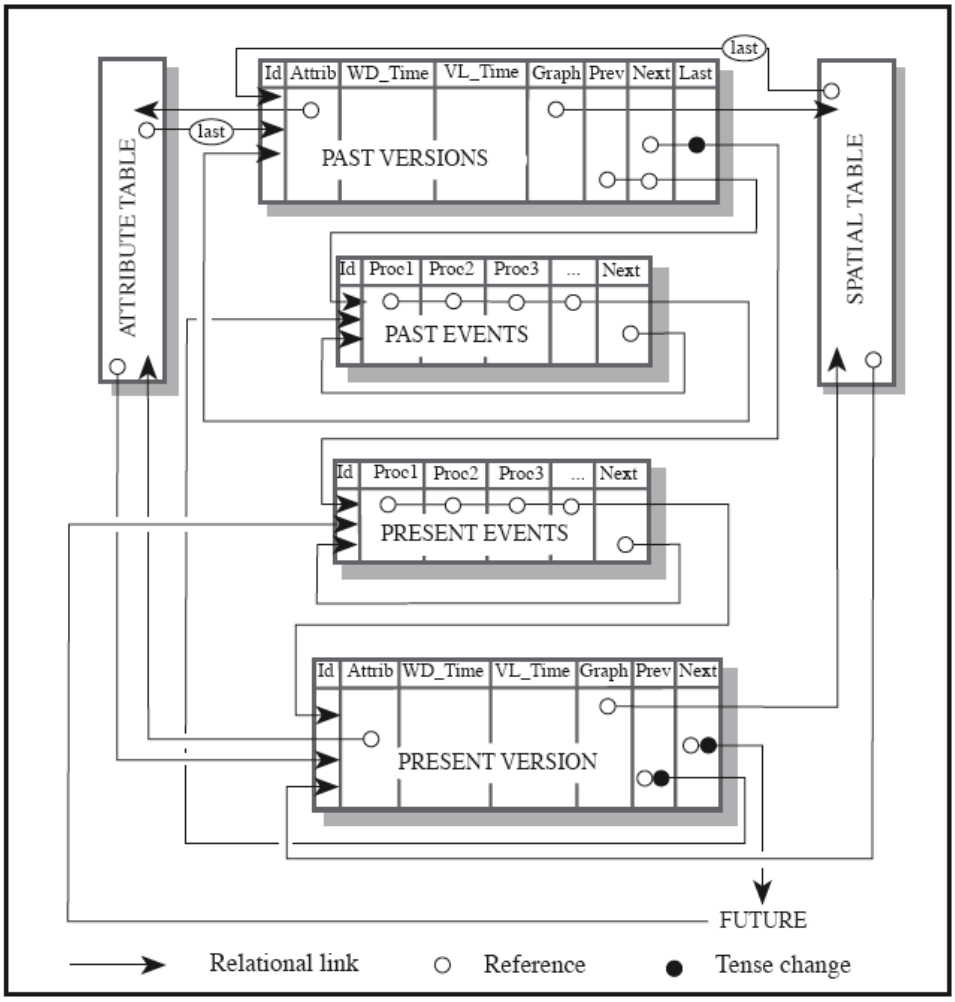


FIGURE 2.26 – Modèle étendu intégrant l'indexation des versions du système par des évènements, d'après [Claramunt 95].

Ce travail propose une double indexation des entités : à la fois temporelle, dans une conception linéaire du temps, mais aussi événementielle. Sur la succession des versions, il propose notamment d'optimiser les requêtes temporelles par l'usage d'un index B+Tree, adapté à la version présente (puisque la validité est marquée par une seule date) ou de R-Tree pour les versions passées ou futures, dont les entités sont datées par des intervalles de validité, se basant en cela sur le travail de [Snodgrass 92]. Ainsi, la formulation de requêtes temporelles sur les versions des entités dans ce système est d'une grande efficacité. Cependant, la valeur ajoutée du modèle est d'associer les versions d'une même entité via des évènements qui sont les liens de causalité entre chaque version. Bien que les processus soient clairement formalisés, la façon d'inférer ces liens entre les versions des entités n'est pas explicitée ou mentionnée. La question de l'identité est donc ici aussi pertinente. En effet, pour un volume de données important, l'acquisition manuelle des liens (ou évènements) entre chaque version du système ne semble pas raisonnable.

2.3 Conclusion

Dans cette conclusion, nous présentons une synthèse des différents aspects de la modélisation spatio-temporelle qui ont retenu notre attention, en relation avec notre problématique.

Nous avons vu que la représentation temporelle dans les SIG passe essentiellement par une modélisation linéaire orientée du temps et utilise les concepts d'instants (des dates) ou d'intervalles (des périodes de durée non nulle) pour localiser et mesurer la durée des entités dans la dimension temporelle. L'importance de conserver à la fois le temps de validité et le temps de transaction, en particulier pour des données statistiques, a été soulignée. Le temps peut également, de façon complémentaire, être modélisé par des événements, des objets de durée non nulle, ayant aussi une emprise spatiale, qui peuvent eux-même être composés de sous-évènements. Leur enchaînement forme une chronique et produit une topologie et une sémantique de la dimension temporelle plus riches que la simple succession de dates.

Par ailleurs, la représentation de l'espace géographique s'est le plus souvent traduite par la projection des entités géographiques (dans leur forme et leur localisation) sur un espace planaire, muni d'une distance euclidienne. Cette représentation a été complètement standardisée dans des normes produites par des organismes comme l'OGC, et sert actuellement de fondement à la représentation de l'information géographique vectorielle dans les bases de données spatiales. Il apparaît cependant que la mesure de l'éloignement (ou réciproquement de la proximité) entre des lieux géographiques n'est pas rendue de façon satisfaisante par la distance euclidienne qui ne tient pas compte de l'anisotropie de l'espace géographique. Il faut donc envisager de représenter les distances géographiques (ou « réelles ») dans un système d'information spatio-temporelle sous la forme d'un graphe, dont les arcs pourraient être typés suivant le type de distance employée (distance temps, distance perçue, distance sociale, ...). De cette analyse émerge aussi la constatation que la modélisation de hiérarchies spatiales, qu'elles soient régulières ou non, est un élément important pour tenir compte des effets d'échelle (et profiter de la propriété d'agrégation de certains attributs thématiques), mais aussi finalement pour modéliser la distance entre unités. En effet, des unités régies par une même unité sont finalement aussi « proches » dans tous les sens du terme (similaires, ressemblantes). Il s'agit donc de modéliser la relation « hiérarchique » entre les unités.

Concernant la modélisation du changement du support des variables statistiques, connu sous le nom de « *Split Tract Problem* », nous avons orienté nos recherches selon deux critères : trouver des modèles qui puissent répondre aux questions (où, quand, quoi ?) de la triade de Peuquet, mais aussi à la question du Comment ? Ces deux perspectives sont différentes, mais complémentaires. Il s'agit, soit de décrire les différents états de l'espace géographique sans tenir compte des logiques de processus ou des évènements qui dirigent cette évolution, soit de fournir une description plus conceptuelle de l'espace, qui formalise les processus d'évolution et les relations spatiales entre entités. La description des processus spatiaux ou des relations spatiales entre entités représente un intérêt certain pour notre objectif car ils donnent à voir le maillage non plus comme un simple objet géométrique support de l'information, mais comme une partition simultanée de l'espace et de la société en sous-ensemble deux à deux disjoints, expression d'un pouvoir qui en régit les limites. Ainsi, les processus qui régissent les transformations du support peuvent eux-mêmes faire l'objet d'étude.

Concernant la description des différents états de l'espace géographique et de l'évolution (ou des mutations) des entités territoriales qui le constituent, il apparaît que la paradigme identitaire permet de répondre efficacement aux trois questions (où, quand, quoi ?). Cependant, il se heurte au problème de l'identification, opération qui consiste à apparier plusieurs versions d'entités au cours du temps pour leur reconnaitre une seule et unique identité. Dans cette recherche, nous nous sommes donc intéressés à la

problématique de l'appariement d'objets géographiques issus de bases de données spatiales hétérogènes, qui peut apporter des éléments de réponse à la question de l'identification. Celle-ci ne peut en aucun cas reposer sur un unique critère, et il nous est apparu que la théorie des croyances, fondée sur la combinaison de critères pondérés par leur degré d'incertitude (ou de croyance), pouvait être utilisée pour l'identification.

Par ailleurs, s'il existe de nombreux travaux proposant une formalisation plus ou moins avancée des événements (ou processus d'évolution), et proposant une implémentation qui prouve son efficacité sur quelques exemples, nous constatons que ces travaux n'abordent pas le problème de la saisie des événements. Il s'agit ici de proposer des méthodes d'acquisition de ce type d'information, ou d'être capable d'inférer automatiquement cette connaissance à partir des données usuellement fournies en entrée du système par un utilisateur. Or, les processus de redistribution territoriale sur des zonages produisent des contraintes topologiques et géométriques qui pourraient éventuellement être utilisées pour calculer un événement de redistribution, scission ou fusion territoriale.

Chapitre 3

Description de l'information statistique territoriale

L'information statistique présente une complexité indéniable puisqu'elle est issue de producteurs différents qui emploient chacun des procédures d'agrégation différentes, qu'elles soient spatiales ou thématiques, avant même de diffuser ces données. S'ajoute à cela le fait que différents organismes habilités à diffuser de l'information statistique territoriale, tels qu'Eurostat ou l'ONU peuvent procéder à des opérations de réajustement et de transformation de données conduisant à des incompatibilités avec les données diffusées par les producteurs nationaux.

Ce problème de cohérence des données statistiques dans le domaine social, économique ou agricole est connu depuis longtemps des statisticiens qui manipulent ces données, et qui, comme [Wilks 39], soulignent le manque d'homogénéité des échantillons, la non-comparabilité et l'inexactitude de ces données.

> « Les conditions et les hypothèses sur lesquelles sont basées les méthodes statistiques sont bien mieux satisfaites du point de vue mathématique par les échantillons d'individus produits par l'industrie ou la science que par ceux produits pour fabriquer les statistiques dans le domaine social, économique ou agricole. »

En dépit des efforts d'harmonisation entre les différents producteurs, il s'avère impossible à ce jour d'obtenir un consensus commun durable sur des définitions, nomenclatures et méthodes de production des indicateurs statistiques. Par ailleurs, concernant les données déjà existantes, il serait impossible de les mettre à jour en fonction de cet hypothétique consensus mondial. En effet, dans l'éventualité où la reprise des micro-données à la source (c'est-à-dire chez les producteurs) en vue de leur agrégation suivant un consensus établi sur des catégories thématiques et des nomenclatures géographiques standardisées soit possible, il sera toujours à craindre que les dates de collecte et les zonages de collecte utilisés ne concordent pas. Pour surmonter ce problème, les métadonnées ont été proposées comme l'élément fondamental du processus d'*intégration statistique*, [Colledge 98], qui vise à produire des données mutuellement comparables (et compatibles). La solution devrait donc passer par l'emploi de descripteurs des données, dans un format partagé par tous (une norme), qui puissent offrir une meilleure compréhension du contenu de ces données : les **métadonnées**.

Cette conclusion est aussi celle à laquelle parviennent différents travaux de recherche, notamment ceux de [Shoshani 82, McCarthy 82, Dean 96, Kokolakis 01], qui recommandent depuis longtemps un usage intensif des métadonnées. Ces travaux ont, par la suite, été relayés par les instances internationales

sous forme de guides méthodologiques à l'intention des producteurs de données, [UN/ECE 95].

Enfin, depuis quelques années, un cadre légal existe en Europe avec la directive INSPIRE [1], publiée par le [Parlement européen 07], qui oblige les producteurs à permettre la découverte des données via les métadonnées. La directive INSPIRE, approuvée par le Conseil des ministres de l'Union Européenne et par le Parlement Européen puis publiée au Journal officiel des Communautés européennes (JOCE) le 25 avril 2007, est entrée en vigueur le 15 mai 2007. Elle vise à favoriser la production et l'échange des données nécessaires aux différentes politiques de l'Union européenne dans le domaine de l'environnement pris dans un sens large. Elle crée plusieurs obligations :

- la fourniture des données selon des règles de mise en œuvre communes,
- la constitution de catalogues de données (métadonnées),
- l'application de règles d'interopérabilité,
- l'accès gratuit aux métadonnées,
- l'accès aux données pour les acteurs réalisant une mission entrant dans le cadre d'INSPIRE,
- les services pour permettre ces accès,
- l'existence d'une organisation adaptée pour s'assurer de la bonne mise en œuvre de la directive.

La directive regroupe ces obligations sous le vocable de « Infrastructure de données géographiques ». L'ensemble de ces obligations devra être réalisé dans le cadre des *normes et standards internationaux*, et du point de vue opérationnel selon les règles de mise en œuvre en cours d'élaboration sous l'égide de la Commission européenne. Il s'agit donc de structurer les métadonnées selon un standard reconnu par la Commission européenne.

La section suivante présente une revue des standards de métadonnées qui seraient utiles pour la description et l'échange de l'information statistique territoriale.

3.1 Usage des métadonnées

Dans cette section, nous examinons les différents standards de métadonnées, pour vérifier s'ils répondent complètement à la problématique de l'intégration statistique. En introduction, nous rappelons ce que sont les métadonnées, et à quoi servent les standards de métadonnées. Nous expliquons pourquoi il n'existe pas un standard unique, mais plusieurs standards.

Les métadonnés sont définies comme « des données sur les données » ou « données qui renseignent sur certaines données et qui permettent leur utilisation pertinente », d'après [Bergeron 92]. Selon l'ONU, [UN/ECE 95], qui complète cette définition dans le cas des données statistiques, les métadonnées doivent répondre à deux besoins. Il s'agit, d'une part de définir le contenu des données (en fournissant des définitions, mais également en décrivant le processus de production des données), et, d'autre part, d'expliquer pour quel usage les données ont été produites. Les métadonnées interviennent lors du processus de diffusion des données entre acteurs autour d'un système d'information.

Suivant le type de données échangées (qu'elles soient relatives au commerce, à la finance, à l'environnement, à la santé, ou la sécurité publique), les descriptions et le vocabulaire employées pour décrire ces données peuvent varier infiniment. Ainsi, pour être utiles, les métadonnées doivent pouvoir être partagées et comprises : il s'agit d'assurer à la fois *l'interopérabilité syntaxique* et *l'interopérabilité sémantique* entre tous les acteurs participant au processus de collecte, production et diffusion des données. C'est

1. `http://inspire.jrc.ec.europa.eu/`

dans cette optique que les normes sont conçues et publiées dans des *registres de métadonnées*.

Un registre de métadonnées est, selon la définition qu'en donne le *Dublin Core Metadata Glossary* [2] dans la version finale du 24 février 2001, un « système de gestion des métadonnées », c'est-à-dire un système formel qui fournit l'information d'autorité sur la sémantique et la structure de chaque élément. Pour chaque élément, le registre en donne la définition, les qualificatifs qui lui sont associés, ainsi que les correspondances avec des équivalents dans d'autres langues ou d'autres schémas. Un registre de métadonnées a les caractéristiques suivantes :

- c'est une zone protégée où seules des personnes habilitées peuvent faire des modifications ;
- il enregistre des éléments qui incluent à la fois la sémantique et les classes de représentation ;
- les zones sémantiques d'un registre de métadonnées contiennent la signification d'un élément avec des définitions précises ;
- les zones de représentation d'un registre de métadonnées definissent comment la donnée est représentée dans un format spécifique comme dans une base de données ou une structure de format de fichier comme XML.

La tenue d'un registre de métadonnées se base généralement sur un référentiel d'éléments de métadonnées qui comporte les types de termes employés par des organismes ayant une communauté d'intérêts.

Il existe une norme pour la représentation des métadonnées dans un registre de métadonnées : la norme ISO 11179, [ISO 04b]. Constituée de six parties, elle explique, en substance, que chaque élément dans un registre de métadonnées doit :

- être classifié dans un schéma de classification (partie 2 de la norme 11179) ;
- être défini par la formulation de règles de définitions de données (partie 4 de la norme 11179) ;
- être identifié de façon unique avec le registre (partie 5 de la norme 11179) ;
- être nommé selon les principes de nommage et d'identification (partie 5 de la norme 11179) ;
- être inscrit selon les règles formulées du registre (partie 6 de la norme 11179).

Il existe plusieurs registres de métadonnées (et de normes associées) en fonction du domaine traité par les données décrites. Le gouvernement des États-Unis d'Amérique prend soin d'appliquer cette norme pour tous ses registres de métadonnées. Par exemple, le tableau 3.1 liste les registres qui se déclarent conformes à cette norme aux États-Unis d'Amérique.

Cependant, l'Europe tarde à mettre en œuvre de tels registres de métadonnées. L'Union Européenne a adopté en 2002 une résolution sur l'utilisation des URI et des métadonnées du Dublin Core, mais ne tient pas de registre de métadonnées conforme aux recommandations de la norme ISO 11179. Le seul registre existant concerne les données environnementales, avec le thésaurus *General Multilingual Environmental Thesaurus* (GEMET) maintenu par l'*European Environment Information and Observation Network* (EEIONet) de l'Agence Européenne de l'Environnement. Un autre cas d'implémentation de registre de métadonnées conforme à la norme 11179 existe au Royaume-Uni pour l'observation du cancer (UK Cancer grid). Finalement, un grand nombre d'organisations (gouvernementales ou non) se contentent du référentiel Dublin Core.

2. `http://dublincore.org/documents/usageguide/glossary.shtml`

Organisation	US National Cancer Institute
Registre	Cancer Data Standards Repository (caDSR)
URL	http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr
Organisation	US Environmental Protection Agency
Registre	Environmental Data Registry
URL	http://www.epa.gov/edr
Organisation	US Health organizations (multiple)
Registre	US Health Information Knowledgebase (USHIK)
URL	http://ushik.ahrq.gov
Organisation	U.S. Department of Homeland Security (DHS) and U.S. Department of Justice (DOJ)
Registre	US National Information Exchange Model (NIEM)
URL	http://www.niem.gov/
Organisation	US Department of Justice
Registre	Global Justice XML Data Model (GJXDM)
URL	http://www.it.ojp.gov/topic.jsp?topic_id=43

TABLE 3.1 – Liste des registres conformes à la norme ISO 11179 aux États-Unis.

3.2 Le Dublin-Core

Le Dublin-Core, créé à l'initiative des États-Unis d'Amérique en 1995, [DCMI 95], est la plus connue des normes de métadonnées. Il s'agit d'une norme généraliste, standardisée auprès de l'ISO en 2003 et révisée en 2009, ISO 15836, [ISO 09]. Ce standard structure l'information en quinze rubriques (ou éléments), qui ont des finalités distinctes. Le tableau 3.2 énumère l'ensemble de ces rubriques, et leur contenu.

Ce standard insiste particulièrement sur les aspects légaux (droits de propriété et d'usage) concernant les données (qui sont en fait vues comme des ressources). Il est utilisé aujourd'hui de façon systématique par les administrations américaines pour le référencement de tous les documents qu'elles produisent. Cependant, cette norme étant généraliste, nous pouvons remarquer l'absence ou le flou qui entoure des éléments très importants pour notre objectif qui est de décrire de façon précise l'hétérogénéité de l'information statistique territoriale, une information à référence spatiale ou temporelle. Par exemple, l'élément Couverture qui retranscrit la couverture spatio-temporelle des données est peu précis. Également, cette norme n'attache que très peu d'importance à la description de la qualité des ressources décrites, et il n'existe aucune section qui lui soit dédiée. C'est pourquoi, nous nous intéressons dans la suite aux normes existantes plus spécialisées.

3.3 Les normes de l'information géographique

L'information statistique territoriale étant à références spatiale et temporelle, il semble naturel d'étudier les normes de métadonnées dédiées aux données géographiques. Parmi celles-ci, nous nous intéressons plus particulièrement à la norme ISO 19115, [ISO 03c], car elle représente une synthèse de

Elément	Contenu
Titre	Titre principal du document.
Créateur	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document.
Sujet	Phrases de résumé, ou codes de classement.
Description	Résumé, table des matières, ou texte libre.
Éditeur	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document.
Contributeur	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document. Chaque contributeur fait l'objet d'un élément Contributeur séparé.
Date	Date d'un évènement dans le cycle de vie du document.
Type de ressource	Genre du contenu.
Format	Format physique du document (CD, livre, page HTML, etc.).
Identifiant	identificateur de la ressource non ambigu : il est recommandé d'utiliser un système de référencement précis, afin que l'identifiant soit unique au sein du site (par exemple les URI ou les numéros ISBN).
Source	Ressource dont dérive le document : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI.
Langue	Langue de la ressource.
Relation	Lien avec d'autres ressources. De nombreux raffinements permettent d'établir des liens précis, par exemple de version, de chapitres, de standard, etc.
Couverture	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle (dates, périodes).
Droits	Droits de propriété intellectuelle, *copyright*, droits de propriété divers.

TABLE 3.2 – Liste des rubriques du standard Dublin-Core, d'après [DCMI 95].

différentes normes qui l'ont précédée, telle que la norme CSDGM, établie en 1998 par le *Federal Geographic Data Committee* (FGDC) aux Etats-Unis. En effet, la norme ISO 19115, publiée en 2003, est issue de travaux internationaux sur le partage des données de nature environnementale, notamment ceux du FGDC aux Etats-Unis, en 1994, et ceux du Comité Européen de Normalisation (CEN) en Europe, avec le CEN/TC 287 ENV 12657 en 1997-1998. Les travaux ont ensuite convergé en 2003 vers cette norme spécialisée, ISO 19115, qui est aujourd'hui recommandée par la directive INSPIRE pour la diffusion de données géographiques. Un nombre important de systèmes dédiés à la mutualisation de l'information géographique mettent en œuvre cette norme. Nous pouvons citer les systèmes d'information distribués et partagés qui sont très bien développés en Espagne, [IDEE 08], ou qui émergent en France pour l'information environnementale, [Barde 05]. En s'appuyant sur cette norme, et les outils de catalogage qui la supportent (tels que MDWeb [3], ou GeoNetwork [4], qui sont des outils puissants et reconnus, [Desconnets 07]), ces systèmes d'information publient des catalogues de métadonnées disponibles sur le Web, autorisant ainsi la découverte des données par les métadonnées.

Le standard ISO 19115 inclut les 15 éléments proposés dans le Dublin-Core, et intègre les éléments

3. http://www.mdweb-project.org/
4. http://geonetwork-opensource.org/

proposés par le FGDC dans la norme CSDGM (la version 3 de la norme CSDGM est devenue un profil de la norme ISO 19115). Cette norme modulaire est extensible et peut faire l'objet d'adaptation dans des extensions (ou profils). Seuls les éléments colorés en orange dans la figure 3.1 sont obligatoires : ce sont les éléments du noyau de la norme. D'un point de vue pratique, les métadonnées de la norme ISO 19115 se présentent dans un fichier de données au format semi-structuré XML (*eXtensible Markup Language*), respectant le schéma de la norme publié en ligne sur `http://www.isotc211.org/2005/gmd/`, qui est adjoint au fichier des données.

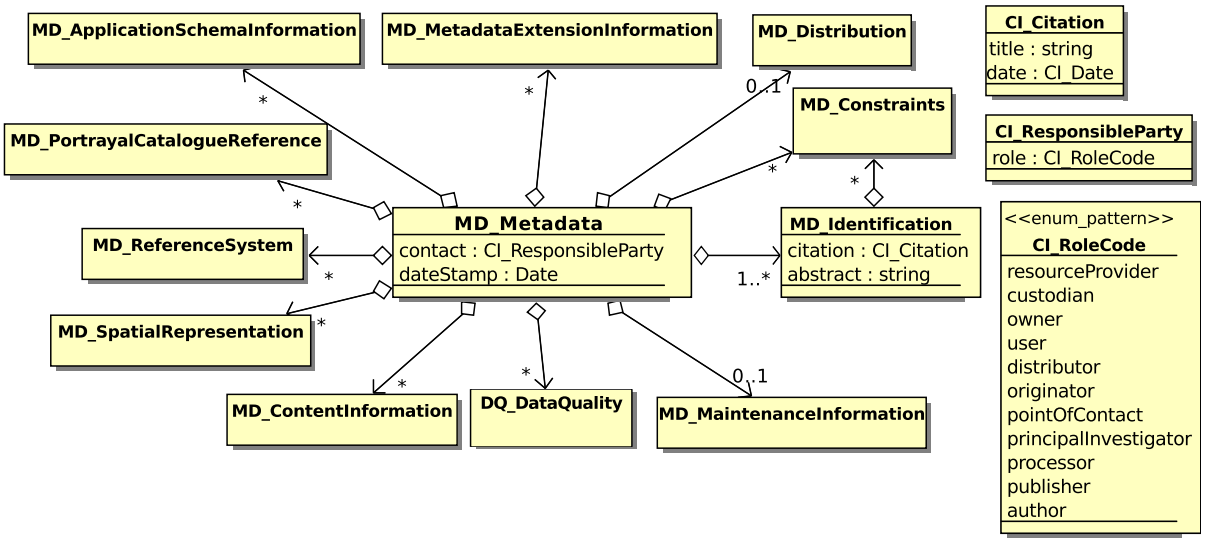


FIGURE 3.1 – Les différentes rubriques de la norme ISO 19115, modélisées d'après le schéma publié sur `http://www.isotc211.org/2005/gmd/`.

Certains projets (comme MEDIS, [Wilde 03], ou la région de la Wallonie [5]) proposent leur propre profil de la norme ISO 19115, adapté à leurs besoins.

3.3.1 Etude approfondie de la norme ISO 19115

Nous décrivons ici de façon détaillée les différentes rubriques, en apportant quelques commentaires concernant leur compréhension et facilité d'édition, que ce soit en général ou bien relativement à l'information statistique.

– **MD_Metadata** regroupe les éléments pour la gestion des métadonnées de la fiche : date d'édition de la fiche (dateStamp), standard, et language utilisé, et un attribut hierarchyLevel qui doit être compris comme un qualificatif du niveau de la métadonnée dans l'arbre des métadonnées. Un contact doit être mentionné, (CI_ResponsibleParty) pour lequel seul le rôle est un champ obligatoire. La description des différents rôles (voir CI_RoleCode, dans la figure 3.1) peut faire hésiter entre « originator » et « author » ou bien « pointOfContact ». Par ailleurs, si l'auteur des métadonnées se cantonne à donner ce rôle, l'information est inutile, voire absurde. Pour décrire le contact, des champs supplémentaires sont proposés : ce peut être le nom d'un individu, d'une

5. `http://cartographie.wallonie.be/NewPortailCarto/index.jsp?page=ProfCatalogueGeneseMeta&node=31`

organisation, les deux. Par souci de généricité, la norme ne rend rien obligatoire, et laisse l'utilisateur choisir ce qui lui est le plus adapté. Cependant, ce type de choix répugne en général aux auteurs de métadonnées.

- **MD_Identification** apporte les éléments principaux pour décrire la ressource : un titre, une date de création ou de mise à jour des données, une description textuelle condensée (abstract), l'emploi visé de la donnée (purpose), etc. Toutefois, dans de nombreux exemples de fiches, on constate que purpose et abstract sont confondus.
- **MD_Constraints** définissent les restrictions sur la diffusion ou l'usage des données décrites dans la fiche, comme des métadonnées elles-mêmes. Il est ainsi possible de rendre des données non publiques, mais de laisser connaître leur existence via les métadonnées. Ou bien d'occulter complètement l'existence de ces données en rendant les métadonnées, elles-aussi, non publiques. Le *copyright* est délivré sous forme textuelle, dans le champ (useLimitation) qui n'est pas obligatoire. Dans le même esprit que celui qui prévaut dans le Dublin-core, les contraintes légales (MD_LegalConstraints) peuvent être distinguées des contraintes de sécurité (MD_SecurityConstraints). À ces deux types de contraintes est associée une liste de codes spécifiques, et si les contraintes de sécurité sont clairement hiérarchisées, les codes associés aux contraintes légales sont plus ambigus. Ainsi, l'auteur des métadonnées doit-il s'interroger sur ce qui convient le mieux entre copyright, patent, patentPending, trademark, intellectualPropertyRights, licence, restricted ou otherRestrictions. Le lecteur de ces métadonnées ou celui en charge de leur diffusion, peut, lui, se questionner sur le niveau de diffusion autorisé s'il ne dispose que de ces contraintes légales codifiées de façon ambiguës.
- **MD_Distribution** décrit les modalités de distribution de la ressource : qui contacter (nom, adresse, téléphone du distributeur), où se trouve la ressource, sur quel type de support (numérique ou papier), le format d'enregistrement, et les moyens d'accès (en ligne, par courrier) et à quel prix éventuellement.
- **MD_MaintenanceInformation** doit informer sur la fréquence de mise à jour des données (régulière ou non), l'obsolescence des données. La spécification de la portée des mises à jour se fait à travers l'usage des éléments MD_Scope, et MD_ScopeDescription : un code spécifie le type des éléments concernés par la mise à jour [6] - type qui peut être le jeu de données entier, le fond géomérique, certaines *features*, des attributs, etc. - et l'utilisateur peut ensuite énumérer les éléments. Mais l'énumération n'est pas spécifiquement liée au type des éléments choisi. De plus, un type d'élément peut-être précisé sans énumération, la portée reste donc vague. L'absence d'attributs spatio-temporels pour cet élément est regrettable car de tels attributs permettraient de discerner les parties mises à jour en fonction de critères spatio-temporels.

Les trois rubriques suivantes sont adaptées à l'information géographique, plus particulièrement à l'information issue d'images aériennes ou satellitaires, comme le Corin Land Cover.

- **MD_ReferenceSystem** donne toutes les informations relatives à la description des systèmes de référence spatial (géographique et vertical) et temporel utilisés, le système géodésique, ellipsoïde, etc. Les descriptions s'appuient sur les normes de référence suivantes : [ISO 02a, ISO 03a, ISO 07a].
- **MD_SpatialRepresentation** apporte une description détaillée des représentations vectorielles ou matricielles, en précisant la résolution, les caractéristiques du géoréférencement, et la nature des objets géométriques utilisés.
- **MD_ContentInformation** précise le contenu avec des caractéristiques plus techniques portant surtout sur les images (aériennes ou satellitaires), qui font appel à des notions de télédétection

6. L'intention est ici de décrire à quel niveau de détail se situe l'information [Devillers 04].

(type de capteur, longueur d'onde, dimensions de l'image, etc.).
Dans le cadre général des échanges de données statistiques, ces rubriques ont peu d'intérêt pour les utilisateurs, qui connaissent les unités territoriales par un code lié à une nomenclature, ou une désignation dans leur langue, mais connaissent à peine leur géométrie et peuvent ignorer comment seront cartographiées les données.

Les deux rubriques suivantes sont « difficiles à renseigner et peu claires pour les utilisateurs », d'après [Barde 05].

- **MD_PortrayalCatalogueReference** présente un catalogue des règles de (re)présentation utilisées dans la ressource.
- **MD_ApplicationSchemaInformation** permet de spécifier comment s'organisent les données et les métadonnées dans une structure arborescente de composition des métadonnées.

La rubrique **MD_MetadataExtensionInformation** est une rubrique capitale pour l'extension de l'ISO 19115 car elle permet de spécifier quels sont les nouveaux éléments qui ont été ajoutés au modèle, et de renseigner les propriétés des éléments créés.

Notre intérêt se concentre sur les éléments disponibles pour décrire la *qualité interne* de données spatio-temporelles. La norme ISO 19115 permet de formaliser les informations de qualité à travers l'élément **DQ_Quality**, dont la structure est présentée dans la figure 3.2.

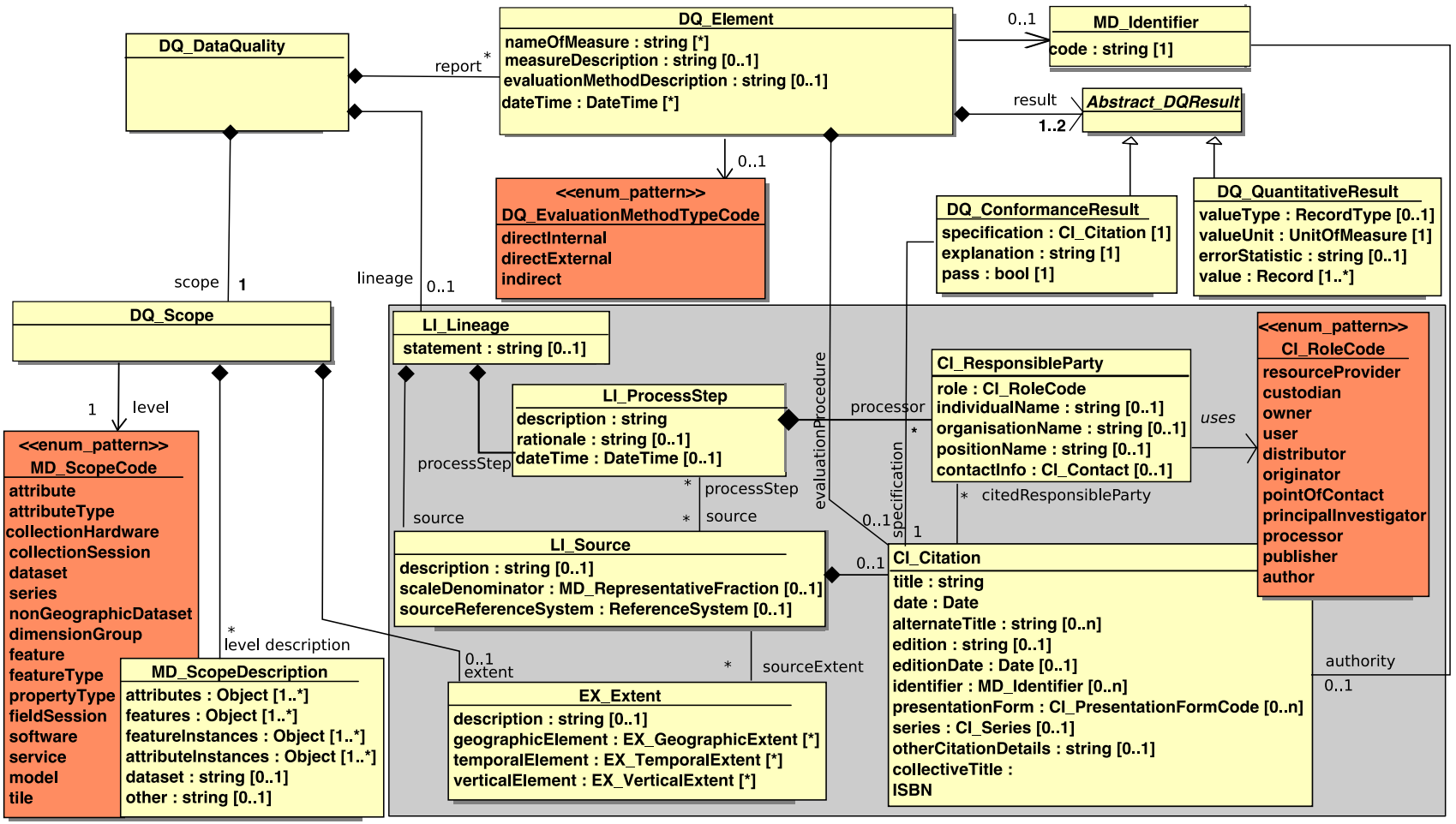


FIGURE 3.2 – L'élément DQ_Quality défini par la norme ISO 19115, modélisé d'après le schéma publié sur `http://www.isotc211.org/2005/gmd/`.

Cette partie de la norme est en fait une expression des besoins de modélisation de la qualité exprimés dans les normes ISO 19113, [ISO 02b], et ISO 19114, [ISO 03b], qui sont des normes essentiellement descriptives. Pour décrire la qualité, la norme se base sur les propositions d'une commission mise en place par le comité exécutif de l'*International Cartographic Association*, en 1991, en vue de documenter la qualité de données spatiales. Ses travaux établissent que la qualité de données spatiales se mesure à l'aune de sept critères : la précision de la position spatiale et temporelle, la précision des attributs thématiques, la complétude, la cohérence logique et sémantique et le lignage. Nous reprenons ici la définition de ces critères, telle qu'elle est fournie dans [Servigne 05].

La précision de la position spatiale (ou exactitude [7] spatiale en réalité) donne le degré de conformité des données par rapport au terrain nominal, en définissant les écarts de valeurs de position respective entre les données du système d'information et le terrain nominal.

La précision temporelle indique si d'une part les données sont à jour et d'actualité, mais également renseigne sur la capacité du système à gérer les versions de données, en distinguant les différentes périodes de validité des données, des dates d'enregistrement des données dans le système. Les informations relatives à la fréquence des mises à jour,

La précision sémantique (ou exactitude sémantique) informe sur les écarts de valeurs des attributs non spatiaux aux valeurs réelles. Elle concerne autant les données quantitatives que les données qualitatives (par exemple, une donnée classée dans la mauvaise catégorie présente un défaut de précision sémantique).

La complétude (ou exhaustivité) se mesure, soit au niveau du modèle de données (est-ce qu'il rend compte de toute la réalité du terrain souhaitée ?), ou bien au niveau des données (objets et attributs). Par exemple, un modèle qui ne prévoit pas de champ numéro de téléphone portable pour enregistrer les coordonnées d'un utilisateur peut présenter un défaut de complétude au niveau du modèle, selon l'usage prévu du système. Au niveau des données, l'absence anormale ou la présence anormale de données sont à vérifier (par exemple, données en doublon).

La cohérence logique vérifie le respect des contraintes d'intégrité définies par le modèle d'information pour la saisie des données. Par exemple, si les données de population sont spécifiées comme des entiers strictement positifs, une valeur négative ne respecterait pas la cohérence logique du schéma.

La cohérence sémantique traduit la qualité avec laquelle les objets géographiques sont décrits par rapport au modèle utilisé, et met en jeu la mesure de la distance sémantique au terrain nominal. Par exemple, la classe « hôpital » comprend-elle les cliniques ? Son évaluation passe forcément par l'étude de l'adéquation des spécifications du système à la réalité que l'on souhaite représenter, et décrit le nombre d'objets, de relations et d'attributs correctement encodés par rapport à l'ensemble des règles de spécification.

Le lignage doit permettre de retracer les procédures d'acquisition, les sources, et les méthodes employées pour transformer les données brutes, et obtenir par dérivation la donnée décrite, [Clarke 95]. Le lignage se représente comme la succession des transformations appliquées aux différentes sources de données dont la combinaison produit la donnée documentée, voir figure 3.3. Idéalement, la description d'une transformation doit permettre de répéter l'opération si besoin ; elle doit donc inclure les paramètres utilisés. À travers ce critère, deux objectifs sont visés : assurer que les méthodes de production respectent les normes en vigueur, et donc s'assurer que les données sont comparables, mais également rassurer l'utilisateur sur la nature des sources employées afin qu'il soit certain que les données répondent à ses besoins.

7. Il faut noter ici une confusion. Précision est pris ici au sens d'exactitude des données, c'est-à-dire de la véracité ou conformité à la réalité - mais le terme précision est devenu couramment et à tort employé dans le sens d'exactitude.

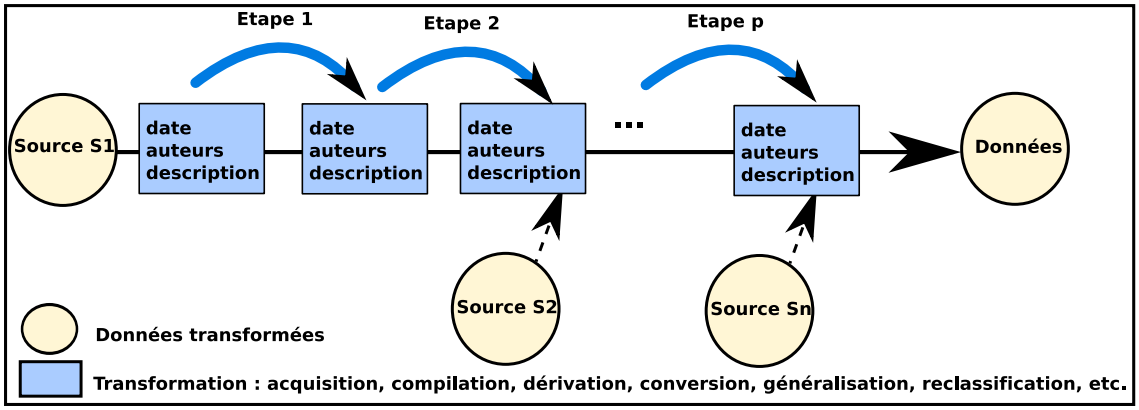


FIGURE 3.3 – Le lignage illustré.

Les informations de lignage (ou de traçabilité) sont conservées dans l'élément **LI_Lineage** de la norme qui utilise de façon conjointe un élément LI_ProcessStep pour décrire une étape de transformation, et LI_Source pour décrire les « ingrédients » de la transformation. L'élément LI_Lineage s'applique à tout ou partie du jeu de données : l'emprise spatio-temporelle des éléments concernés par cette transformation est décrite dans l'élément EX_Extent.

À l'exception du lignage, les critères de qualité sont quantifiables et peuvent donner lieu à des rapports d'évaluation **Abstract_DQResult**. Le résultat de l'évaluation peut-être quantitatif (DQ_QuantitativeResult), ou bien une indication de la conformité des données vis à vis d'une spécification (DQ_ConformanceResult). De plus, comme l'expose [Hangouët 05], tout résultat d'évaluation de la qualité est sans valeur s'il n'est pas associé à un contexte qui décrit le type de méthode employée, le mode d'échantillonage, l'objectif de l'évaluation, la date et le(s) auteur(s) de cette évaluation : ce sont les informations dites de méta-qualité, qui sont regroupées dans l'élément **DQ_Element**. Une méthode d'évaluation de la qualité peut être typée sur la base d'une typologie qui distingue les méthodes directes des méthodes indirectes [Servigne 05]. Les méthodes indirectes consistent à déduire ou estimer une mesure de la qualité des données à partir de métadonnées et plus spécifiquement des informations de généalogie ou d'emploi des données. Les méthodes directes consistent en une comparaison des données avec d'autres données, soit du jeu de données (c'est alors une méthode directe et interne), soit de données externes.

3.3.2 Les limites de la norme ISO 19115

Il s'avère que les échanges de données statistiques territoriales ne se font généralement pas dans le respect de la directive INSPIRE, et, si des métadonnées existent, elles ne suivent pas le standard ISO 19115. L'ignorance de cette norme peut être expliquée pour deux raisons que nous développons ci-dessous : la complexité de la norme et son inadéquation en l'état à l'information statistique.

Une lecture approfondie de cette norme a montré qu'elle est riche, car elle se veut générique, et par conséquent peut-être trop complexe. En effet, il existe un décalage certain entre une recherche qui est en avance de quelques années, et les producteurs, utilisateurs ou collecteurs de données qui peinent encore à comprendre l'objet des métadonnées, et plus encore à les mettre en œuvre. En Europe notamment, ceux-ci ne se sont intéressés à la question des métadonnées que lorsque le projet de réglementation INSPIRE a démarré, dans les années 2005. En revanche, aux Etats-Unis, le Dublin-core, qui existe depuis 1995, est

utilisé à large échelle par les administrations. Sur de nombreux forums ou dans des rapports techniques internes, les discussions portent encore sur les modalités d'interprétation des différentes rubriques et de la mise en œuvre, qui est jugée complexe. Ainsi, la note provisoire d'implémentation des règles INSPIRE (le projet de norme ISO 19139 concernant les métadonnées), publiée en 2007 par [C.E. 07] a suscité beaucoup de doute quant à sa mise en œuvre, [Aisenor 07]. Bien que la norme ISO 19115 ait été officialisée dès 2003, la note d'implémentation définitive vient seulement d'être émise, [Parlement européen 10]. Or, les producteurs de données sont les acteurs désignés pour la production de ces métadonnées : s'ils ne comprennent pas comment remplir certaines rubriques, elles seront négligées. Certaines parties de la norme doivent donc être simplifiées ou mieux expliquées en vue de la rendre opérationnelle.

Cette norme, conçue à l'origine pour documenter des images satellites, ne correspond pas d'emblée au format tabulaire des données statistiques socio-économiques territoriales. La référence spatiale correspond dans ces fichiers le plus souvent à un code d'unité territoriale, dont les limites ne se résument pas à une boîte englobante. Ainsi, le mode de définition des éléments EX_Extent et MD_SpatialRepresentation doit être modifié. De même, l'identification de chaque indicateur doit comporter plus d'éléments que le résumé et le nom de l'indicateur : il faut rappeler son code et son unité de mesure. Enfin, contrairement à un jeu de données géographiques plus classique (des documents géoréférencés au format raster, ou vectoriel, portant sur une thématique unique) chaque valeur du jeu de données possède un lignage spécifique, puisque les valeurs de certaines unités territoriales du jeu de données sont issues de sources spécifiques. Par ailleurs, l'aspect multi-dimensionnel de l'information statistique est difficilement pris en compte car il n'existe pas d'élément prévu pour décrire les catégories, leurs bornes, la propriété de leur domaines respectifs (partition ou bien recouvrement).

Cependant, cette structuration de l'information, bien qu'essentielle, ne permet pas de solutionner entièrement le problème de l'interopérabilité des données statistiques. En effet, si deux indicateurs sont basés sur des définitions différentes ou bien emploient des méthodes de transformations particulières qui rendent les données non comparables, l'information se retrouvera dans les rubriques MD_Identification ou LI_ProcessStep dans un format non structuré, uniquement compréhensible par un humain.

Ce constat nous a amené à considérer des formats comme SDMX, conçu spécifiquement pour l'information statistique, (sans être forcément à référence spatiale). Ce format mixe métadonnées et données dans un même support, et semble proposer une structuration plus poussée de la description des catégories.

3.4 SDMX, un modèle pour l'échange de données statistiques

Les langages semi-structurés ont été avancés depuis quelques années comme une solution au problème de non-interopérabilité entre Systèmes d'Information Statistiques (SIS), [Meyer 04]. La raison principale est que ces langages (basés sur XML et des schémas XSD) permettent d'embarquer une description du format des données dans les données. Ainsi, la flexibilité et la souplesse d'utilisation des données s'en trouveraient accrues. C'est la raison pour laquelle le Statistical Data Model eXchange (SDMX) [8] est aujourd'hui promu par l'OCDE, mais également Eurostat, ou d'autres organismes de production de données statistiques à caractère commercial (et non plus forcément territorial) comme les banques et les organismes financiers. Ainsi, le Fond Monétaire International [9] ou la Banque Centrale Eu-

8. http://sdmx.org/
9. http://www.imf.org/external/index.htm

ropéenne [10] font partie des promoteurs de ce standard. Déjà standardisé en 2005 dans sa première version 1.0 avec la norme ISO TS 17369, [ISO 05], la seconde version de SDMX présentée ici est en cours de standardisation.

En tant que successeur de GESMES (pour *Generic Statistical Messages*) [Kent 97], SDMX est conçu pour faciliter l'échange de données, en précisant l'identité de l'émetteur des données, le format des données envoyées, ainsi que la nature des données employées. GESMES est un standard conçu par Eurostat, dérivé de EDIFACT (*Electronic Data Interchange for Administration, Commerce and Transport*) [11], qui vise à spécifier le format électronique dans lequel les données statistiques et leurs métadonnées devraient être transférées et qui s'est développé dans les systèmes d'information bancaires. GESMES comme SDMX contribuent à promouvoir le développement de systèmes d'information dits « *actifs* », où les métadonnées sont conçues pour traiter les données automatiquement, et non pas uniquement pour que les utilisateurs comprennent les données. Par exemple, il s'agit d'être en mesure de vérifier si des données sont conformes à leurs spécifications (puisque le type des valeurs, et la fourchette de valeurs autorisées sont incluses dans les métadonnées) ou de traduire automatiquement les unités de mesure des données. En effet, entre deux systèmes d'information SI_1 et SI_2, si SI_1 envoie les données exprimées dans l'unité u_1 (l'euro par exemple), et SI_2 veut réutiliser ces données converties dans l'unité u_2 (le dollar par exemple), il faut que :

- SI_2 sache que l'unité postée par SI_1 était u_1 (donc sache dans quel champ de métadonnées trouver cette information) ;
- SI_2 reconnaisse ce que veut dire u_1 (si u_1 est une chaine de texte au format libre, ceci n'est pas assuré).

Ceci implique de s'entendre sur un certain nombre de *concepts* (ou descripteurs) contenus dans les métadonnées, mais également sur la façon de renseigner chaque concept (leur *représentation*). Ce second point n'est pas vraiment assuré dans la norme ISO 19115 pour l'instant car les champs descriptifs sont pour la plupart textuels, et ne font pas l'objet d'harmonisation. C'est un problème d'interopérabilité sémantique [Barde 05].

Cependant, contrairement à GESMES qui est un standard pour les métadonnées à caractère essentiellement « logistique », SDMX présente également un caractère « documentaire » comme la norme ISO 19115. En effet, les concepts et les codes utilisés par les parties s'échangeant l'information pour décrire l'information sont eux-même échangés avec les données ou mutualisés dans des registres accessibles par tous sur le Web, avec leur description textuelle en langue naturelle.

SDMX inclut un modèle de diffusion des données qui repose sur la technologie des services Web et propose aux producteurs de données de s'enregistrer comme « Fournisseur » auprès d'un registre public, et de publier les métadonnées (les fichiers de structure (*Data Structure Definition*, DSD), les concepts et les listes de codes qu'ils auront définis) sur ce registre. Les utilisateurs de données (ayant un rôle de « Demandeur ») consultent ces registres (l'équivalent d'annuaires) pour récupérer l'URL d'accès au service Web d'un producteur délivrant les données souhaitées, ainsi que les modalités de formulation de la requête sur le service, voir figure 3.4. Dans ce modèle, il n'est plus nécessaire que le Fournisseur ou le Demandeur se connaissent, et la donnée est accessible en permanence, tant que le service de délivraison électronique fonctionne.

10. `http://www.ecb.int/home/html/index.en.html`
11. `http://www.unece.org/trade/untdid/welcome.htm`

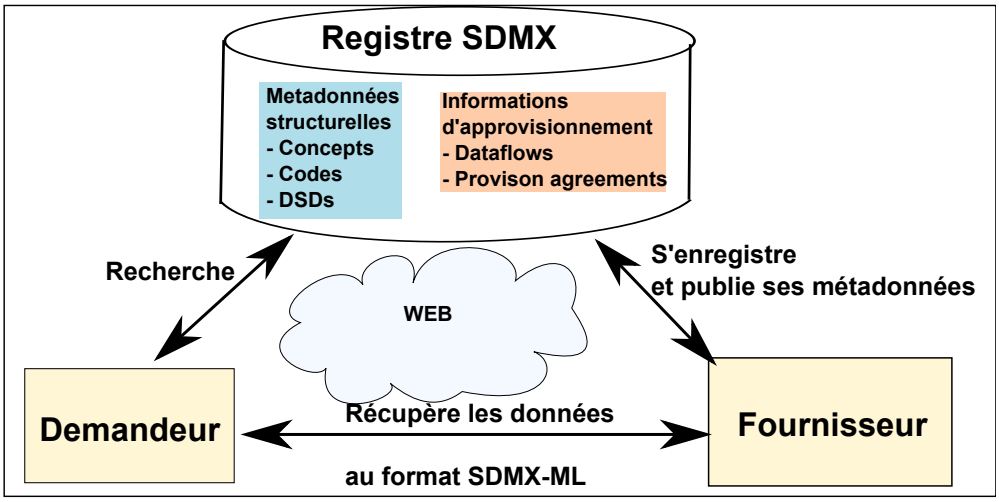


FIGURE 3.4 – Architecture pour l'échange d'information proposée par SDMX.

Les auteurs de SDMX se sont concertés pour définir une liste de concepts [12], termes [13], et codes [14] identifiés et partagés dans un registre, en vue de faciliter l'intégration statistique. Les concepts, référencés par un identifiant unique peuvent être valués, soit par des codes indépendants de la langue, soit par du texte. Par exemple, le statut (OBS_STATUS) d'une valeur statistique est décrit par une liste de codes (A, B, E, F, I, M, P, S) qui signale si la valeur est normale (A), manquante (B), estimée (E), etc., et le registre définit exactement la signification de ces codes. De même, la fréquence de publication des données est définie par un code (CL_FREQ) : A (Annuel), S (Semestriel), Q (Trimestriel), M (Mensuel), W (Hebdomadaire), B (semaine travaillée), D (Journalier), N (Minutes). Concernant les devises d'échange (CL_CURRENCY), ou le format des dates (TIME_FORMAT), la norme se refére aux standards ISO en vigueur : la norme ISO 4217 [15] et la norme ISO 8601.

La liste des concepts et des codes qui leur sont associés peut également être étendue dans le fichier DSD, et être utilisée comme une dimension descriptive de l'information. Par exemple, une catégorie TRANCHE_AGE sera définie comme un concept, avec la liste des différentes tranches d'âge codifiées, décrites chacune par un champ texte. Une observation peut y faire référence. Le fichier de données fait référence à cette description au niveau, par exemple, de la valeur observée, lorsque sont présentées les données :

<Obs TIME_PERIOD="2009" OBS_VALUE="14 702" OBS_STATUS="A" TRANCHE_AGE="15-64" SEX="M"/>

3.4.1 Utilisation de SDMX

L'échange de données statistiques nécessite la définition du format des données dans un fichier externe, le *Data Structure Definition (DSD)*, qui structure l'information à l'aide de ces codes et ces concepts

12. http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf
13. http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf
14. http://sdmx.org/wp-content/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf
15. http://www.currency-iso.org/iso_index/iso_tables/iso_tables_a1.htm

harmonisés [16]. Ce fichier sert à produire une grammaire (le schéma XSD standard) des données que le fichier de données utilise et référence. Les étapes de production de données dans le standard SDMX sont résumées dans la figure 3.5.

FIGURE 3.5 – Etapes de production des données dans le standard SDMX.

Le format prévu pour structurer les données correspond à la hiérarchisation des informations présentes dans les formats tabulaires que s'échangent les acteurs. En effet, plusieurs niveaux de description sont prévus, et les informations rattachées à un niveau inférieur spécialisent les informations décrites au niveau supérieur, la racine de cette hiérarchie étant le jeu de données.
Voici une description de ces niveaux :

1. les observations (*observation*) qui concernent la mesure d'un indicateur sur une unité statistique.
2. les séries (*series*) qui sont définies comme la mesure d'un même phénomène dans le temps, et regroupent les observations d'un même indicateur, mais à des dates différentes ;
3. les groupes (*group*) dont la définition est encore un peu vague mais pourrait s'appliquer par exemple à des séries publiées à des fréquences différentes : journalières, mensuelles et annuelles [17] ;
4. le jeu de données (*dataset*) qui regroupe des séries d'observations d'indicateurs liés à une certaine thématique ;

Le modèle utilisé pour décrire les valeurs statistiques s'inspire du vocabulaire du domaine des entrepôts de données [Rafanelli 90], comme présenté dans la section A.3.2 page 17 du préambule. Ainsi, les valeurs des indicateurs statistiques associées à des unités spatiales sont du niveau *observation* (ou mesure), et se rattachent à des *dimensions*, dont la combinaison permet d'identifier de façon unique une valeur. Les dimensions correspondent aux catégories statistiques, comme le sexe, l'âge, la classe socio-professionnelle, mais aussi l'*unité statistique*. Les dimensions doivent toujours êtres renseignées par des codes, en vue d'assurer l'interopérabilité sémantique, et donc l'interprétation « active » des métadonnées. Le modèle SDMX n'est pas spécifiquement conçu pour l'information statistique territoriale. Ainsi l'unité statistique peut être une entreprise comme une unité géographique ; toutefois, lorsque les unités sont géographiques, SDMX préconise l'usage de nomenclatures géographiques en vigueur, telles que la norme ISO 3166 ou la NUTS pour codifier les unités. À chaque observation est attachée un *Attribut* dont le rôle est purement descriptif, mais qui doit être associé à son concept. Un attribut peut aussi être utilisé à des niveaux supérieurs de description (comme la série, le groupe ou le jeu de données), et la grammaire doit définir son caractère obligatoire ou facultatif. Par défaut, le niveau *series* hérite des dimensions décrites au niveau du *group*.

16. Selon la documentation de SDMX, la grammaire .XSD n'est pas directement exigée car le fichier .DSD est l'expression d'une grammaire considérablement allégée qui facilite le travail de conception au Fournisseur.

17. Voici ce que dit la documentation : « group of series. A well-known example is the sibling group which contains a set of series which are identical except that they are measured with different frequencies ». On peut donc comprendre que les groupes sont une option permettant de regrouper entre elles des séries d'un même indicateur, mais mesurées avec des fréquences différentes. Dans la DSD, ce niveau est facultatif.

3.4.1.1 Exemple

Nous fournissons ici un exemple complet, qui montre à la fois la souplesse de ce format et la complexité de sa mise en œuvre. L'exemple est issu du domaine bancaire : il vise à décrire des données sur le taux de change entre deux devises, données publiées par la Banque Centrale Européenne (ECB en anglais).

Dans un premier temps, le Fournisseur doit décrire la structure de ses données dans un fichier contenant la « *Data Structure Definition* ». De façon simplifiée, la structure d'un tel fichier est la suivante :

- une entête (*Header*) qui rapporte des informations équivalentes à l'élément MD_Metadata de la norme ISO 19115 ;
- les concepts (*Concept*) utilisés pour décrire ces données ;
- l'énumération des listes de codes (*Codelist*) employées pour valuer les dimensions ou les attributs ;
- la structure des informations (*KeyFamily*), précisant à quel niveau chaque *Dimension* ou *Attribut* se rattache.

Nous commentons ici en détail le contenu du fichier qui est en annexe, page 267.

L'entête (Header) (3.1) permet d'identifier cette structure de données, à laquelle on attache un identifiant unique (ID), un nom (Name) ainsi qu'une date (Prepared). Le créateur de cette structure de données (Sender) peut être référencé dans une liste de codes pré-établis dans un des registres SDMX : il est alors identifié par son id dans cette liste.

```
<Header>
  <ID>IREF000506</ID>
  <Test>false</Test>
  <Name>ECB structural definitions</Name>
  <Prepared>2006-10-25T14:26:00</Prepared>
  <Sender id="4F0"/>
</Header>
```

Listing 3.1 – Entête d'un fichier DSD.

Chaque élément Concept (3.2) définit un des concepts utilisés pour identifier et décrire des données.

```
<Concept agencyID="ECB" id="COLLECTION">
  <Name xml:lang="fr">Collection d'indicateurs</Name>
  <Name xml:lang="en">Indicators collection</Name>
</Concept>
```

Listing 3.2 – Définition d'un concept.

Chaque élément CodeList (3.3) définit une liste de codes (et leur signification) qui seront utilisés pour valuer un attribut. Chaque élément CodeList contient au moins 2 attributs :

- l'ID de l'organisme (agency) responsable de cette liste de code (dans l'exemple, "ECB"),
- l'ID de cette liste de code (dans l'exemple, "CL_EXR_SUFFIX").

```
<CodeList agencyID="ECB" id="CL_EXR_SUFFIX">
  <Name xml:lang="fr">Liste des codes associés à la variation des
     taux de change</Name>
  <Code value="A">
```

```
      <Description xml:lang="fr">Moyenne (ou mesure standardisée)
         sur la période</Description>
    </Code>
    <Code value="E">
      <Description xml:lang="fr">Valeur en fin de période</
         Description>
    </Code>
  </CodeList>
```

Listing 3.3 – Définition d'une liste de codes.

L'élément KeyFamily (3.4) définit ensuite la hiérarchie des informations. Ses attributs sont :
- l'ID de l'organisme (agency) responsable de cette structure est obligatoire (dans l'exemple, "ECB"),
- l'ID de ce jeu de données (dans l'exemple, "ECB_EXR1"),
- l'URI de l'espace de nommage de ce jeu de données (dans l'exemple, "`http://www.ecb.int/vocabulary-/stats/exr/1`"),
- le nom du jeu de donnée dans une langue (dans l'exemple, "Taux de change").

```
  <KeyFamily agencyID="ECB" id="ECB_EXR1"
    uri="http://www.ecb.int/vocabulary/stats/exr/1">
    <Name xml:lang="fr">Taux de change</Name>
    <Components>
    ...
    </Components>
  </KeyFamily>
```

Listing 3.4 – Hiérarchisation de l'information.

Dans la balise Components sont spécifiées les informations relevant des quatre différents niveaux d'information avec d'abord la listes des dimensions utilisées pour décrire les valeurs statistiques, puis déterminer quelles dimensions sont au niveau du groupe (Group). PrimaryMeasure identifie le concept qui porte la valeur mesurée pour chaque unité, et c'est par convention le concept OBS_VALUE. Ensuite sont précisés tous les attributs par leur niveau de rattachement, leur concept, liste de code et leur caractère obligatoire ou facultatif.

```
  <Dimension conceptRef="FREQ" codelist="CL_FREQ"
     isFrequencyDimension="true"/>
  <Dimension conceptRef="CURRENCY" codelist="CL_CURRENCY"/>
  <Dimension conceptRef="CURRENCY_DENOM" codelist="CL_CURRENCY"/>
  <Dimension conceptRef="EXR_TYPE" codelist="CL_EXR_TYPE"/>
  <Dimension conceptRef="EXR_SUFFIX" codelist="CL_EXR_SUFFIX"/>
  <TimeDimension conceptRef="TIME_PERIOD"/>

  <Group id="Group">
    <DimensionRef>CURRENCY</DimensionRef>
    <DimensionRef>CURRENCY_DENOM</DimensionRef>
    <DimensionRef>EXR_TYPE</DimensionRef>
    <DimensionRef>EXR_SUFFIX</DimensionRef>
  </Group>
  <PrimaryMeasure conceptRef="OBS_VALUE"/>

  <Attributes>
```

```
        <Attribute conceptRef="TIME_FORMAT" attachmentLevel="Series"
          assignmentStatus="Mandatory" isTimeFormat="true">
          <TextFormat textType="String" maxLength="3"/>
        </Attribute>
        <Attribute conceptRef="OBS_STATUS" attachmentLevel="Observation"
          assignmentStatus="Mandatory" codelist="CL_OBS_STATUS"/>
        <Attribute conceptRef="DECIMALS" attachmentLevel="Group"
          assignmentStatus="Mandatory" codelist="CL_DECIMALS">
          <AttachmentGroup>Group</AttachmentGroup>
        </Attribute>
      </Attributes>
```

Listing 3.5 – Spécification des dimensions - des groupes et des attributs.

Dans ce fichier de structure, seuls les concepts ou les codes qui n'ont pas déjà été définis dans la liste des concepts standardisés proposés par SDMX sont expliqués. Par exemple, les concepts EXPR_TYPE et EXR_SUFFIX sont des concepts particuliers à ces données, non standardisés dans des registres. Pour l'exemple proposé, le tableau 3.3 décrit la structure de l'information utilisée (les concepts retenus pour chaque niveau, et leur domaine de valeur).

TABLE 3.3 – Structure de l'information pour l'exemple proposé.

Niveau	Concept	Représentation	Statut
Group	CURRENCY	CL_CURRENCY	Dimension
	CURRENCY_DENOM	CL_CURRENCY	Dimension
	EXR_TYPE	CL_EXR_TYPE	Dimension
	EXR_SUFFIX	CL_EXR_SUFFIX	Dimension
	DECIMALS	CL_DECIMALS	Attribute
	UNIT	CL_UNIT	Attribute
	UNIT_MULT	CL_UNIT_MULT	Attribute
Series	FREQ	CL_FREQ	Dimension
	CURRENCY	CL_CURRENCY	Dimension
	CURRENCY_DENOM	CL_CURRENCY	Dimension
	EXR_TYPE	CL_EXR_TYPE	Dimension
	EXR_SUFFIX	CL_EXR_SUFFIX	Dimension
	TIME_FORMAT	un des formats ISO 8601	Attribute
	COLLECTION	CL_COLLECTION	Attribute
Observation	TIME_PERIOD	chaîne de caractères conforme à TIME_FORMAT	Dimension
	OBS_VALUE	nombre (nombre de digits définis par DECIMALS)	Dimension
	OBS_STATUS	CL_OBS_STATUS	Attribute
	OBS_CONF	CL_OBS_CONF	Attribute

3.4.1.2 Fichier d'échange SDMX-ML

Le fichier DSD est ensuite transformé en schéma XML standard (dans un fichier XSD [18]) qui est utilisé comme grammaire du fichier de données produit. En effet, les données sont dans un fichier XML (que la documentation de SDMX définit parfois comme fichier SDMX-ML). L'exemple complet est en ligne sur `http://www.ecb.europa.eu/stats/eurofxref/eurofxref-sdmx.xml`, mais nous en commentons quelques extraits dans ce qui suit.

L'entête du fichier (3.6) identifie de façon unique le document (ID), lui donne un titre ("Euro foreign exchange reference rates"), une date de publication ("2006-11-23T08:26:29"), et fournit l'auteur de cette publication ("European Central Bank"). Cette section est très semblable à la rubrique MD_Metadata de la norme ISO 19115.

```
<Header>
  <ID>EXR-HIST_2006-11-29</ID>
  <Test>false</Test>
  <Name xml:lang="en">Euro foreign exchange reference rates</Name>
  <Prepared>2006-11-23T08:26:29</Prepared>
  <Sender id="4F0">
    <Name xml:lang="en">European Central Bank</Name>
    <Contact>
      <Department xml:lang="en">DG Statistics</Department>
      <URI>mailto:statistics@ecb.int</URI>
    </Contact>
  </Sender>
</Header>
```

Listing 3.6 – Entête du fichier de données.

La suite du fichier (3.7) décrit chaque niveau : *dataset* (Dataset) - *group* (Group) - *series* (Series) - *observation* (Obs).

```
<DataSet xmlns="http://www.ecb.int/vocabulary/stats/exr/1"
  xsi:schemaLocation="http://www.ecb.int/vocabulary/stats/exr/1
     ecb_exr1_compact.xsd"
  datasetID="ECB_EXR1">
  <Group CURRENCY="AUD" CURRENCY_DENOM="EUR" EXR_TYPE="SP00"
    EXR_SUFFIX="A" DECIMALS="4" UNIT="AUD" UNIT_MULT="0"
    TITLE_COMPL="ECB reference exchange rate, Australian dollar/Euro
       "/>
  <Series FREQ="D" CURRENCY="AUD" CURRENCY_DENOM="EUR" EXR_TYPE="SP00"
    EXR_SUFFIX="A" TIME_FORMAT="P1D" COLLECTION="A">
    <Obs TIME_PERIOD="1999-01-04" OBS_VALUE="1.9100" OBS_STATUS="A"
       OBS_CONF="F"/>
    <Obs TIME_PERIOD="1999-01-05" OBS_VALUE="1.8944" OBS_STATUS="A"
       OBS_CONF="F"/>
    ...
  <Group CURRENCY="CZK" CURRENCY_DENOM="EUR" EXR_TYPE="SP00"
     EXR_SUFFIX="A" DECIMALS="3" UNIT="CZK"
```

18. En ligne sur : `https://stats.ecb.europa.eu/stats/vocabulary/exr/1/2006-09-04/sdmx-compact.xsd`

```
        UNIT_MULT="0" TITLE_COMPL="ECB reference exchange rate, Czech
            koruna/Euro, 2:15 pm (C.E.T.)"/>
        <Series FREQ="D" CURRENCY="CZK" CURRENCY_DENOM="EUR" EXR_TYPE="SP00"
            EXR_SUFFIX="A" TIME_FORMAT="P1D" COLLECTION="A">
          <Obs TIME_PERIOD="1999-01-04" OBS_VALUE="35.107" OBS_STATUS="A"
            OBS_CONF="F"/>
          <Obs TIME_PERIOD="1999-01-05" OBS_VALUE="34.917" OBS_STATUS="A"
            OBS_CONF="F"/>
          ...
        </Series>
      </DataSet>
```

Listing 3.7 – Corps du fichier de données.

Le jeu de données (DataSet) référence la grammaire qu'il utilise (xsi :schemaLocation="....xsd") décrivant tous les concepts et les valeurs de codes utilisés dans ce fichier. Le premier groupe d'observations concerne le taux de change entre le dollar australien (CURRENCY="AUD") et l'euro (CURRENCY_DENOM="EUR"), en valeur comptante (EXR_TYPE="SP00") exprimé en valeur moyenne sur la période (EXR_SUFFIX="A"), avec une précision de 4 décimales (DECIMALS="4"), et un facteur multiplicateur de 1 (UNIT_MULT="0"). Le second groupe est identique sauf que ce sont les observations pour le taux de change entre la couronne Tchèque (CURRENCY="CZK") et l'euro, et que la précision est donnée avec 3 digits (DECIMALS="3").

Les mesures sont publiées tous les jours (FREQ="D"), et datées suivant le format de la norme ISO 8601 correspondant à P1D (TIME_FORMAT="P1D"). Chaque jour, le taux de change est donc précisé avec sa date (TIME_PERIOD="1999-01-04"), sa valeur (OBS_VALUE="1.9100"), le statut de la mesure (OBS_STATUS="A") - qui est normale ici - et les contraintes de confidentialité associées à la valeur (OBS_CONF="F") - F signifie libre.

3.4.2 Les limites de SDMX

La prise en charge dans ce standard de l'aspect multi-dimensionnel de l'information statistique est donc meilleure que dans le format ISO 19115. En ce qui concerne la définition des transformations, le concept DOC_METHOD s'utilise pour décrire les méthodes de mesure, la définition et les transformations opérées sur les données, et le concept COMPARABILITY pour fournir un commentaire sur l'équivalence de cette donnée avec une autre. Le focus de ce standard porte sur la publication de séries statistiques temporelles. De ce fait, la gestion des révisions de publication et de leur fréquence est mieux assurée que dans la norme ISO 19115 avec un code prévu au niveau de la valeur décrivant le statut de chaque valeur, mais également la fréquence de publication des valeurs ainsi que leur niveau de confidentialité.

Néanmoins, ce format ne résout pas le problème d'hétérogénéité des catégories, comme le rapporte l'expérience à grande échelle menée par [Oakley 05], membres du Bureau Australien des Statistiques (ABS), producteur officiel en Australie, dans le cadre d'un projet visant à exporter au format SDMX les statistiques économiques nationales. Il est rapporté, en particulier, le cas de concepts de SDMX qui ne trouvent pas leur pendant dans la base de données statistiques d'ABS, parce qu'ABS utilise une classification nationale spécifique, qui, par exemple, codifie les activités de pêche, agriculture et d'exploitation forestière avec la lettre A, alors que SDMX classe séparément les activités de pêche (avec le code AYB)

et les activités d'agriculture, chasse et exploitation forestière, codées AYA.

Sur le plan opérationnel, le défaut majeur de SDMX est la complexité de sa mise en œuvre pour des non-informaticiens. En effet, les acteurs du domaine ont encore l'habitude tenace d'échanger les données sous format tabulaire, ce qui ne nécessite pas de trop grandes compétences en informatique, et autorise une lecture facile des données. La mise en œuvre d'un échange d'information au format semi-structuré mêlant données et métadonnées nécessite le développement de supports adéquats pour la lecture et l'écriture des données dans ce format, ainsi que la définition de grammaires appropriées à chaque producteur. Ceci exige une expertise informatique trop élevée de la plupart des utilisateurs.

Par ailleurs, ce standard manque encore de maturité, et la liste des termes codifiés mérite d'être étendue et testée de façon plus poussée. Ainsi, les valeurs de codes n'ont été définies que pour quelques champs, comme CL_OBS_STATUS (statut de la valeur), ou CL_CURRENCY (devise de la valeur associée). Les codes des unités de mesure sont absents, comme les codes des agences officielles de production de données ou le statut civil (ou marital) des personnes.

3.5 Aller plus loin que les métadonnées et résoudre l'interopérabilité sémantique

Dans cette section, nous abordons les travaux relatifs au problème de l'interopérabilité sémantique que ne résolvent pas les métadonnées. En effet, à partir des informations structurées dans le profil de la norme ISO 19115, ou bien le modèle SDMX, il n'est pas encore possible de restituer simplement dans un formalisme exploitable par des automates les processus de transformation agissant sur les données ou bien leur sémantique. Si ces informations sont généralement décrites dans des rubriques structurées, le cœur de l'information est encore décrit par du texte écrit en langage naturel. Ce problème a déjà été souligné dans d'autres travaux [Barde 05], et ceci offre des pistes à explorer pour aller plus loin que les métadonnées et offrir ainsi une méthode pour mieux documenter la qualité de l'information en vue de construire un système d'information statistique plus intégré.

3.5.1 Calcul d'une ontologie de domaine pour résoudre l'interopérabilité sémantique

Face au problème d'incohérence sémantique entre différentes sources de données, un nombre croissant de travaux s'oriente vers l'usage d'ontologies, tels ceux de [Pattuelli 03, Comber 10]. Ces travaux proposent des méthodes pour l'extraction semi-automatisée de métadonnées à partir de textes, mais également pour l'alignement sémantique de données issues de sources hétérogènes. [Comber 05] propose également des méthodes de construction d'ontologies à partir de métadonnées, dans le cadre de l'analyse de données d'occupation du sol.

Les travaux de [Wadsworth 06] et de [Comber 05, Comber 10] étudient plusieurs approches pour l'alignement des différentes catégories d'usage de sol, en vue de comparer des cartes d'occupation du sol produites par différents organismes, à dix ans d'intervalles. Ces catégories, qui sont qualitatives et non-ordonnées, présentent le même niveau d'hétérogénéité que les catégories socio-professionnelles. Parmi les différentes approches étudiées pour leur alignement, il ressort qu'une analyse détaillée de la description de la catégorie (lorsqu'elle comporte plus de 100 mots) par une technique de fouille de données textuelle permet d'établir une matrice de recouvrement entre catégories, que les auteurs démontrent

supérieure aux autres techniques qui nécessitent l'intervention d'experts. La technique consiste à établir la liste des mots employées dans la description de chaque catégorie, puis à calculer leur poids sémantique dans chaque catégorie à l'aide d'une mesure de fréquence inverse dans le document (*Inverse Document Frequency*, IDF, telle que discutée par [Robertson 04]). Le niveau de recouvrement entre chaque catégorie est ensuite calculé via l'usage de la théorie de l'analyse sémantique latente probabiliste introduite par [Hofmann 99], qui stipule que la similarité sémantique entre deux concepts (ici les catégories) peut-être mesurée par la quantité d'information qu'ils partagent (les mots). Cette approche permet également d'identifier les concepts qui structurent une classification données, et les termes qui s'y rapportent. Les auteurs notent toutefois que les concepts identifiés varient d'une exécution à l'autre, et que la méthode doit encore être stabilisée.

Cependant, identifier que des indicateurs se rattachent à un même concept ne résout pas le problème d'équivalence des valeurs qui a été souligné dans la section A.1.5 page 6. L'ontologie sert essentiellement à résoudre et à raisonner sur les problèmes d'équivalences entre catégories. Il s'agit aussi de pouvoir ensuite raisonner sur les valeurs au niveau de leur transformations (modalités de calcul, réajustement, estimation).

3.5.2 Travaux relatifs à la capture d'un lignage

Pour raisonner au niveau des transformations, il faut se doter de formalismes de représentation. La description des transformations en vue de retrouver des données originelles à partir de données transformées est une ambition affichée par [Woodruff 97], qui proposent un formalisme sous forme de graphe (direct et acyclique) du flot de données dans une base de données, chaque nœud modélisant une fonction de transformation, et chaque arc correspondant à une donnée particulière. Le graphe est défini par l'utilisateur, via une interface graphique. L'utilisateur spécifie chaque fonction f de transformation, par son nom, et son type, ainsi que les paramètres d'entrée (des attributs de tuples), par leur type et leur nom, et l'indicateur produit par leur type. Cependant, le dictionnaire des fonctions de transformation proposé est pauvre par rapport à l'ensemble des opérateurs qui sont recensées dans la littérature, car ne sont distingués que deux types de fonction : les agrégats (min, max, compte, moyenne) et les scalaires (produit, quotient, somme, etc.). « Scalaire » est de plus défini dans un sens différent du sens mathématique usuel, et l'adjectif s'applique aux fonctions qui permettent de retrouver une valeur à partir des attributs d'un unique tuple. Ces travaux s'appliquent donc essentiellement sur un modèle de données relationnel, et ces résultats ne s'exportent pas aisément dans un contexte moins structuré. Mais l'idée de structurer sous forme de graphe les relations entre indicateurs, et d'employer une interface graphique nous semble extrêmement pertinente.

Basé sur un formalisme de plus haut-niveau que constitue une ontologie, les travaux de [Brilhante 06] suggèrent également d'associent les indicateurs à leurs formules de calcul dans une base de connaissances. Mais, d'une part, les modalités de construction et d'acquisition de ces opérateurs ne sont pas mentionnées, et, d'autre part, les formules ne sont pas rédigées dans un formalisme mathématique standardisé.

3.6 Conclusion

Face au problème de la variabilité sémantique des données statistiques, il faut envisager de mieux décrire l'information statistique et c'est pourquoi les métadonnées sont considérées depuis longtemps comme une des solutions à ce problème. Les métadonnées sont des données qui renseignent sur la qualité des données, comprise au sens large : elles rapportent par exemple la source des données, le nom du producteur, l'année de production, la définition, la méthode de calcul employée, et les modalités d'échange des données. Les métadonnées interviennent au niveau du flux d'échange des données, et doivent aider les utilisateurs à comprendre si les données correspondent à leurs besoins. Elles sont normalement produites par les producteurs des données. Pour assurer leur compréhension et l'interopérabilité entre les différentes sources de données, des normes définissant la syntaxe comme le contenu des métadonnées ont été établies. Dans d'autres domaines que celui de l'information statistique (pour les données de santé par exemple), l'usage et la production de métadonnées est ainsi devenue courante, et elle respecte ces normes.

Cependant, pour l'information statistique, on note une faible adhésion des producteurs aux métadonnées. Dans ce domaine, le standard SDMX, qui modélise la structure multi-niveau des jeux de données, et permet de prendre en compte l'aspect multidimensionnel des données semble le plus adapté. Avec SDMX, l'objectif est de proposer des métadonnées opérationnelles, c'est-à-dire prévues pour traiter immédiatement les données auxquelles elles sont associées dans un même support. Cependant, pour des producteurs encore peu au fait des métadonnées, décrire et partager leurs données avec SDMX est contraignant parce qu'ils doivent renoncer au format tabulaire des données, et s'ajuster au niveau technologique que requiert l'emploi de langages semi-structurés comme XML. Il existe une autre norme, la norme ISO 19115, initialement prévue pour l'information géographique, et *a priori* la plus adaptée pour l'information statistique en dehors de SDMX. La norme ISO 19115, promue par la directive INSPIRE, permet de produire des métadonnées sans contraindre les utilisateurs à changer le format de leurs données. L'étude de la norme ISO 19115 révèle en revanche certaines difficultés. Par exemple, la signification de certains champs n'est pas toujours évidente et il manque des directives claires pour leur compréhension et leur renseignement. Même comprises, certaines rubriques sont difficiles à remplir car elles mettent en jeu l'écriture et la formalisation de processus complexes, notamment celles qui concernent le lignage des données ou l'évaluation de la qualité des données. Enfin, cette norme ne correspond pas d'emblée à la structure d'un jeu de données, qui présente différents niveaux d'information.

La démocratisation des métadonnées pour l'information statistique passe notamment par des simplifications et une adaptation de cette norme, voire la spécification d'outils appropriés pour la capture et l'édition de ces métadonnées, afin de la rendre plus opérationnelle. Il s'agit aussi de s'interroger sur l'usage qui peut être fait d'une partie de ces informations supplémentaires. En effet, l'interopérabilité sémantique n'est pas encore acquise car de nombreux champs de métadonnées ne sont pas codifiés. Dans ce sens, l'établissement d'une ontologie de domaine statistique pourrait être un apport certain pour l'usage et la compréhension de ces données très hétérogènes. De même, il semble utile de représenter les processus de transformation que les données ont subis, de façon structurée, et des travaux se sont penchés sur la question, sans cependant aboutir à une solution suffisamment générique pour être réutilisée dans le contexte de notre recherche.

Chapitre 4

Analyse de la qualité des données par recherche de valeurs exceptionnelles

Un système intégrant une information statistique territoriale issue de sources hétérogènes peut se transformer en vulgaire boîte à données si ce système n'offre pas des outils pour regarder et comprendre les données. Regarder s'entend ici comme visualiser de façon intuitive les différentes dimensions dans lesquelles s'inscrivent ces données : la dimension spatiale, temporelle, thématique. Par exemple, un indicateur sur la surface moyenne des exploitations agricoles peut se rattacher à une question environnementale d'usage du sol, ou bien à une question économique portant sur les revenus des agriculteurs. Cependant, suivant le pays et l'époque, le rapport entre revenu et surface agricole peut varier, exister ou ne pas exister, ou être lié également au type de production. Comprendre les données implique de replacer chaque chiffre dans son contexte, spatial, temporel et thématique. Cet objectif s'inscrit dans celui d'une discipline établie par Tukey il y a trente ans, [Tukey 77], appelée Analyse Exploratoire des Données (*Exploratory Data Analysis*, EDA). L'EDA se présente comme une philosophie pour détecter et décrire des formes, des tendances et des relations entre les données. Ce domaine fertile a produit de nombreux outils et méthodes pour l'exploration des données, mais sans tenir compte forcément de l'hétérogénéité de ces données.

Nous posons donc ici une autre question. En effet, au regard de l'hétérogénéité entourant les données présentes dans le système, il semble primordial d'évaluer la part de différentiation qui revient réellement à la thématique traitée, de celle qui incombe à la provenance des données, et qui est la conséquence des différentes logiques de production des données. Nous nous intéressons par conséquent à la qualité *interne* des données, telle que la définit la norme ISO 19115, c'est-à-dire de ce qui relève des processus de transformation et de la provenance des données. Les métadonnées permettent de conserver des rapports sur la qualité des données. Il s'agit ici d'étudier par quels moyens ces rapports peuvent être établis, mais également quel doit être le rôle de l'utilisateur par rapport à cette problématique.

Ce chapitre définit en premier lieu le concept de qualité auquel nous nous intéressons, et ce que précisément nous voudrions cerner à travers l'analyse des données. La deuxième section explique les principes qui régissent la recherche de valeurs exceptionnelles, et illustre ces principes avec quelque unes des méthodes les plus connues. La troisième section décrit les caractéristiques des outils qui mettent en œuvre ces méthodes, et propose une critique de ces outils par rapport à notre objectif, celui de l'analyse de la qualité des données.

4.1 Définition de la qualité

Dans cette section, nous définissons ce que nous entendons par qualité des données, et justifions la méthode qui sera proposée, à savoir l'évaluation de la qualité de l'information statistique territoriale par la recherche de valeurs exceptionnelles.

Il s'agit d'abord de cerner précisément ce qu'est la « qualité ». Bien souvent, le terme « qualité » est associé dans l'esprit des utilisateurs à la précision spatiale des données collectées, mais en réalité, le concept de qualité couvre un spectre beaucoup plus large et touche l'ensemble du processus d'acquisition, gestion, diffusion et utilisation de l'information géographique, [Devillers 05]. La qualité est un concept hautement subjectif, car il correspond à l'adéquation entre la vue que donne un système d'information d'une réalité, et la réalité perçue par les utilisateurs [Wand 96]. Il s'agit donc de distinguer la qualité externe, celle perçue par les utilisateurs, qui correspond à l'adéquation des données à leurs besoins (*fitness for use* en anglais) de la qualité interne des données, qui ne dépend pas de l'usage des données. Par ailleurs, la qualité est un concept multi-dimensionnel se prêtant à de multiples interprétations, et de nombreux termes et critères peuvent être définis, comme l'exposent les travaux de [Wand 96] ou [Vaisman 07]. Le tableau 4.1 montre l'ensemble des adjectifs [1] qui relèvent du domaine de la qualité, en distinguant ce qui relève de la qualité interne de ce qui relève de la qualité externe.

TABLE 4.1 – Dimensions de la qualité des données, d'après [Wand 96].

Qualité	Dimensions
Vue interne (orientée système)	**Les données peuvent être** : exactes (*accuracy*), fiables (*reliability*), ponctuelles (*timeliness*), complètes (*completeness*), actuelles (*currency*), cohérentes (*consistency*), précises (*precision*) **Le système peut être** : fiable (*reliability*)
Vue externe (orientée utilisateur)	**Les données peuvent être** : ponctuelles (*timeliness*), pertinentes (*relevance*), satisfaisantes (*content*), importantes (*importance*), suffisantes (*sufficiency*), utilisables (*usableness*), utiles (*usefulness*), claires (*clarity*), concises (*conciseness*), exemptes d'erreurs (*freedom from bias*), instructives (*informativeness*), détaillées (*level of detail*), nombreuses (*quantitativeness*), de portée plus ou moins grande (*scope*), interprétable (*interpretability*), compréhensible (*understandability*) **Le système peut être** : ponctuel (*timeliness*), flexible (*flexibility*), normé (*format*), efficace (*efficiency*)

Ce tableau démontre la difficulté inhérente à la définition et à la mesure de la qualité, car elle se cache derrière une grande variété de qualificatifs. Il apparaît d'abord qu'en réalité tout ce qui concerne la qualité externe est en dehors du contrôle qui peut être effectué dans le système. Par exemple, la précision requise des données ne peut pas toujours être adaptée à l'emploi qui va en être fait : une précision de l'ordre du millier d'unités suffit pour connaître la situation financière d'une entreprise, alors que la réalisation d'un audit financier requiert une donnée précise à la centaine d'unités près. Ainsi, l'usage qui sera fait des données ne peut être systématiquement anticipé dans le système. Ce n'est qu'au niveau de la conception du système que les besoins des utilisateurs (la qualité externe) peuvent être prise en compte. En revanche, tout ce qui relève de la production des données (acquisition, maintenance, et publication) peut être mis sous contrôle, indépendamment de l'usage qui sera fait des données.

1. la traduction anglaise de l'adjectif est fournie entre parenthèses en raison de l'incertitude de la traduction française.

4.1.1 Les critères de mesure

Notre démarche se focalise donc sur la qualité interne, et examine les méthodes de vérification ou d'évaluation de la qualité en reprenant ici les critères qui ont été définis dans la norme ISO 19115. Nous écartons l'usage de méthodes de mesure de la précision géométrique, puisque nous nous concentrons ici sur la dimension thématique de l'information. Egalement, la cohérence sémantique qui relève plutôt de l'adéquation du modèle aux besoins des utilisateurs n'est pas considérée. Par ailleurs, nous sommes dans un cas d'archivage de différentes versions des données, qui peuvent remonter loin dans le temps. Il ne s'agit donc pas de vérifier si les données sont d'actualité, ou ponctuelles, mais bien de vérifier que l'on archive à la fois leur date de validité et leur date d'acquisition dans le système, ainsi que la date de publication de ces données. Restent donc les critères suivants, dont les désignations correspondantes dans le domaine des entrepôts de données sont fournies, [Vaisman 07], en vue de mesurer la qualité des valeurs statistiques :

- la complétude (*completness*)
- la cohérence logique (*consistency*)
- la précision sémantique ou l'exactitude (*accuracy, correctness*)

Pour chacun de ces critères sont détaillées les méthodes de mesures existantes et les travaux qui s'y rapportent.

Un nombre conséquent de travaux visant à qualifier le niveau d'incomplétude s'attachent à considérer les données dans leur ensemble, lorsqu'elles sont conservées dans une base de données relationnelle, et vérifient la complétude des tables et tuples (enregistrements) de la base. Cependant, dans le cas que nous considérons, les données arrivent dans des lots (les jeux de données) destinés à couvrir une certaine aire d'étude sur certains niveaux de la hiérarchie territoriale, pour une certaine période temporelle et pour une certaine thématique. Dans ces cas, il ne faut pas s'attendre à obtenir un remplissage complet du modèle : toutes les unités, ni tous les niveaux ne peuvent être renseignés pour tous les indicateurs existants ni toutes les dates. De même, certaines données sont produites et conservées en doublon, voire plus. En effet, si l'on prend le cas du chômage par exemple, il peut être intégré au moins deux fois dans le modèle, pour la même période et la même aire d'étude car il peut être présent dans des jeux de données utilisant respectivement comme source EUROSTAT, ou bien l'INSEE. Il s'agit donc pour calculer la complétude de définir le nombre de valeurs attendues pour chaque jeu de données, et de rapporter le nombre de doublons ou de valeurs absentes à ces valeurs attendues. Le travail de [Naumann 04] est ici particulièrement pertinent car il propose une mesure permettant de distinguer la complétude dite extensionnelle, qui mesure la couverture d'un jeu de données, de la complétude dite intentionnelle, qui décrit à travers la densité des données dans quelle mesure les intentions annoncées dans le jeu de données ont été réalisées. De plus, les solutions proposées sont adaptées à l'intégration de sources de données hétérogènes.

La vérification de la cohérence logique des données consiste à contrôler la conformité des valeurs vis à vis du domaine de leurs valeurs (par exemple, des valeurs exprimant un poids en kilogrammes ne peuvent être négatives), ainsi que leur typage. Elle consiste également à contrôler certains invariants définis dans le modèle. Par exemple, dans un modèle d'unités statistiques hiérarchiques comme la NUTS, les valeurs statistiques associées aux unités de niveau inférieur devraient toujours être inférieures à la valeur de l'unité statistique à laquelle elles appartiennent. Ce type de travail relève généralement des programmes d'acquisition de données (les ETL, pour *Extract-Transform-Load*) qui sont destinés à nettoyer les données et à les convertir dans le format attendu du modèle de données. Durant cette phase, les données peuvent être rejetées dès leur entrée dans le système, ou bien acceptées, mais signalées comme non cohérentes, et/ou corrigées.

Lorsque les données ont été nettoyées de ce qu'on appelle les « erreurs d'entrée », il reste alors le délicat travail de déterminer les valeurs inexactes qui relève de la vérification de la précision sémantique. Notre cas d'étude concerne la constitution d'une base de données statistiques territoriales issue de sources multiples, et nous postulons qu'il n'existe pas *a priori* de source(s) plus fiable(s) que d'autre(s). Nous orientons donc nos recherches vers les méthodes de vérification de type directes et internes. En l'absence de données de référence, il est difficile de procéder à un nettoyage systématique des données : une donnée peut être exceptionnellement haute ou basse sans être fausse. Dans cette optique, l'identification de valeurs exceptionnelles, celles qui sont très différentes de leur voisinage (temporel, spatial et thématique) peut être utile, car elle ne nécessite pas de données externes, et permet de repérer rapidement des valeurs suspectes, peut-être inexactes.

4.1.2 L'évaluation de la précision sémantique par recherche de valeurs exceptionnelles

Une valeur exceptionnelle (*outlier* en anglais) est définie de façon basique comme une observation qui *dévie* de la valeur *moyenne* de *l'échantillon* dans lequel elle est observée, [Grubbs 69]. À travers cette définition, il ressort immédiatement que la connaissance des opérations statistiques et du vocabulaire afférent (dévier, moyenne, échantillon, etc.) est nécessaire à la compréhension des méthodes de recherche de valeurs exceptionnelles. C'est pourquoi un rappel des notions statistiques les plus élémentaires est présenté en annexe, dans la section 8.2, page 309.

De façon plus générale, une valeur est exceptionnelle si son *résidu* est très *significatif* par rapport à un *modèle*. Donc trouver une valeur exceptionnelle signifie trois choses :

- proposer un modèle de distribution représentatif de l'échantillon,
- calculer des résidus, c'est-à-dire l'écart des valeurs de l'échantillon à ce modèle,
- et évaluer dans quelle mesure cet écart est significatif : il faut donc classer les valeurs des écarts, et les traiter statistiquement.

Deux aspects de la recherche de valeurs exceptionnelles sont à souligner : d'abord, le plus évident qui réfère à l'usage de méthodes statistiques, et ensuite, un aspect non moins important, qui porte sur le rôle de l'utilisateur dans le choix du modèle. En effet, pour qu'un système trouve des valeurs exceptionnelles, il faut que l'utilisateur soumette des hypothèses sur la distribution des données. C'est pourquoi ce type d'analyse ne peut se mener de façon totalement automatique, et nous conduit à étudier les méthodes proposées dans le domaine de l'analyse exploratoire de données (EDA).

Par ailleurs, le recherche de valeurs exceptionnelles est très similaire à l'estimation de données manquantes : en effet, dans les deux cas, il est nécessaire de mettre en place un modèle et de proposer des hypothèses pour estimer la distribution des données. La recherche de valeurs exceptionnelles peut donc être perçue comme le pendant de l'estimation de valeurs manquantes.

Enfin, il est à noter que les méthodes qui permettent de repérer des individus (au sens statistique du terme) exceptionnels ont de façon symétrique besoin de repérer les ressemblances. C'est pourquoi une partie de la présentation qui suit insiste sur la notion de corrélation et d'auto-corrélation. En probabilités et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques numériques, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. L'autocorrélation est une estimation de la corrélation d'une variable en référence à sa localisation dans l'espace ou dans le temps. On estime si les valeurs sont inter-reliées, et si oui, s'il existe une forme de répartition spatiale ou temporelle des individus.

4.2 Méthodes de recherche de valeurs exceptionnelles

Cette présentation, qui n'est pas exhaustive, vise à expliquer les principes de la recherche de valeurs exceptionnelles par l'usage de méthodes statistiques classiques. Nous classons les méthodes de recherche suivant trois dimensions à explorer : la dimension thématique, la dimension spatiale, et la dimension temporelle. L'adjectif « thématique » est à prendre ici dans le sens aspatial, et atemporel, lorsqu'on effectue une étude statistique sur une ou plusieurs variables sans prendre en compte leur localisation spatio-temporelle.

Par ailleurs, nous soulignons que le type de la variable à analyser joue un rôle crucial : une variable quantitative absolue correspondant à un compte d'effectifs sur des unités, (ou *stock*) ne peut être manipulée de la même manière qu'une variable relative (un *ratio*). Par exemple, dans le cas d'agrégation spatiale, les taux des unités ne s'additionnent pas pour constituer le taux de l'unité englobante. En revanche, les taux peuvent être interpolés, et seulement certaines méthodes d'interpolation fonctionnent pour des stocks. De plus, lors de l'étude de la distribution d'une variable issue de comptes sur des unités territoriales, il faut garder présent à l'esprit de toujours utiliser des variables relatives, des ratios, exprimant un rapport à la quantité observée (que ce soit la surface, ou le nombre d'habitants). Dans le cas contraire, des variables associées à des unités de vaste étendue ou très peuplées comptabiliseraient mécaniquement plus d'effectifs (comme le nombre d'usines, de lits d'hôpitaux, de bureaux de poste, de naissances ou de décès), *a priori*, que des variables associées à de petites unités, soit en termes de surface, soit en termes de population. Dès lors, la recherche de valeurs exceptionnelles se réduirait en réalité à une recherche des unités exceptionnelles en termes de nombre d'habitants ou de surface.

4.2.1 L'étude thématique

Nous présentons un certain nombre de méthodes qui visent à trouver des valeurs exceptionnelles, mais ceci sans tenir compte des dimensions spatiales ou temporelles.

4.2.1.1 La boîte à moustaches

La boîte à moustaches (*Boxplot*), inventé par John Tukey en 1977 est une représentation synthétique extrêmement efficace des principales caractéristiques d'une variable numérique X. Elle permet de situer rapidement le profil d'une donnée x en la comparant à des constantes statistiques calculées pour ce même ensemble (médiane, quartiles, minimum, maximum). Elle repose sur le concept de profondeur statistique : la profondeur de demi-espace s'écrit $D(x) = min\{F(x), 1 - F(x^-)\}$, qui est maximal pour la médiane : D(mediane) = 1/2. Pour un quantile d'ordre α, $D(Q_\alpha) = min\{\alpha, 1 - \alpha\}$. La boîte correspond à la partie centrale de la distribution : la moitié des valeurs comprises entre le premier et le troisième quartile. Les moustaches s'étendent de part et d'autre de la boîte jusqu'aux valeurs suivantes : $Q_1 - 1,5|Q_3 - Q_1|$, et $Q_1 + 1,5|Q_3 - Q_1|$. Les valeurs en dehors des moustaches sont considérées comme des valeurs atypiques. Son usage ne requiert pas d'émettre une hypothèse sur la distribution des données, car c'est une méthode non-paramétrique. Les espaces de part et d'autre de la boîte montrent le degré de dispersion des données et l'asymétrie de la distribution.

Le schéma 6.2 illustre son fonctionnement sur un exemple : après un premier tri des données qui détermine la médiane M et les écarts inter-quartiles Q_1 et Q_3, les frontières hautes et basses de l'ensemble

sont calculées. Pour la recherche de valeurs exceptionnelle, toute valeur en dehors de ces frontières est alors considérée comme anormale : dans l'exemple, c'est la valeur 200 associée à l'unité P11 qui est anormalement haute.

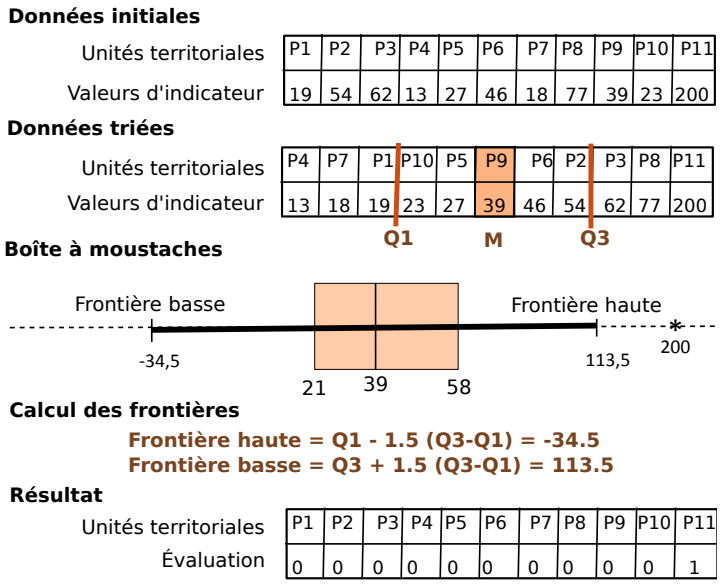


FIGURE 4.1 – Fonctionnement de la boîte à moustaches.

Un autre intérêt de ces diagrammes est de pouvoir faire facilement des comparaisons entre sous-groupes de données car il est plus simple de comparer des diagrammes en boîte que des histogrammes. La figure 4.14 permet de comparer la distribution du Produit Intérieur Brut (PIB) par habitant selon un classement des régions dans une typologie à 6 classes distinguant les régions à forte ou faible urbanisation.

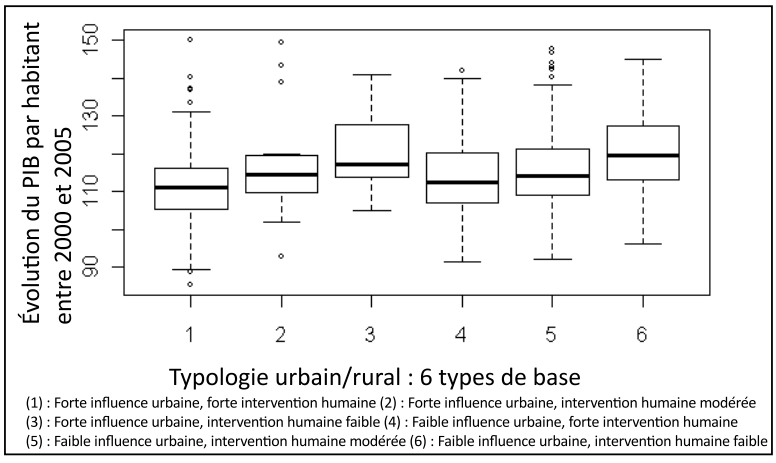


FIGURE 4.2 – Usage de plusieurs boîtes à moustaches pour comparer plusieurs distributions.

4.2.1.2 Les matrices de diagrammes de dispersion

Comme le décrit le chapitre de rappels 8.2 en annexe, page 309, les diagrammes de dispersion sont un moyen efficace de repérer une corrélation éventuelle entre deux variables X et Y. En analyse multivariée, il est courant d'utiliser les matrices de diagrammes de dispersion, voir figure 4.3. Leur représentation n'a qu'un coût nul, et permet de repérer au premier coup d'oeil des formes de corrélation dans les données.

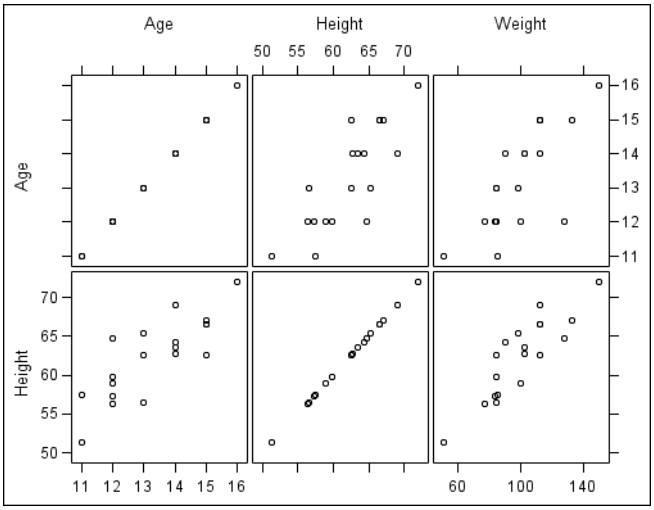


FIGURE 4.3 – Matrices de diagrammes de dispersion.

Par ailleurs, les deux méthodes qui suivent enrichissent ces diagrammes de dispersion avec la construction de frontières permettant de visualiser au premier coup d'oeil des valeurs exceptionnelles.

4.2.1.3 Le bagplot

La méthode *bagplot* généralise le *boxplot* pour un jeu de données multivarié [Rousseeuw 99]. Il s'agit d'abord de déterminer la profondeur de demi-espace, l'équivalent de la médiane en dimension 1. Par exemple, pour déterminer la « médiane » en dimension 2, on va recherche un ensemble (appelé *bag*) contenant 50% des observations. Dans ce cas, la boîte délimitant les observations à l'intérieur des quartiles devient un polygone convexe, et la zone délimitée en clair autour de ce polygone convexe correspond à l'intérieur de la barrière (nommée "boucle"). Les points en dehors de la barrière sont des valeurs exceptionnelles [2], le plus souvent représentée dans une couleur rouge (voir figure 4.4).

Par cette méthode, la distribution des variables X et Y peut être étudiée sans émettre d'hypothèse, car la représentation montre :

- leur situation (le point de profondeur maximale)
- la dispersion des valeurs (figurée par l'aire du bag).
- la correlation (figurée par l'orientation du bag)
- l'asymétrie de la distribution (figurée par situation du point de profondeur maximal par rapport à la boucle)

2. Il a été démontré qu'en dessous de 15 individus, la position de la barrière n'est pas suffisamment stable et la méthode ne permet pas alors de détecter les valeurs exceptionnelles de façon fiable.

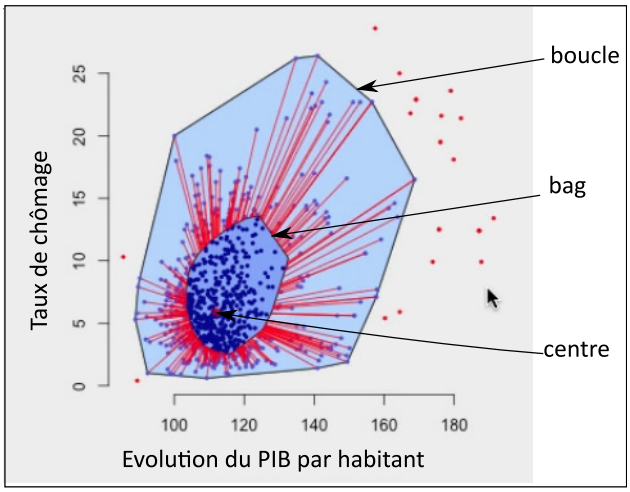


FIGURE 4.4 – Bagplot du PIB et du chômage

– l'importance des queues de distribution (figurée par importance de l'aire située entre la boucle et la limite du bag, et la quantité de points extérieurs à cette frontière).

Dans le cas multivarié, on peut représenter des matrices de *bagplot* en calculant les *bagplot* de chacune des variables deux à deux. Les matrices de *bagplot* sont symétriques, et la diagonale représente les boîtes à moustaches de chacune des variables.

4.2.1.4 La distance de Mahalanobis

La distance de Mahalanobis, introduite par Prasanta Chandra Mahalanobis en 1936 [Mahalanobis 36], est une métrique qui permet de s'affranchir des effets d'échelle (présents lorsque les variables sont mesurées dans des unités différentes, avec des étendues hétérogènes) et de corrélation entre variables. Contrairement à la distance euclidienne, où l'ensemble des points équidistants dans un nuage de dispersion est une sphère, la distance de Mahalanobis étire cette sphère pour respecter les échelles respectives des variables, et prendre en compte la corrélation entre les variables. En pratique, la distance de Mahalanobis d'une série de valeurs de moyenne $\bar{x} = (\bar{x_1}, \bar{x_2}, \bar{x_3}, \ldots, \bar{x_p})^T$ et possédant une matrice de covariance V pour un vecteur à plusieurs variables $x = (x_1, x_2, x_3, \ldots, x_p)^T$ est définie comme suit :

$$D_M(x) = \sqrt{(x - \bar{x})^T V^{-1} (x - \bar{x})}. \tag{4.1}$$

Elle s'utilise pour la recherche de valeurs exceptionnelles dans un ensemble multivarié. En effet, pour des données multivariées avec distribution normale (de type Laplace-Gauss), la distribution des valeurs de la distance suit une loi du khi-deux à p degrés de liberté χ^2_p. Ainsi, en définissant un quantile de cette distribution, à $1 - \alpha$, soit 95% si α vaut 0.05 par exemple, un seuil de test peut-être défini, et l'ellipse définissant la frontière de la distribution a pour équation :

$$\epsilon = \{x : (x_i - \bar{x})^T V^{-1} (x_i - \bar{x}) \leq \chi^2_{p,1-\alpha}\} \tag{4.2}$$

Une valeur supérieure à ce seuil montre que l'observation considérée est en périphérie de nuage, et doit donc être considérée comme une valeur exceptionnelle (voir figure 4.5).

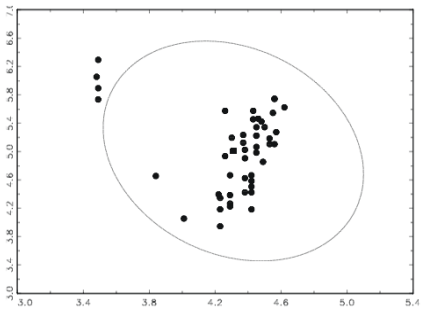


FIGURE 4.5 – Utilisation de la distance de Mahalanobis pour la détection de valeurs exceptionnelles dans une distribution bi-variée.

Cependant, cette mesure est elle-même peu robuste car le calcul de la matrice de variance-covariance est sensible à la présence de sous-ensemble de valeurs extrêmes. Une variante de cette méthode, plus robuste, est présentée dans [Filzmoser 04] avec l'usage d'un seuil adaptatif calculé à partir de la matrice MCD (pour *Minimum Covariance Determinant*). Les estimateurs MCD de localisation et de dispersion sont la moyenne et la matrice de covariance calculées sur l'échantillon de h points parmi n qui minimise le déterminant de la matrice de covariance correspondante.

4.2.1.5 L'Analyse en Composantes Principales

Une autre méthode thématique très classique est l'Analyse en Composantes Principales (ACP) qui propose de déterminer les principaux axes d'une distribution multivariée. L'ACP revient à remplacer les variables $X_1, X_2, \ldots, X_p$ par de nouvelles variables, les composantes principales, $C_1, C_2, \ldots, C_k$ des X_i, non corrélées entre elles, de variance maximale et les plus liées en un certain sens aux X_i : l'ACP est une méthode factorielle linéaire. En pratiquant l'ACP sur des données centrées-réduites [3], les vecteurs C_i sont donnés par la formule 4.3, où les U_i sont les vecteurs propres de la matrice R de variance-covariance des données centrées et réduites (ils sont solutions de l'équation $RU_i = \lambda U_i$).

$$C_i = XU_i \tag{4.3}$$

Le calcul produit ainsi p composantes C_i, chacune associée à sa valeur propre λ_i, mais dont on ne retient que k composantes considérées comme principales (ceci relève d'un choix). En général, la réduction du nombre de variables utilisées pour décrire un ensemble de données provoque une perte d'information. L'ACP procède de façon à ce que cette perte d'information soit la plus faible possible, selon un sens précis et naturel que l'on donne au mot « information ». En fait, on cherche à maximiser le pourcentage d'inertie totale expliquée par ce sous-espace de dimension k, $k \leq p$. L'inertie totale I_g étant la somme des valeurs propres, $I_g = \sum_{i=1}^{p} \lambda_i$, et le rapport $\frac{\lambda_i}{I_g}$ correspondant à la part de variance expliquée par la composante C_i, on cherche à maximiser le rapport 4.4, avec un k le plus petit possible. On retient donc en général les vecteurs associés aux plus fortes valeurs propres.

$$\frac{\sum_{i=1}^{k} \lambda_i}{I_g} \tag{4.4}$$

3. Utiliser les données centrées et réduites a pour conséquence de rendre les distances entre individus invariantes par transformation linéaire séparée de chaque variable, et de s'affranchir des unités de mesure, ce qui est particulièrement intéressant lorsque les variables sont hétérogènes.

L'ACP construit ainsi de nouvelles variables C_k, (avec k que l'on espère très petit devant p), artificielles, et fournit des représentations graphiques permettant de visualiser les relations entre les variables, ainsi que l'existence éventuelle de groupes d'individus et de groupes de variables. Par exemple, dans la représentation dite du « *cercle des corrélations* », les variables X_i sont représentées par des vecteurs dans un repère orthonormé composé des deux composantes principales et les coordonnées des extrémités des vecteurs sont le coefficient de corrélation des variables avec chacune des composantes principales : $(r(X_i, C_1); r(X_i, C_2))$. Cette figure permet de visualiser rapidement quelles sont les variables contribuant principalement à chaque composante, et comment ces variables se comportent en elles (un angle qui tend vers l'angle droit monte une indépendance des variables). Par exemple, sur la figure 4.7, où une ACP a été pratiquée sur des variables décomposant les dépenses des ménages sur différents postes, il apparait que le poste alimentaire est le plus significatif et les achats de viandes sont très corrélés aux achats de légumes, alors que ces dépenses n'ont pas de lien avec les dépenses dans les cantines.

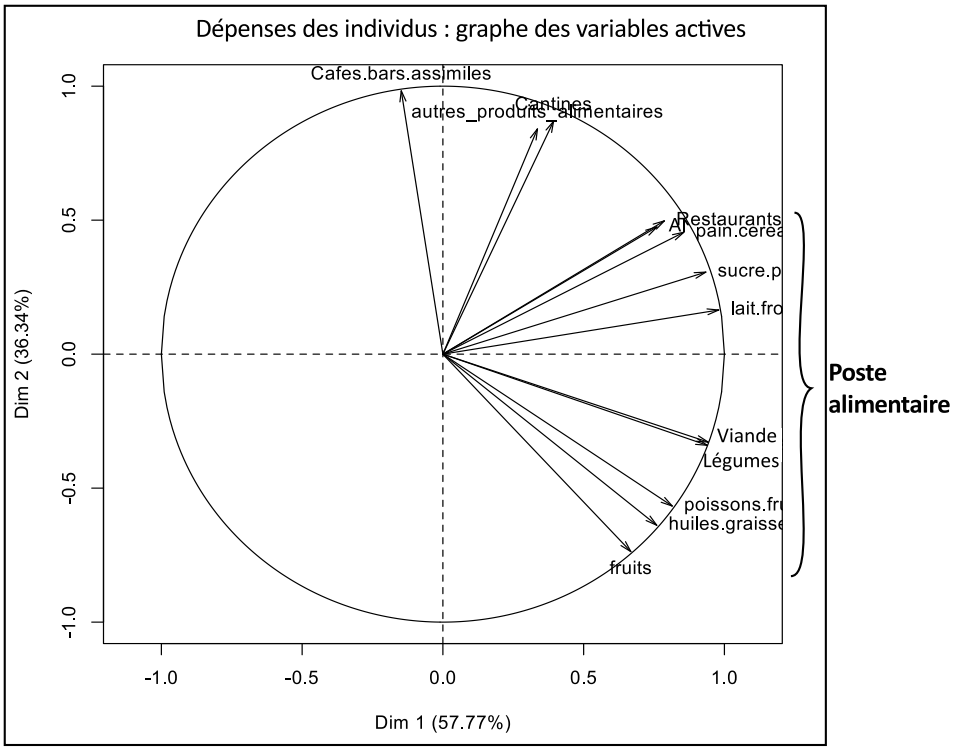


FIGURE 4.6 – Usage du « cercle des corrélations » dans une ACP.

L'ACP est la « mère » de la plupart des méthodes descriptives multi-dimensionnelles qui permettent de détecter des corrélations entre variables dans l'espace thématique.

Mais également, l'ACP s'utilise pour repérer des valeurs exceptionnelles. De l'examen de ces projections, on peut déterminer l'existence et la localisation d'observations exceptionnelles. Une observation est exceptionnelle si elle prend des valeurs extrêmes sur plusieurs variables. Un tel individu j de valeur $(X_{1j}, X_{2j}, \ldots, X_{pj})$ est loin du centre de gravité d'un nuage, et l'on peut évaluer son caractère remarquable par sa distance au centre du nuage dans l'espace complet $\mathbb{R}^p$.

À cette fin, le dessin de *l'ellipse de dispersion d'ordre* k permet de déterminer les individus loin du centre de gravité du nuage. On sélectionne en effet les deux premières composantes C_1, C_2 qui sont deux

variables indépendantes, que l'on renomme X et Y. Pour ces deux composantes, on mesure la moyenne $\bar{x}$, $\bar{y}$, et la variance σ_x, σ_y des individus projetés dans cet espace réduit. L'ellipse de dispersion a pour équation 4.5 :

$$\{x, y : \frac{(x-\bar{x})^2}{\sigma_x} + \frac{(y-\bar{y})^2}{\sigma_y} = k^2\} \tag{4.5}$$

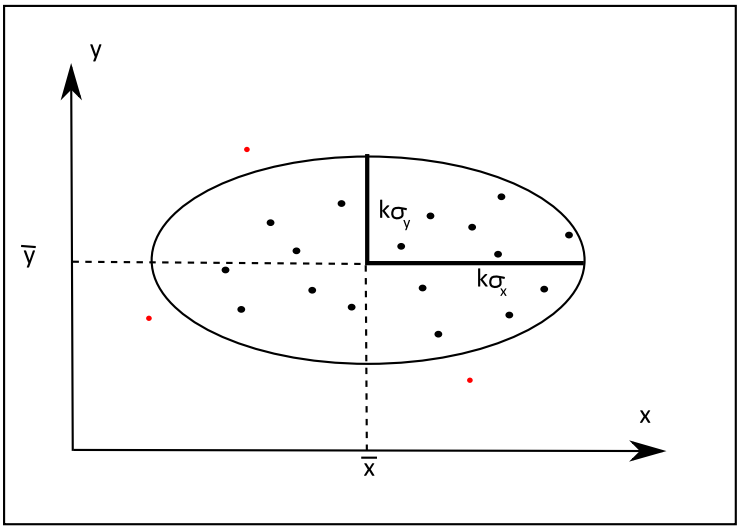


FIGURE 4.7 – L'ellipse de dispersion unitaire (k=1) pour deux variables indépendantes. En rouge, les valeurs exceptionnelles.

Les individus à l'extérieur de cette ellipse sont exceptionnels.

4.2.2 L'étude spatiale

Les méthodes développées en analyse aspatiale ne sont pas adaptées à l'analyse de données spatiales « *because spatial is special* » comme le formule Anselin [Anselin 89].

L'un des premiers supports à l'analyse spatiale est la matrice d'information géographique, introduite par le géographe américain Berry [Berry 68], qui décrit un tableau à trois dimensions (cube) dans lequel les lignes i correspondent aux unités géographiques, les colonnes j correspondent aux caractères permettant de décrire ces unités géographiques, les plans étagés t correspondent aux dates ou périodes pour lesquelles ces attributs ont été mesurés. Un élément quelconque d'une matrice d'information géographique se note X_{ijt} et désigne la « situation du lieu i, pour le caractère j au temps t ». Cette matrice correspond à un tableau d'information géographique, présentée en préambule dans la section A.3.1 page 16, mais intégrant un paramètre temporel. L'analyse des relations entre colonnes permet de découvrir des associations spatiales, l'analyse des relations entre lignes permet d'établir des typologies spatiales, et l'analyse des relations entre plans permet de saisir des dynamiques spatiales. Mais en réalité, en l'état, ce schéma pourrait être réutilisé dans n'importe quelle discipline en remplaçant les lieux par d'autres types d'unités (individus ou groupes en sociologie, firmes ou ménages en économie, etc.).

Dans cette première approche, les variables de localisation des données statistiques sont intégrées comme des variables n'ayant pas d'effet statistique sur les données, et les données sont supposées être indépendantes dans la dimension spatiale. Cette hypothèse de non auto-corrélation spatiale des données

implique que les résidus d'un modèle de distribution d'une variable devrait présenter une forte homogénéité spatiale. Or, les hypothèses portant sur l'indépendance des variables et la non auto-corrélation des données sont fausses, car suivant la première loi de la géographie de Tobler, « *chaque phénomène est relié à tous les autres, mais des phénomènes proches dans l'espace auront tendance à être d'avantage liés que des phénomènes éloignés* », et il est donc fréquent d'observer une auto-corrélation spatiale entre les caractères observés, dont la force dépend de la distance de localisation entre ces caractères.

Pour passer d'une analyse statistique spatialisée à une véritable analyse géographique, il faut donc introduire dans le schéma de Berry une quatrième dimension qui a trait à la position géographique des lieux les uns par rapport aux autres, c'est-à-dire à leurs attributs de localisation et aux relations de proximité que l'on peut en déduire. Cette quatrième dimension peut, dans le cas le plus simple, prendre la forme d'une matrice de pondération (ou matrice de voisinage), notée W, pour *Weight* en anglais, dont les éléments prennent la valeur 1 pour les i, j voisins, et 0 autrement. La notion de voisinage peut être définie soit en termes de relation, soit en termes de distance entre unités spatiales ou en termes de co-appartenance. Le premier cas retient l'aspect topologique du voisinage, le deuxième utilise ses caractéristiques métriques, le dernier cas s'intéresse aux relations de hiérarchie entre les unités.

L'association spatiale ou *l'autocorrélation spatiale* mesure l'intensité de la relation entre la proximité des lieux et leur degré de ressemblance [Pumain 97]. Si la présence d'une valeur forte pour une variable X rend sa présence dans les lieux voisins plus au moins probable, on dira que la variable manifeste une autocorrélation spatiale. L'autocorrélation est positive si les lieux proches ont tendance à se ressembler davantage que les lieux éloignés, elle est négative si les lieux proches ont tendance à être plus différents que les lieux éloignés. Elle est nulle quand aucune relation n'existe entre la proximité des lieux et leur degré de ressemblance, voir figure 4.8.

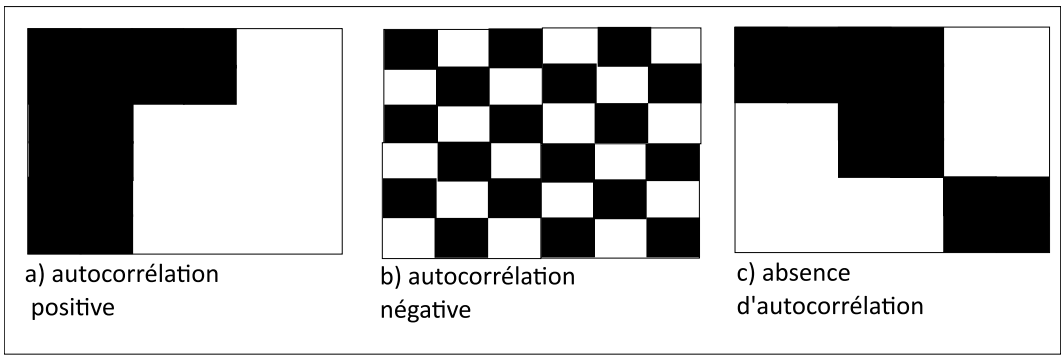


FIGURE 4.8 – Illustration de l'autocorrélation spatiale.

Avec l'analyse spatiale, c'est précisément des *associations spatiales* locales ou globales, des phénomènes de concentration spatiale (*clustering*) ou bien de dispersion qui sont à rechercher. Les statistiques doivent donc être adaptées pour prendre en compte l'espace et les *relations spatiales* comme des composantes à part entières de l'analyse. Depuis plus de cinquante ans, la recherche dans le domaine de l'analyse statistique spatiale a construit des méthodes adaptées, [Cressie 91], dont nous proposons un aperçu. Il s'agit pour nous d'introduire progressivement pour le lecteur la notion d'autocorrélation, ses méthodes de mesure, et de montrer comment elles peuvent s'utiliser enfin pour la recherche de valeurs exceptionnelles.

4.2.2.1 Les indices globaux d'autocorrélation spatiale

Les coefficients de corrélation spatiale sont construits de telle manière qu'il soit possible de répondre à la question : la variation de X entre unités géographiques proches est-elle plus ou moins grande que la moyenne des variations observée entre l'ensemble des unités de la zone étudiée prises deux à deux ? Ces coefficients sont exprimés sous la forme de rapports. Le dénominateur est, à une constante près, une mesure générale de la dispersion statistique de la distribution de X, le plus souvent sa variance σ_X^2. Le numérateur est, en général, soit une mesure de la dispersion statistique des valeurs prises par le caractère dans les unités voisines i et j, soit une mesure de la covariation des valeurs prises par le caractère dans les unités contiguës. Dans un contexte multivarié, ces écarts sont des mesures de distance entre profils, mesurant la distance entre les profils de deux unités voisines dans le premier cas, ou la distance de chaque profil au profil moyen dans le second.

Parmi toutes ces statistiques, l'*indice de Moran* (connu sous le raccourci de I de Moran) ne constitue donc qu'une possibilité, mais jusqu'à présent la plus robuste [Cliff 81]. Il s'agit du rapport de la covariation des valeurs de X dans les unités voisines et de la variance de X, et le voisinage choisi correspond à la contiguïté d'ordre 1. Par conséquent, sa forme est proche de celle du coefficient de corrélation, et en notant m le nombre de liaisons entre les unités (ou nombre total de paires de voisins), on a :

$$I = \frac{\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.6}$$

Le I de Moran mesure donc la covariation d'un point et de ses voisins, en ramenant le résultat à la variance de l'ensemble des points. Le résultat du calcul du I de Moran est d'interprétation facile puisqu'il s'interprète approximativement comme un coefficient de corrélation classique. Il varie entre -1 (autocorrélation spatiale négative) et +1 (autocorrélation spatiale positive). On notera cependant que la valeur du I de Moran peut parfois être supérieure à 1 ou inférieure à -1. La valeur zéro marque l'absence d'autocorrélation spatiale négative ou positive, à une échelle globale.

Toutefois, le calcul de Moran est déséquilibré, car certaines observations sont plus représentées que d'autres : certaines localités centrales ont en effet plus de voisins que d'autres, situées par exemple sur les limites du territoire ou dans des zones éparses. On choisit alors de corriger ce biais en calculant un indice Moran corrigé du nombre variable de localités prises en compte, imposant donc un poids identique à chaque observation. La définition classique de l'indice de Moran considère en effet que plus le nombre de voisins est important, plus l'individu aura de poids dans la matrice de pondération. Le nombre de paires de voisins (m) est alors égal à $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. Cela ne se justifie que rarement. Au contraire, on préfère que chaque individu ait le même poids, c'est-à-dire que sa contribution à l'indice d'autocorrélation spatiale soit la même, qu'il ait un ou plusieurs voisins. Pour cela, il faut standardiser la matrice W en ligne. Cette opération consiste à pondérer le nombre de voisins j de chaque localité i pour que chaque ligne de la matrice (qui décrit les voisins de chaque individu i) soit égale à 1. En d'autres termes, si un point a 5 voisins, chaque voisin comptera pour un cinquième du total. Dans ce cas, le nombre de paires de voisins (m) est égal au nombre des localités i (n), comme si chaque individu n'avait qu'un voisin. Par rapport à cette remarque, D'Aubigny a proposé un indice de Moran corrigé (ainsi qu'un indice de Geary corrigé), [D'Aubigny 06], et a démontré qu'alors ces deux indices corrigés sont équivalents.

Le *coefficient de Geary* se présente comme un rapport entre la variance des écarts des valeurs de X entre unités spatiales contiguës et la variance de X :

$$G = \frac{n-1}{2m} * \frac{\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.7}$$

Si $G = 1$, il n'y a pas d'autocorrélation spatiale. Si $G > 1$, l'autocorrélation est négative. Enfin, si $G < 1$, l'autocorrélation est positive. Des tests sont associés à chacun des coefficients, et permettent de définir sous certaines hypothèses les valeurs du coefficient significatives d'une absence d'autocorrélation.

Ces deux indices reposent sur l'usage d'une contiguïté d'ordre 1 (les couples d'unités voisines sont séparées par une frontière). Cependant, on peut envisager de faire varier l'ordre de la contiguïté, et de représenter alors une courbe d'autocorrélation spatiale dans un *corrélogramme* qui fournit une représentation des « ondes » d'interdépendance entre unités spatiales différentes. Ainsi, Lebart, [Lebart 69], utilise le coefficient de Geary pour mesurer l'autocorrélation du niveau d'urbanisation des départements français, avec trois indicateurs : les dépenses par habitant, les revenus par habitant, et le taux de population urbaine. Il fait varier les mesures pour des ordres de contiguïté allant de 1 à 9 (chaque couple d'unités voisines est alors séparée par au plus 9 frontières) et ces suites de mesure montrent l'existence d'une autocorrélation positive aux ordres de contiguïté faible qui diminue jusqu'à l'ordre 4 (donc les unités proches se ressemblent, mais de moins en moins), puis l'autocorrélation s'inverse et devient négative pour des ordres supérieurs à 7 (les départements éloignés ont tendance à avoir des valeurs très dissemblables). Cette variation de l'autocorrélation en fonction des différents ordres de contiguïté traduit l'existence d'une forte composante régionale de l'intensité de l'urbanisation en France au début des années soixante.

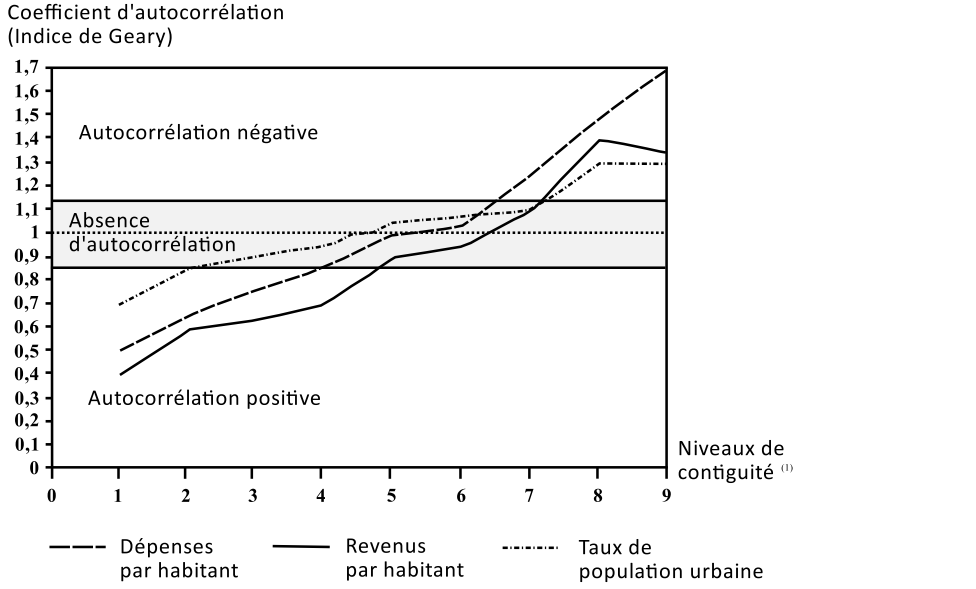


FIGURE 4.9 – Variation de l'autocorrélation spatiale mesurée par l'indice G sur des ordres de contiguïtés allant de 1 à 9, d'après [Lebart 69].

Ces indices peuvent être généralisés en remplaçant la matrice de contiguïté par une matrice de voisinage établie à partir de distances entre les unités [Decroly 96] : $w_{ij}(d_k) = 1$ si i et j sont distants de d_k. On calcule alors le coefficient de Bravais-Pearson, qui a la même interprétation que celle de Moran : proche de 0, il n'y a pas d'autocorrélation, sinon elle est positive ou négative suivant le signe du coefficient.

La courbe que propose Lebart est en fait un cas particulier de *variogramme*. Outre le terme générique de variogramme, qui renvoie à la représentation en ordonnée de la semivariance, on trouve aussi les termes de corrélogramme, qui renvoie à la covariance, ou d'autocorrélogramme pour des indices comme le I de Moran, et, de façon générale, en référence à l'autocorrélation spatiale représentée. On peut aussi rencontrer les termes de covariogramme et de semivariogramme lorsque la mesure de l'autocorrélation spatiale utilisée est la covariance ou la semivariance. Ces diagrammes donnent une mesure de l'échelle de la structure spatiale. Ainsi, on peut dégager des distances à l'intérieur desquelles la variable étudiée sera autocorrélée et mieux comprendre la dimension spatiale des structures ainsi découvertes. De même, le variogramme permet la comparaison pour un même ensemble d'objets géographiques de la structuration spatiale de différentes variables, ainsi que la comparaison des structures spatiales de différents ensembles d'objets géographiques [Oliveau 04]. Ainsi, au lieu de considérer la variation isotropique de la corrélation, on peut l'observer suivant des directions différentes. Par exemple, dans le cas de données portant sur la modernisation de l'Inde connues au niveau des villages, Oliveau décompose le calcul de l'indice I de Moran suivant les quatre directions cardinales, et obtient alors quatre mesures directement dépendantes de l'orientation des données, voir figure 4.10. L'interprétation qu'il en propose montre la richesse de cet outil pour la découverte de schémas d'organisation spatiale.

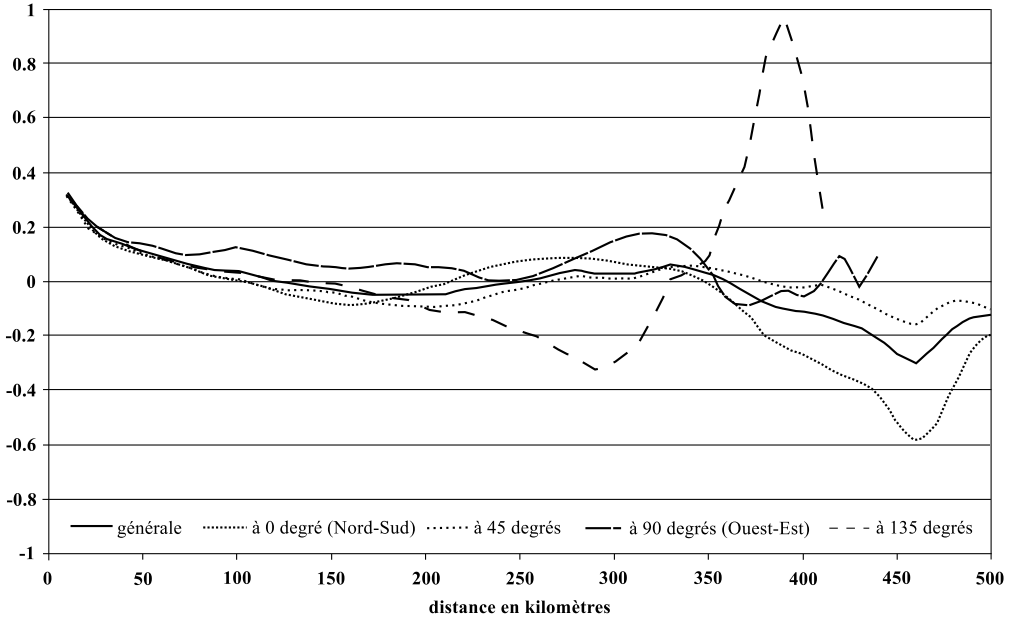


FIGURE 4.10 – Variogramme anisotrope de l'indice de modernisation en Inde, d'après [Oliveau 04].

La limite de l'exploration des données spatiales par des indices globaux d'autocorrélation spatiale est ici atteinte. En effet, si le corrélogramme permet d'aller plus loin, il ne permet pas d'être en mesure de connaître le détail de la structure spatiale de la variable étudiée [Charre 95]. Par ailleurs, les indices globaux peuvent être « aveugles » face à des phénomènes très fortement structurés, mais situés dans de petites zones au sein de grands espaces. De plus, comme le coefficient de corrélation linéaire ρ, ils ne sont pas robustes : ils dépendent fortement des valeurs extrêmes qui augmentent la variance des données.

Ainsi, la valeur de l'indice global d'autocorrélation spatiale est plus influencée par la présence de valeurs d'extrêmes que par les valeurs voisines proches de la moyenne. Par ailleurs, l'autocorrélation spatiale augmente mécaniquement avec l'agrégation des données (c'est l'effet d'échelle du MAUP).

4.2.2.2 Les indices locaux d'autocorrélation spatiale

Les indices G_i et G_i^* proposés par Getis et Ord [Getis 92], [Getis 95], comme les indices de Moran individuels, appelés indices locaux d'association spatiale ou LISA par leur inventeur Anselin, [Anselin 95], ont pour objectif d'affiner les méthodes de repérage et de description des mesures d'autocorrélation spatiale. Les indices G_i et G_i^* sont des indices locaux que l'on peut éventuellement transformer à nouveau en indices globaux. Le principe des LISA est de désagréger des indices globaux existants (I de Moran ou c de Geary par exemple). Les indices locaux mettent en évidence des associations locales de valeur. Le terme d'association locale apporte une légère nuance à celui d'autocorrélation spatiale locale. Quand l'autocorrélation spatiale mesure la plus grande similarité statistique entre les valeurs d'une variable associée à deux individus par rapport à la moyenne de l'échantillon, le terme d'association soulignerait plutôt le regroupement spatial des individus dont les valeurs de la variable étudiée sont extrêmes (*hot spots* ou *cold spots*). La nuance est faible, et les deux termes sont généralement employés comme synonymes [Oliveau 04].

L'idée du LISA repose sur un constat simple : puisque les indices globaux sont une somme pour tous les i, on peut définir pour chaque individu i une statistique locale de la forme : $\Gamma_i = \sum_j W_{ij}V_{ij}$ avec $V_{ij} = (x_i - \bar{x})(x_j - \bar{x})$ dans le cas de l'indice de Moran. L'indice gamma global Γ est la somme de tous les indices locaux Γ_i moyennant un facteur γ, suivant l'équation 4.8

$$\sum \Gamma_i = \gamma\Gamma \text{ avec } \gamma = \sum_i \sum_j W_{ij} \frac{\sum_i (x_i - \bar{x})^2}{n} \tag{4.8}$$

Ce facteur de proportionnalité γ permet d'apprécier la valeur d'un gamma local Γ_i en fonction du gamma global Γ. Ainsi, il est aisé de mettre en évidence des valeurs atypiques, en comparant la valeur du gamma local à celle de la moyenne des gamma locaux (c'est-à-dire le gamma global divisé par le nombre d'individus i).

Anselin propose de plus une méthode de visualisation des indices LISA dans [Anselin 96], qui permet de détecter rapidement la présence de *hot-spots*, comme de valeurs exceptionnelles. Il utilise pour cela un nuage de points de Moran (*Moran scatterplot*) représentant les valeurs croisées des unités spatiales pour leur valeur et la valeur moyenne de leur voisinage. Ainsi, comme sur la figure 4.15, l'axe des abscisses porte les valeurs de la variable X, et l'axe des ordonnées les valeurs moyennes de la variable pour les individus voisins, noté Wx [4]. Quatre quadrants sont ainsi constitués : intuitivement, les points dans les quadrants 2 et 4 sont des valeurs exceptionnelles car elles sont soit faibles au milieu de valeurs fortes, soit fortes au milieu de valeurs faibles. L'hétérogénéité locale se caractérise par le regroupement de points en dehors du nuage principal : il s'agit souvent d'une poche locale d'autocorrélation spatiale. À l'intersection des quadrants, au centre du nuage de points, les valeurs sont proches de la moyenne de l'échantillon et les valeurs de leurs voisins aussi. Elles sont donc peu significatives. De plus, l'indice global de Moran peut s'interpréter comme le coefficient directeur de la droite de régression liant les

4. Le « *spatial lag* » en anglais.

valeurs de X aux valeurs Wx : si cette droite est tracée, il est aisé de percevoir le degré de non-linéarité ou de linéarité de la dépendance entre Wx et X.

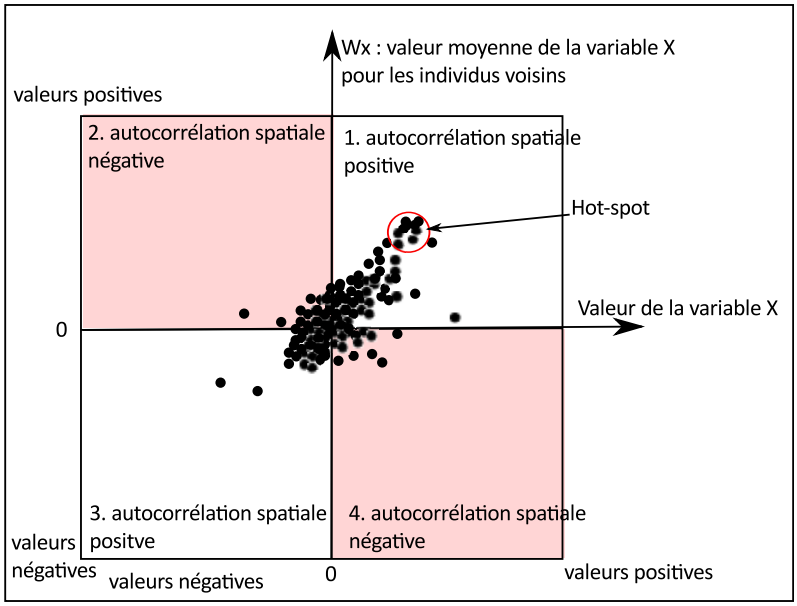


FIGURE 4.11 – Représentation schématique de nuage de points de Moran.

On peut toutefois s'interroger sur la pertinence de ce découpage en quadrants. Si l'on prend un découpage par bandes, en diagonales par exemple, du même échantillon, la diagonale représente les valeurs qui ressemblent à leur voisinage, tandis que la largeur de la bande représente l'écart toléré à cette moyenne locale. Dans ce cas, les valeurs exceptionnelles seront les valeurs en dehors de cette bande (voir figure 4.12)

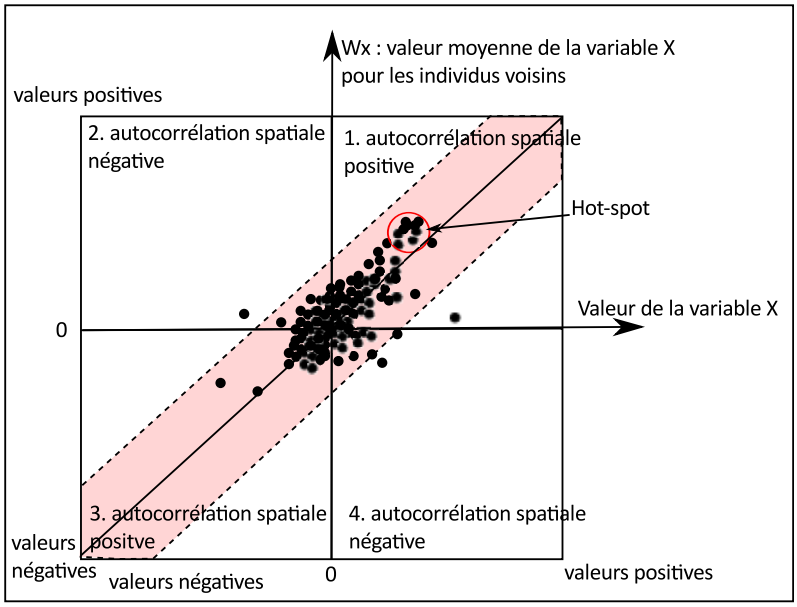


FIGURE 4.12 – Représentation schématique de nuage de points de Moran, modifiée par rapport à la détection de valeurs exceptionnelles.

D'autres représentations graphiques de l'autocorrélation spatiale, pour des indices locaux, ont été proposés dont, par exemple, des diagrammes en camembert figurant les décalages spatiaux (*spatial lag pies*) par Anselin [Anselin 93], ou le diagnostique d'autocorrélation locale [Nass 92] basé sur des tests de permutation dans la matrice de voisinage.

Nous allons maintenant montrer comment ces coefficients d'autocorrélation peuvent être utilisés pour construire une régression spatiale entre deux variables tenant compte des effets spatiaux, et comment la cartographie des résidus propose une visualisation des valeurs exceptionnelles.

4.2.2.3 L'analyse des résidus géographiquement pondérée (GWR)

La régression géographiquement pondérée, *Geographically Weighted Regression* - GWR- en anglais, est une adaptation de la régression linéaire multiple classique qui permet de tenir compte de la variation dans l'espace des corrélations [Fotheringham 02]. Un modèle de régression linéaire global cherche à décrire la relation entre p variables X_k et Y en ajustant une droite estimant les valeurs de Y en fonction des X_k selon l'équation $\hat{y} = b_0 + \sum_{k=1}^{p} b_k x_k$. On cherche à résoudre une contrainte aux moindres carrés sur le vecteur b, c'est-à-dire à minimiser la distance entre les $\hat{y_i}$ estimés et les y_i mesurés, formulée par l'équation 4.9.

$$Min\{\sum_{i=1}^{n}(y_i - \hat{y_i})^2\} \tag{4.9}$$

et dont la solution usuelle est :

$$b = (X^T X)^{-1} X^T Y \tag{4.10}$$

La GWR modifie ce modèle en présumant que la relation entre les variables X_i et Y n'est pas constante dans l'espace, en conséquence de quoi les paramètres de la droite de régression peuvent varier pour chaque point i. Elle utilise donc une matrice de voisinage W qui permet de pondérer les valeurs de la régression avec une fonction décroissante de la distance. Elle propose donc comme solution :

$$b = (X^T W X)^{-1} X^T W Y. \tag{4.11}$$

Ainsi, en dimension 2, lorsque Y n'est ajusté qu'à une seule variable X, Pour chaque point i de calcul (variant de 1 à n), les paramètres locaux de la droite sont ajustés par la méthode des moindres carrés, à partir des valeurs des variables X et Y connues pour les points j de données (variant de 1 à m), chacune affectée d'un coefficient w_{ij} décroissant selon la distance d_{ij} entre les points i et j. Ainsi, les valeurs localisées à proximité du point de calcul pèsent plus que les valeurs éloignées dans la détermination des paramètres locaux de la droite de régression d'équation $\hat{y} = b_0 + b_1 x$. Ces derniers sont donc donnés par les équations :

$$b_1 = \frac{cov(x,y)}{\sigma_x^2} = \frac{\frac{1}{m}\sum_{j=1}^{m} w_{ij}(x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.12}$$

$$b_0 = \frac{\sum_{j=1}^{m} w_{ij}.y_j - b_1 \sum_{j=1}^{m} w_{ij}.x_j}{\sum_{j=1}^{m} w_{ij}} \tag{4.13}$$

Plusieurs modèles peuvent être employées pour déterminer les w_{ij}. Un modèle logistique conviendra si la variable Y ne prend que des valeurs binaires 0 ou 1. Si les données sont des comptes positifs (des stocks), les auteurs proposent d'utiliser un modèle de Poisson. Pour des variables à valeurs réelles, on choisit souvent un modèle gaussien, de portée (*bandwith* en anglais) b, et alors :

$$w_{ij} = \exp(-(\frac{d_{ij}}{b})^2) \tag{4.14}$$

L'étude des résidus ϵ_j qui sont les écarts entre les valeurs prévues par le modèle $\hat{y}_i$ et les valeurs y_j mesurées permet de se rendre compte de l'écart au modèle moyen prédit. Ces écarts rendent compte du caractère exceptionnel d'une valeur par rapport à son voisinage, et lorsqu'ils sont cartographiés, facilitent la lecture des valeurs exceptionnelles (voir figure 8.8). En effet, si R est le coefficient de corrélation entre la série y_i et $\hat{y}_i$, $i = 1..n$, alors R^2 représente la part de variance expliquée par la régression, R^2 est appelé le coefficient de détermination. Si $R^2 = 1$, l'ajustement est parfait. Comme on connait la loi de probabilité de R^2, on peut tester la significativité des résidus de la régression (sont-ils le fait du hasard, ou bien montrent-ils la présence de valeurs exceptionnelles pour Y ?).

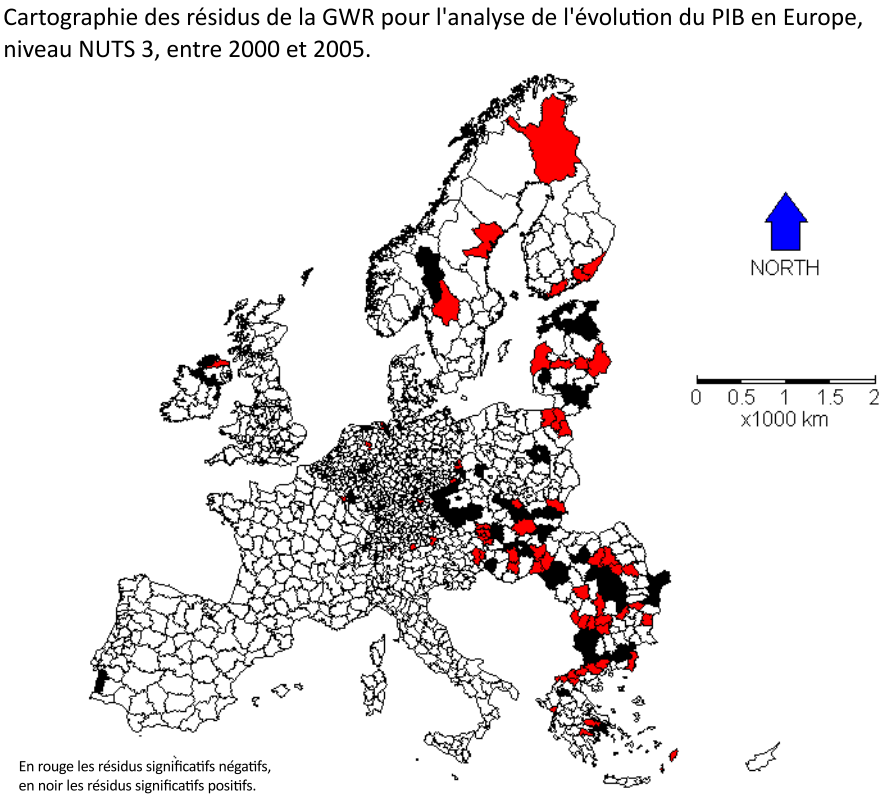


FIGURE 4.13 – Usage de la cartographie des résidus de la GWR pour détecter des valeurs exceptionnelles. Source : [Harris 10].

Cependant, cette méthode est très sensible au MAUP comme le démontre Charleux [Charleux 05], et en particulier aux effets d'échelles : plus les individus sont agrégés dans un faible nombre de classes spatiales (passage, par exemple, d'une échelle spatiale locale de type commune à une échelle spatiale moins fine de type département), plus la variance diminue. Si l'échelle du maillage pouvait être contrôlée (cas de carroyages par exemple), alors cette sensibilité aux effets d'agrégation statistique pourrait être favorable à l'étude de configurations spatiales dépendantes des échelles. Cependant, dans le cas réel des maillages territoriaux qu'intègre la NUTS, la taille et la forme des unités est très hétérogène à chaque niveau, et donc brouille la lecture des cartes de résidus.

4.2.2.4 Recherche de valeurs exceptionnelles par l'analyse multi-niveau

Le traitement de données zonales, qui est extrêmement sensible aux effets d'échelles [Openshaw 77, Openshaw 81, Openshaw 96], [Marceau 99] du fait de ces processus d'agrégation, peut énormément bénéficier d'analyses spatiales multi-niveaux, qui intègrent et mettent en évidence des effets d'échelles. Ce point de vue est défendu par Mathian et Piron qui proposent une méthodologie pour l'analyse multi-niveau [Piron 93], [Mathian 01] qui passe par l'analyse de la variance. Elles soulignent que la prise en compte des différents niveaux d'analyse naturellement créés par les jeux d'emboîtement existant entre les différents niveaux de zonages peut être particulièrement utile à l'analyse des différenciations. Ces différents niveaux agissent comme des filtres qui laissent à voir des structures et des formes spatiales différentes, co-existant simultanément.

Les travaux de Haggett sur *l'analyse de la variance* servent de base à cette approche [Haggett 73] qui consiste à considérer que les relations d'appartenance d'unités d'un niveau inférieur à d'autres unités de niveau supérieur permettent de fabriquer des classes X_i, c'est-à-dire des sous-ensembles dans un ensemble d'unités formant un zonage d'un certain niveau, comme par exemple les départements regroupés par régions. Les classes X_i sont formées d'unités appartenant à la même unité de niveau supérieur. Il s'agit alors d'étudier la variable Y, quantitative à caractère continu (quantitatif continu ou quasi continu, qualitatif ordinal), connue sur ces unités et de calculer trois types de variance :

- la variance globale de toutes les unités du zonage,
- la variance intra-classe qui est la variance interne à chaque classe du zonage,
- la variance inter-classe qui est la variance entre les classes du zonage.

Intuitivement, la relation entre une variable Y et la variable catégorielle X est forte si pour l'ensemble des classes de X, les moyennes de Y sont très différentes, et les variances autour de ces moyennes petites [Dumolard 03], donc si la variance intra-classe est faible, et la variance inter-classe forte. [Dumolard 03] définit la *somme des carrés des écarts* (SCE) aux moyennes de classe et à la moyenne générale comme la variance de l'échantillon multipliée par son effectif. Les équations 4.15, 4.16, et 4.17 résument les relations mathématiques qui existent entre ces différents écarts pour une variable Y associée à une variable de classe X, si n est le nombre d'unités du zonage, k le nombre de classes X_g (ou modalités de X) de ce zonage, n_g le nombre d'éléments chaque classe X_g, $\bar{y_g}$ la moyenne de Y dans chaque classe X_g, $\bar{y}$ la moyenne globale de Y toutes classes confondues, et σ_y la variance globale de Y sur le zonage.

$$SCE_{global} = SCE_{intra} + SCE_{inter} \tag{4.15}$$

$$SCE_{global} = \sum_{i=1}^{n} (Y_i - \bar{y})^2 = n * \sigma_y^2 \tag{4.16}$$

$$SCE_{inter} = \sum_{g=1}^{k} n_g (\bar{y_g} - \bar{y})^2 \tag{4.17}$$

Ce modèle s'utilise si la distribution des valeurs de Y dans chaque classe est à peu près gaussienne, si les variances intra-classes sont sensiblement égales et si les effectifs de chaque classe sont à peu près égaux. Ce qu'on nomme « analyse de la variance » est en réalité une analyse sur l'inégalité des moyennes. Elle peut servir à repérer des valeurs exceptionnelles si on est en présence d'une liaison forte entre Y et X, et que de plus l'écart au profil moyen de sa classe pour une unité est très fort (par exemple, supérieur à 1,5 fois l'écart-type de sa classe). Un autre moyen de mesurer le caractère exceptionnel d'une unité consiste à calculer le rapport à la moyenne de sa classe, multiplié par 100 par exemple (c'est une mesure d'écart). Si ce rapport est loin de l'indice 100, alors l'unité est exceptionnelle dans sa classe.

Ainsi, l'analyse de la variance permet de repérer des valeurs qui sont exceptionnelles, non pas seulement par rapport à un voisinage déterminé par une distance dans le plan que constituent les unités du même niveau, mais également par rapport à la relation d'appartenance à des unités de niveau supérieur.

4.2.3 L'étude temporelle

Pour traiter les séries de données, un nombre important d'outils statistiques ont été développés. Ici, nous ne présentons qu'une courte introduction aux principes qui régissent ces outils. Cependant, dans le domaine de l'information territoriale, les séries suffisamment fournies sont rares du fait des nombreuses recompositions territoriales. Il devient dès lors nécessaire de construire ces séries en transvasant dans un zonage de référence, celui de l'étude, l'ensemble des indicateurs qui ont été mesurés sur les différentes versions de ce zonage. Nous proposons donc une revue pour les méthodes de transfert d'indicateurs entre supports non alignés.

4.2.3.1 L'analyse des séries temporelles

L'analyse des séries temporelles part du principe que les phénomènes observés suivent une régularité dans le temps, qui comprend à la fois une tendance et des effets saisonniers. Par exemple, la température diminue en hiver et augmente en été, mais la tendance globale peut être au réchauffement climatique. Cette tendance peut être la stabilité, et la série est alors dite stationnaire. Il existe aussi des séries dont la tendance est non-linéaire, comme les cours boursiers, mais le phénomène est imaginé continu et la valeur d'une mesure semble plutôt dépendre de quelques valeurs précédentes. Dans la figure 4.14, l'exemple 1 illustre une chronique de tendance linéaire, ayant une variabilité saisonnière de période 4 (unités de temps), tandis que l'exemple 2 montre une chronique qui présente une tendance non linéaire, sans variabilité saisonnière.

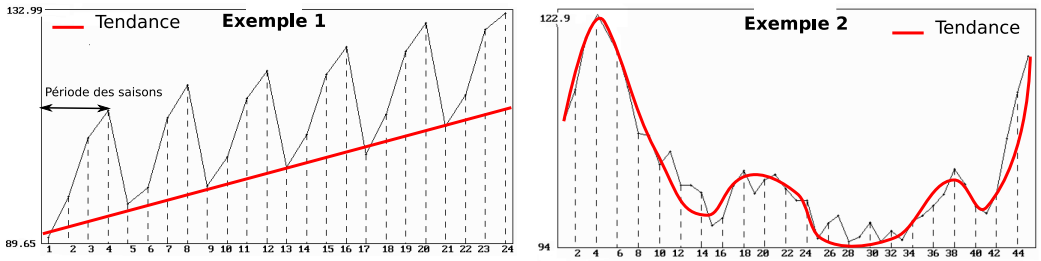


FIGURE 4.14 – Deux exemples de chroniques.

Il existe deux manières de procéder pour l'estimation des séries temporelles et la détection de valeurs

exceptionnelles. Soit, comme dans l'exemple 1, la série présente une tendance linéaire, et, peut-être, des variations saisonnières. Il s'agit alors de retrouver par calcul la tendance globale, ainsi que les éventuelles variations saisonnières, afin de calculer la moyenne temporelle à laquelle chaque mesure doit être comparée, en tenant compte de la saisonnalité du phénomène observé. Dans ce cas, une valeur v' mesurée au temps $t + P$ diffère normalement peu d'une valeur v mesurée au temps t à laquelle les effets de la tendance globale sont appliqués, avec une variation saisonnière de périodicité P. Soit, comme dans l'exemple 2, la série ne semble pas présenter de tendance linéaire ni de saisons : alors les modèles exponentiels sont utilisés, et ils prédisent chaque valeur de la série en fonction de quelques valeurs précédentes. Si la valeur mesurée s'écarte de beaucoup de la valeur mesurée, la valeur est dite exceptionnelle. La littérature dans le domaine est abondante et pour de plus amples explications sur la prédiction temporelle, on peut se référer aux ouvrages suivants [Anderson 71], [Droesbecke 89] et [Box 94] (pour les modèles ARIMA) et [Droesbecke 94] (pour les modèles ARCH).

Dans ce paragraphe, nous nous limitons à une idée très simple concernant l'étude spatiale de la variation des valeurs, en considérant leur dérivée première, et ce que cette étude peut apporter à la recherche de valeurs exceptionnelles. En effet, si on s'intéresse au rapport existant entre une variable X mesurée au temps t_1, et sa mesure au temps t_2, sur l'ensemble des unités formant un zonage, il est possible de calculer la variation moyenne de ces unités, sur l'ensemble des unités, mais également de considérer cette variation comme une nouvelle variable (de taux) spatialisée, sur laquelle les méthodes d'analyse de la variance, comme les méthodes de détection de auto-corrélation locale et globale s'appliquent. Il est ainsi possible de repérer les zones qui évoluent très fortement, dans un sens négatif ou positif, ainsi que les unités qui évoluent de façon exceptionnelle par rapport à leurs voisines, ou à leur classe d'appartenance définie par la relation hiérarchique avec un niveau supérieur.

Cette idée semble facile à mettre en œuvre, mais le premier problème à surmonter concerne la constitution de séries temporelles longues. Du fait des changements et des transformations perpétuelles des zonages, qui servent de support à la collecte ou l'analyse des données, les données entre deux périodes différentes se sont pas toujours comparables puisqu'elles sont relevées sur des zones ayant des emprises spatiales différentes. Par exemple, au niveau NUTS2 au Portugal, entre 1999 et 2003, la redistribution qui intervient entre les unités PT12, PT13, PT14 de la version 1999 et les unités PT16, PT17 et PT18 de la version 2003 (voir la figure 4.15) rend non comparable la population qui peut avoir été recensée sur l'unité PT12 puis l'unité PT16 (car l'unité PT16 inclut PT12, mais également un morceau de PT13). Pour être en mesure de reconstruire des séries temporelles sur un zonage choisi pour l'étude, il faut mettre en oeuvre des méthodes de transfert entre maillages non alignés, qui s'inscrivent dans le cadre de réponses apportées au « *problème du changement de support* ».

4.2.3.2 Le problème du changement de support

La représentation de l'information géographique est nécessairement portée par un objet géométrique qui en constitue le *support*. Le support peut être de formes différentes : ponctuel, linéaire, polygonal. Le support polygonal a deux aspects : *régulier* ou *irrégulier*. Ainsi, un support matriciel (ou grille) est un support polygonal régulier dans lequel on observe la répétition d'un motif géométrique. *A contrario*, un support polygonal irrégulier est composé de cellules de tailles et de formes variées. Ce type de support, polygonal et irrégulier, est aussi généralement nommé maillage (s'il forme une partition complète de l'espace d'étude) ou zonage. Enfin, deux maillages sont *non-alignés* s'ils ne présentent aucune forme d'emboîtement, c'est-à-dire que leurs frontières se chevauchent, comme ce peut être le cas des départements avec les bassins versants, ou bien des frontières de deux départements à des époques différentes

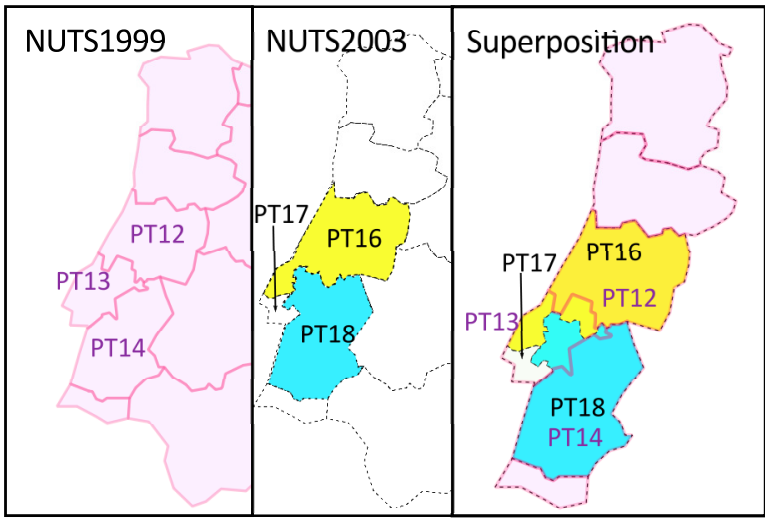


FIGURE 4.15 – Un évènement de redistribution au Portugal entre les versions de NUTS 1999 et 2003 au niveau régional. Les régions codées PT12, PT13 PT14 deviennent les régions PT16, PT17 et PT18.

par exemple. Dans ces cas, il est indispensable d'harmoniser les supports à des fins d'analyse et/ou d'interprétation de la donnée, et donc de transférer les données (ou indicateurs ou variables) dans un support commun.

Cette harmonisation est identifiée dans la littérature anglo-saxonne sous l'expression *Change Of Support Problem* (COSP) [Arbia 89], qui signifie problème de changement de support en français. En nous basant sur la terminologie adoptée par [Markoff 73], nous parlons de changement de support lorsqu'une variable est connue sur un support *source* et qu'elle est transférée, à l'aide de méthodes plus ou moins complexes, sur un support *cible*. Le problème de transfert de données est connu dans la littérature anglaise sous différents noms : « *spatial rescaling* » [Nordhaus 02], ou « *downscaling* », parce que la principale difficulté vient de la désaggrégation d'une variable sur un support d'échelle plus fine, mais non-aligné avec le support source. Les techniques de transfert portent les noms de « *cross-area aggregation* » [Fotheringham 00], ou « *areal interpolation* » [Flowerdew 89], [Fisher 95], ou encore « *polygon overlay problem* ».

Les méthodes de transfert de données d'un support à un autre sont donc basées sur la création d'un maillage intermédiaire, inclus dans le maillage cible. Les opérations possibles sont les suivantes :

- la *désagrégation* est une transformation qui permet de ventiler la donnée sur des objets (ponctuels ou polygonaux) de taille plus réduite, complètement inclus dans les mailles du support d'origine, grâce à la fabrication d'un support constituant le plus petit commun dénominateur spatial entre le support source et le support cible, par intersection des deux supports.
- l'*agrégation* de données est une transformation obtenue par le cumul des données selon des mailles incluses spatialement dans des mailles de taille supérieure. Elle peut survenir après une désagrégation pour obtenir une estimation de la variable sur un autre maillage ou bien lors du passage d'une échelle locale à une échelle globale, dans le cas de hiérarchie.
- l'*interpolation* permet, à partir d'un support ponctuel (ou parfois polygonal), de reconstruire une surface de densité ayant pour support un maillage régulier, qui peut être aussi fin que voulu.

L'interpolation s'inscrit généralement dans une séquence d'opérations où la donnée est réduite à un semis de points, puis interpolée, puis réagrégée sur le support final. En effet, il est possible et aisé de résumer un polygone en son centre. La donnée portée par le polygone est alors entièrement transférée au centre, qui peut être le centre de l'enveloppe géométrique (centroïde), ou bien un centre géographique localisé dans la maille, choisi par l'utilisateur pour sa pertinence dans l'étude : la plus grande ville, le point d'altitude le plus bas, etc. La surface de tendance calculée a souvent pour but de donner une représentation continue de mesures discrètes ou ponctuelles. Cette surface est nécessairement portée par une discrétisation de l'espace, fine, sur laquelle les éléments peuvent être sommés en fonction de leur inclusion dans le maillage cible, (seulement si les variables sont des stocks). Le graphe de la figure 4.16 présente schématiquement ces trois opérations. Les transformations dépendent pour beaucoup de la nature des supports source et cible. Mais, ces transformations ont aussi un effet sur l'analyse qui peut être menée sur les données, en raison du changement de taille et de forme du support. En effet, plus les unités spatiales d'analyse sont grandes, moins la variation est importante entre les valeurs prises par les variables d'une unité à l'autre (effet d'échelle - *scale effect*), et la délimitation des entités spatiales influe sur la lecture du phénomène (effet de découpage - *zoning effect*) [Openshaw 79].

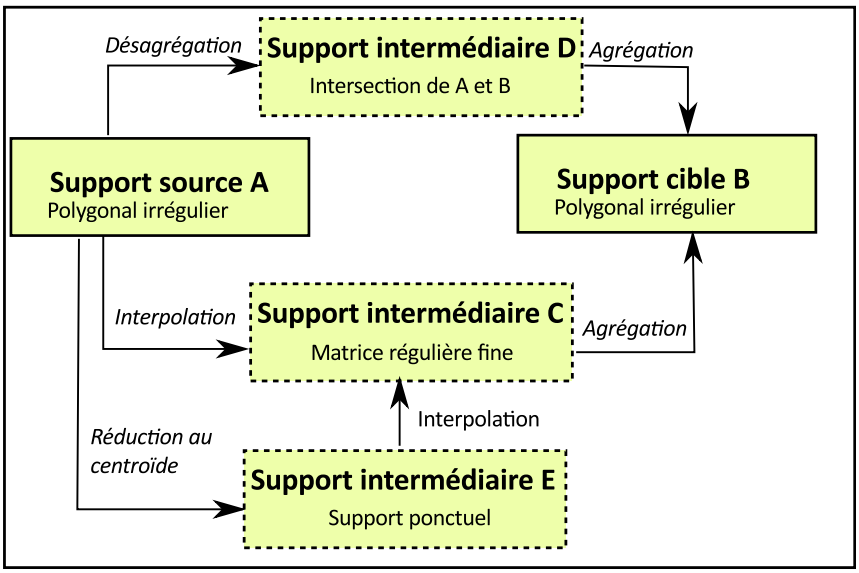


FIGURE 4.16 – Graphe des transformations de support pour le transfert des données.

4.2.3.3 Aperçu des principales méthodes de « transfert »

Nous présentons ici un aperçu synthétique des différentes familles de techniques de transfert, classées suivant le type d'opération de transformation que le type des supports source et cible impose. Rappelons toutefois que le type de la variable à transférer joue un rôle : une variable quantitative absolue (un compte d'effectifs sur des unités) ne peut être manipulée de la même manière qu'une variable relative (un ratio), en particulier dans les opérations d'agrégation. Les taux ne s'additionnent pas pour constituer le taux de la maille supérieure. Enfin, le transfert s'effectue dans un contexte dans lequel on peut choisir de faire intervenir d'autres variables qui ont une influence sur la spatialisation du phénomène : on les appelle *variables auxiliaires*.

4.2.3.3.1 Opérations d'agrégation - possibles uniquement pour des stocks Pour l'agrégation, nous aurions pu citer les techniques de régionalisation [Pumain 97], d'optimisation [Openshaw 88] par recuit-simulé, ou de *clustering*. Toutefois ces techniques visent essentiellement à reconstruire un nouveau maillage suivant certaines contraintes posées par l'utilisateur (similarité, homogénéité, continuité, etc.). Or, notre objectif est de transférer une variable vers un maillage cible dont les contours sont définis, et dans ce cadre, la technique d'agrégation de données matricielles vers un maillage irrégulier est très opportune. Elle correspond sur la figure 4.16 à la transformation de C vers B. L'idée est d'agréger des variables connues sur un support matriciel vers un maillage irrégulier. En superposant le maillage cible à la matrice, il est possible de cumuler les effectifs des cellules selon le maillage cible, et de connaître ainsi la valeur de la variable sur le nouveau support, moyennant une marge d'erreur. Cette erreur dépend à la fois de la taille des cellules du maillage intermédiaire, et de la procédure adoptée pour réallouer une cellule à une unité du maillage cible. Par exemple, soit la cellule appartient majoritairement au polygone, dans ce cas on lui affecte l'effectif total porté par la cellule ; soit on réalloue proportionnellement à la surface d'intersection avec le polygone (voir figure 4.17). Pour réduire autant que possible l'erreur, le maillage intermédiaire doit être le plus fin possible. Par ailleurs, une grande part de variabilité dans l'information est perdue puisqu'elle est ré-agrégée dans des unités de plus grande taille et de formes différentes (scale and zoning effect). Ce type de méthode est très largement répandu dans le domaine environnemental où les données sur des supports raster (matriciels) abondent, lorsque les données doivent être transférées dans un maillage socio-économique comme la Nomenclature des Unités Territoriales Statistiques (NUTS) [Paramo 05], [Gómez 09].

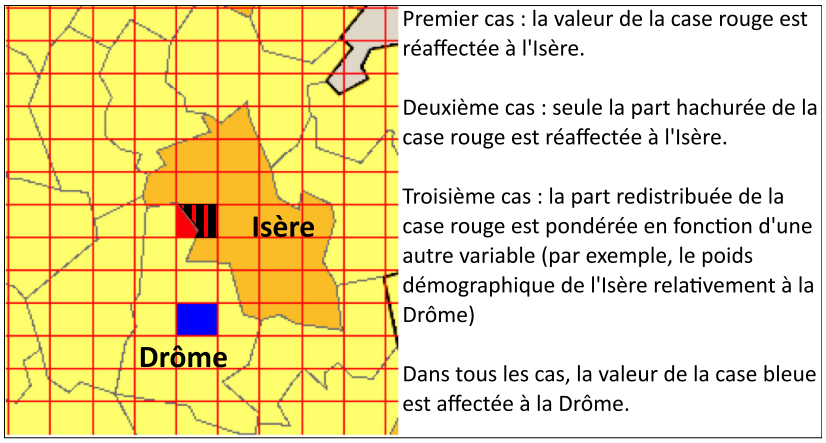


FIGURE 4.17 – Agrégation d'une grille dans un support maillé irrégulier.

4.2.3.3.2 Opérations de désagrégation

4.2.3.3.2.1 Pondération surfacique simple (*Simple Areal Weighted*) Cette méthode permet de transférer une variable d'un maillage à un autre sans qu'ils soient emboîtés, par la constrution d'un support intermédiaire D, résultant de l'insersection des supports source et cible. Elle suppose l'hypothèse de l'uniformité de densité de la variable dans chaque maille. La ventilation de la variable se fait proportionnellement à la surface d'intersection d'une maille source avec le maillage cible [Goodchild 80]. Elle est facile à mettre en œuvre et ne nécessite pas un grand échantillon de données. L'algorithme est adopté par la plupart des systèmes d'information géographique (SIG). Par contre, l'hypothèse d'uniformité de densité de la variable est simpliste et ne permet pas de rendre compte des variations locales.

4.2.3.3.2.2 Pondération surfacique contrôlée par régression (*Modified Areal Weighted - Regression*) Cette méthode admet que des relations de corrélation spatiale peuvent exister entre les variables (confère [Droesbecke 06]). On identifie donc une variable auxiliaire corrélée spatialement avec la variable à transférer, et qui soit connue sur les maillages source et cible. Cette méthode introduit dans la ventilation de la variable sur le support d'intersection D une contrainte correspondant à la similarité de répartition entre la variable à transférer et la variable auxiliaire, dite « *prédicteur* ». Dans une première étape, une étude statistique formalise la ressemblance de dispersion spatiale entre les deux variables sur le support source. La seconde étape résout une équation de régression (qui peut être linéaire ou non) entre le prédicteur, dont la dispersion est connue sur le support cible, et la variable à estimer [Goodchild 93]. La figure 4.18 illustre sur un exemple fictif cette méthode avec un transfert de la variable population entre le maillage des NUTS, niveau 3, et celui des bassins versants sur l'espace européen. Ici, la variable auxiliaire envisagée est l'altitude moyenne, et on imagine assez facilement que la densité de population décroit avec l'altitude (les gens habitent majoritairement en plaine, près du niveau de la mer). Comme on connaît l'altitude moyenne des départements (unités de la NUTS3 en France) et des bassins versants, cette méthode permet de déduire aisément la population sur les bassins versants. L'utilisation d'une variable auxiliaire peut améliorer de façon conséquente les résultats, mais il n'est pas aisé de trouver des données complémentaires qui soient fortement corrélées à la variable étudiée. Ainsi, dans notre exemple, l'altitude des bassins versants n'est elle-même pas assez homogène pour être considérée comme une bonne variable auxiliaire.

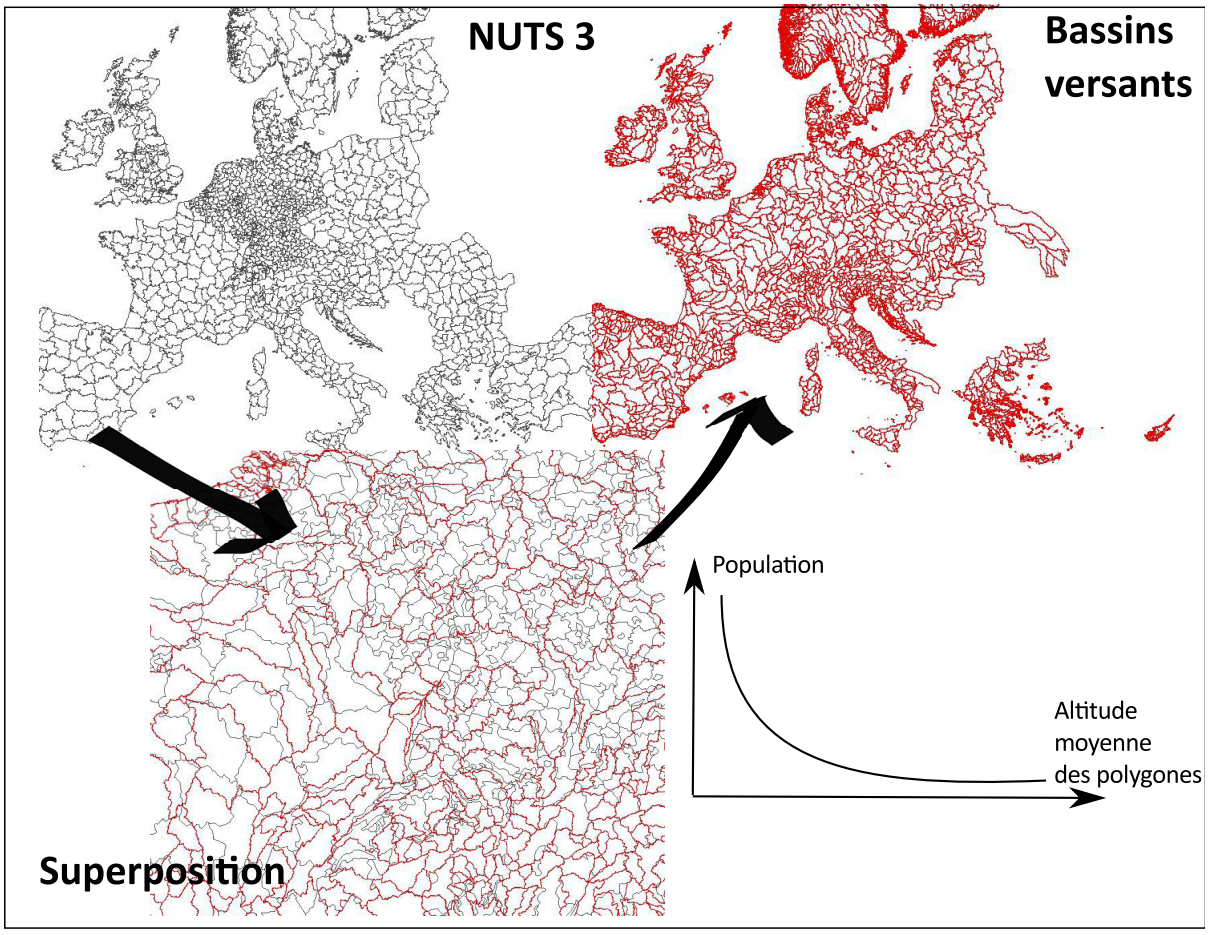


FIGURE 4.18 – Utilisation d'une régression sur variable auxiliaire pour le transfert de variable.

4.2.3.3.2.3 Pondération surfacique contrôlée par zones de contrôle (*Modified Areal Weighted - control zones*) Cette alternative de la pondération surfacique simple se base sur une donnée connue sur un « *maillage de contrôle* », pour établir des estimations sur le maillage cible. En effet, il est possible de calculer la répartition spatiale de la donnée à transférer à l'aide de la donnée auxilliaire connue sur le maillage de contrôle ou sur des objets géographiques de granularité plus fine que celle du maillage cible. Ce troisième maillage est appelé « *zones de contrôle* » (*control zones*) par Goodchild93. On considère que ces zones de contrôle ont une densité constante, [Langford 92]. Dans le domaine de l'environnement, la tendance actuelle [Reibel 07] est d'utiliser les données d'image satellitaires (matrices d'occupation du sol) comme zones de contrôle pour le transfert des variables. En effet, ces matrices sont traitées pour reconnaître les types majoritaires d'occupation du sol sur des pixels de 1 ha, ce qui facilite le transfert d'un nombre important de variables : la population est liée aux zones urbaines, la quantité de pesticides aux zones agricoles, etc.

4.2.3.3.3 Opérations d'interpolation Les méthodes d'interpolation sont très nombreuses [Arnaud 00]. Cependant, sur le plan thématique, le schéma de transfert par interpolation ne s'applique pas sans précaution au cas de données zonales. En effet, avec les variables socio-économiques comme la population, le PIB, on est en situation d'évaluer des caractères qui ne sont pas continus *stricto-sensus* : on assigne aux individus des classes spatiales (les unités d'un zonage), car le niveau élémentaire de l'information ne peut être observé (voir la figure 3 page 10), mais le caractère observé n'est pas un phénomène continu sur l'espace géographique. Ceci n'est pas anodin car, si dans une unité est observée la quantité d'habitants ou d'usines, alors cette quantité augmente avec la taille de l'unité. Ainsi, en construisant une surface continue par l'interpolation avec des méthodes classiques d'interpolation telles que triangulation, moyenne locale, interpolation polynomiale (splines, Bezier) ou la méthode de Shepard, on aura tendance à faire croire que les disparités observées sont le fait du phénomène étudié, alors qu'en réalité, elles seront la résultante de l'hétérogénéité du maillage, [Grasland 06]. Dans [Grasland 06], les auteurs s'appuient sur la comparaison de cartes obtenues à partir d'un maillage de niveau NUTS 2 ou NUTS 3 en employant la méthode de Shepard pour démontrer leur propos. Toujours sur le plan thématique, les auteurs indiquent aussi qu'en revanche, le passage par fréquence, c'est-à-dire en rapportant la quantité observée à la surface des unités, ou bien au nombre d'habitants par unité par exemple, on gomme l'effet de taille des unités, et qu'alors les valeurs de fréquence deviennent comparables et même interpolables. Enfin, les auteurs de [Grasland 06] précisent qu'en revanche la méthode pycnophylactique et la méthode du potentiel s'appliquent parfaitement à des stocks car elles tiennent compte dans leur principe de cette discontinuité de l'information de base.

Toutefois, sur le plan méthodologique, les méthodes d'interpolations classiques (telles que triangulation, moyenne locale, interpolation polynomiale (splines, Bezier) ou la méthode de Shepard) sont très utilisées [Zaninetti 05], même si elles sont plus adaptées à l'estimation de variables continues dans l'espace pour lesquelles il existe des points de mesures (comme la température, ou l'altitude). Ces méthodes sont les plus usuellement implémentées dans les SIG. Donc nous en présentons certaines des plus connues, sachant que pour leur usage, il faut prendre garde à n'utiliser que des variables quantitatives relatives. Nous laissons de côté les familles de méthodes suivantes, pour lesquelles sont données les références principales :

- les splines [Dubrule 83],
- la triangulation par les polygones de Thiessen (Triangular Irregular Network, TIN),
- les arbres hiérarchiques bayésiens, [Mugglin 00],
- régression par cartes topologiques [Badran 97].

4.2.3.3.3.1 Surface de tendances (*Trend Surface Analysis*) Les surfaces de tendances sont une interpolation globale polynomiale au sens des moindres carrés. Elle vise à ajuster et à lisser une surface mathématique à partir des valeurs des points échantillonnés. Dans la figure 4.16, cette méthode correspond au passage de A à E (réduction au centroïde), puis E à C (calcul de surface). La surface créée peut être d'un ordre variable, mais plus l'ordre est élevé, plus l'interprétation des résultats est complexe. Cette méthode est aisée à mettre en œuvre, mais le lissage ainsi effectué peut être excessif et alors masquer des variations locales. Cette méthode n'est pas adaptée si la variable étudiée a une distribution complexe sur le territoire [Zaninetti 05].

4.2.3.3.3.2 Moyennes mobiles spatiales - *Inverse Distance Weighting (IDW)* Les filtres spatiaux (*focal functions*) constituent une autre famille de méthodes déterministes, mais locales. L'IDW est un type de filtre spatial, où une grille de carroyage (matrice) aussi fine que possible est appliquée sur l'espace d'étude pour simuler une surface continue. Dans la figure 4.16, cette opération correspond au passage de A à C, où la valeur des cellules de C prend initialement une valeur proportionnelle à la surface intersectée avec les unités du maillage A. Le calcul de la variable dans chaque cellule de la grille est ensuite pondéré par la valeur des cellules de son voisinage à l'aide d'une fonction décroissante de la distance : ce peut-être une fonction inverse ou bien une exponentielle négative. Le voisinage est défini par l'utilisateur en précisant un rayon de voisinage, ou portée. La taille du rayon (portée) a un effet sur le niveau de lissage des données : trop grand, on ne voit plus les phénomènes locaux ; trop petit, on les met en exergue. La mise en œuvre nécessite peu d'itérations et repose sur une hypothèse d'auto-corrélation spatiale. La carte résultante est « bruitée » et peut nécessiter une agrégation par classification.

4.2.3.3.3.3 Krigeage(s) Le krigeage est une méthode fréquemment utilisée pour éviter les questions de MAUP et de COSP, alors qu'à l'origine, cette technique, inventée par un ingénieur sud-africain, M. Krige, visait à déterminer la distribution spatiale de minerais à partir d'un ensemble de forages miniers (mesures ponctuelles donc). C'est cependant Matheron [Matheron 63] qui a formalisé l'approche en utilisant les corrélations entre les forages pour en estimer la répartition spatiale. Elle modélise la variabilité spatiale d'observations ponctuelles pour créer une surface continue. Le krigeage consiste à déterminer la pondération de chacun des points environnants selon le degré de similarité entre les valeurs. Celle-ci est établie à partir de la covariance entre les points, fonction de la distance entre ceux-ci. Cette modélisation de la similarité spatiale est intégrée dans des opérations d'interpolation spatiale. Le krigeage est ainsi sensible à la distribution spatiale du semis de points, à la portée choisie, et au choix de la fonction d'interpolation. Il existe plusieurs formes de krigeage qui dépendent de la connaissance que l'utilisateur a des caractéristiques statistiques de la variable étudiée : évolution ou non de la variable avec le temps (stationnarité), connaissance ou non de la moyenne de la variable, pour les plus générales [Cressie 91]. Basée sur un algorithme probabiliste, cette méthode fournit les erreurs d'estimations. Par contre, sa mise en œuvre requiert un grand échantillon en entrée (plus de 100 individus) et de nombreux jeux de données ne présentent pas de dépendance spatiale claire.

4.2.3.3.3.4 Interpolation pycnophylactique L'interpolation pycnophylactique, développée par [Tobler 79] vise à générer une surface lissée à partir de données connues sur un support polygonal en préservant les effectifs (ou masse) ou le volume sur les régions d'origine. Elle ressemble aux filtres spatiaux dans sa première étape : une grille de carroyage très fine est superposée à la zone d'étude et les cellules prennent, au départ, la valeur exacte des effectifs de la région à laquelle elles appartiennent (passage de A à C dans la figure 4.16). Une fenêtre de filtre est déplacée sur la grille et remplace les valeurs

de chaque cellule par la moyenne pondérée des cellules voisines (4 ou 8 cellules, au choix de l'utilisateur). Mais spécifiquement, après chaque itération, la méthode pycnophylactique contraint l'ajustement des nouveaux effectifs de manière à garantir la continuité avec les mailles voisines et conserver sur chaque maille de départ la masse (les effectifs) mesurée. Le processus se termine lorsque tout ajustement supplémentaire est plus petit que la tolérance spécifiée par l'utilisateur et que la surface est lissée (comme l'illustre la figure 4.19). L'interpolation pycnophylactique a l'avantage de conserver les masses, ou effectifs, et d'éviter d'avoir à réduire les mailles irrégulières à leur simple centroïde pour estimer une surface continue. En revanche, c'est un modèle qui gomme les changements brusques (par exemple des poches locales de forte densité de population à l'intérieur d'une maille) [Rase 01], car elle s'appuie sur l'hypothèse d'une progression uniforme de la densité d'effectifs.

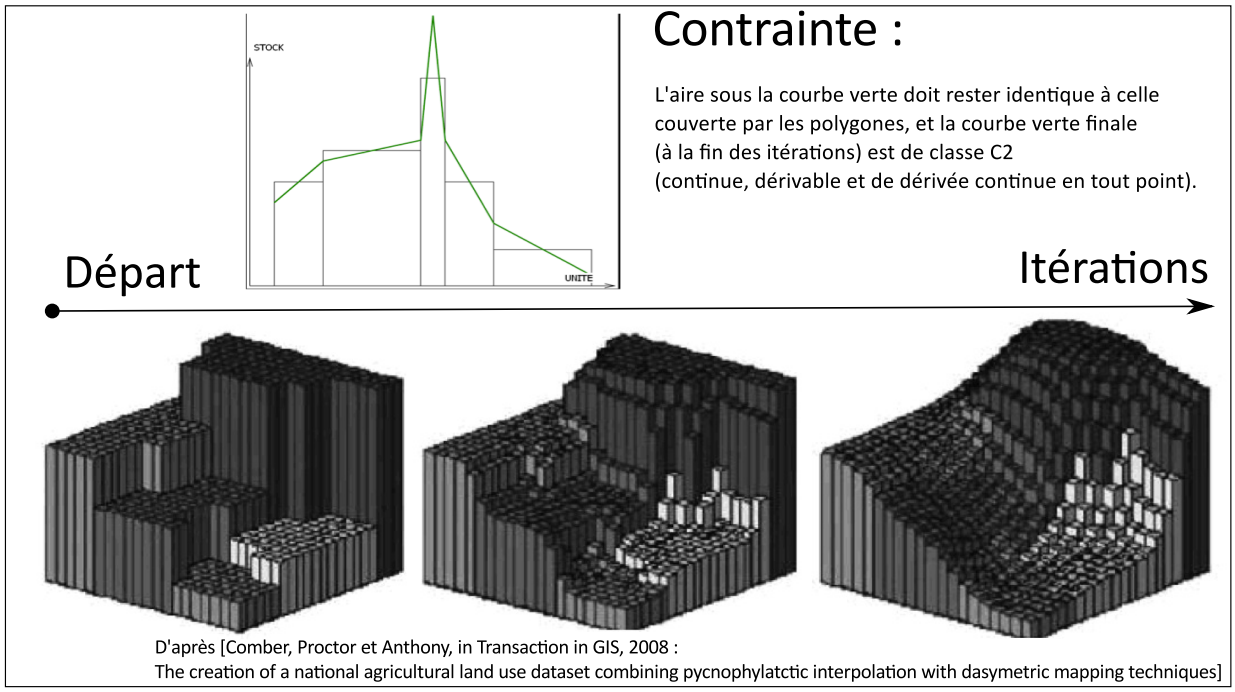


FIGURE 4.19 – Fonctionnement de la méthode pycnophylaticque.

4.2.3.3.3.5 Les lissages par la méthode des noyaux - la méthode du potentiel Les méthodes de lissage par noyaux sont plutôt réservées au lissage d'une distribution d'objets ponctuels supposée exhaustive qu'à l'interpolation d'une surface de densité à partir d'un échantillon de mesures ponctuelles [Zaninetti 05]. Elles s'appliquent également à des localisations pondérées par une variable de dénombrement, qui peut résulter de l'agrégation des objets géographiques dans un maillage d'unités territoriales administratives [Bracken 89]. Il ne faut pas lisser des taux avec ce type de méthode, mais des variables quantitatives positives possédant les propriétés d'additivité. La *méthode du potentiel* qui fait partie de cette famille de méthodes, est utilisée sur ce type de données [Grasland 00]. Les lissages par noyaux produisent une surface de densité qui peut être interprétée comme la densité de probabilité d'une loi continue sous-jacente au phénomène étudié. Pour cela, une *fonction d'interaction spatiale* f doit être définie, correspondant au modèle de diffusion du phénomène, qui produit une estimation continue aréolaire de densité cumulée autour des points du semis. Également, l'utilisateur doit choisir la *portée* (ou rayon d'influence) p de l'aire estimée autour de chaque point du semis.

Les lissages sont des méthodes dont les deux paramètres traduisent des hypothèses (économiques, sociologiques) associées aux interactions entre les acteurs sur le territoire, qui sont observées par la variable mesurée. La fonction d'interaction spatiale f intègre les hypothèses concernant les lois de diffusion dans l'espace associées au phénomène étudié : la fonction gaussienne, la loi uniforme, la fonction exponentielle, la loi de Pareto sont des exemples de modèle d'interaction (voir figure 4.20). La fonction uniforme, proche de l'estimateur naïf de densité, est la plus rudimentaire. La fonction gaussienne filtre les variations locales abruptes, et fournie une estimation continue, étendue aux limites de la zone d'étude, la décroissance de f se faisant selon une exponentielle négative. La loi de Pareto convient pour un modèle à interaction de longue portée car elle propose la décroissance suivant une puissance inverse. Par exemple, cette méthode permet de modéliser et d'étudier la propagation des épidémies touchant l'homme : leur diffusion pourra se faire soit sur de longues distances, soit sur de courtes distances, suivant le rayon d'action de l'élément contaminant [Grasland 05a]. La portée p modélise la distance moyenne d'action d'une masse sur son voisinage. La portée peut être interprétée comme l'échelle spatiale de représentation choisie : plus la portée est petite, plus l'échelle d'analyse est fine. Lorsque la portée grandit, ce sont les structures globales de répartition qui sont mises à jour, et le phénomène est généralisé (voir les figures 4.21 et 4.22).

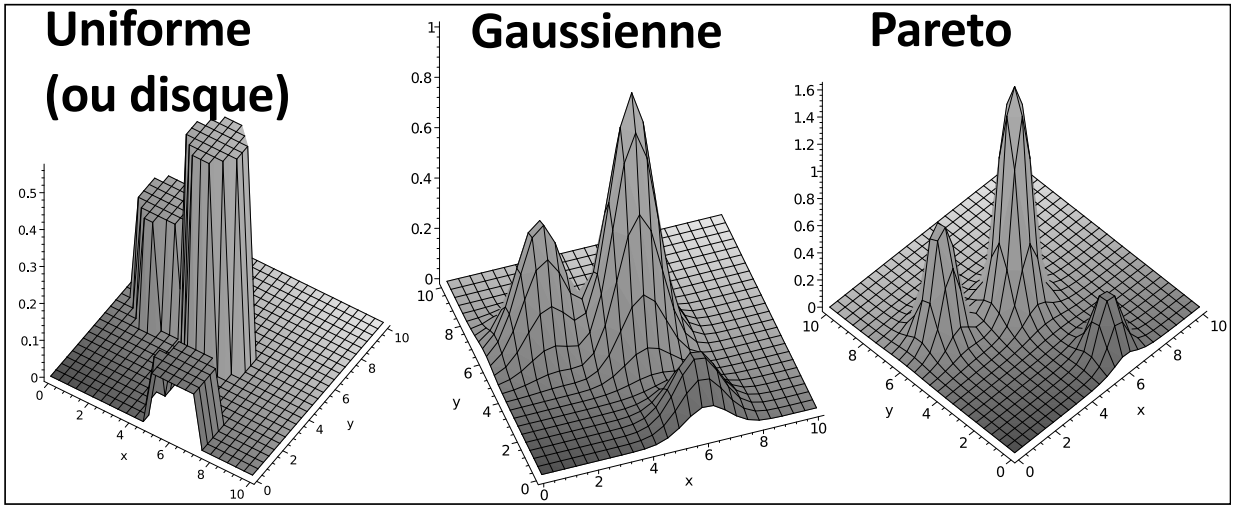


FIGURE 4.20 – Surfaces produites par trois fonctions d'interaction différentes à partir des mêmes données.

Ce type d'interpolation permet d'obtenir des représentations cartographiques continues de phénomènes spatiaux discrets, qui sont affranchies du maillage de collecte initial des données. En effet, la surface de densité qui est calculée ne dépend pas de la forme, ni de la finesse du maillage de collecte des données [Grasland 06]. Du point de vue méthodologique, de type d'interpolation s'apparente aux méthodes de traitement du signal par déconvolution du signal échantillonné. Cependant, il ne résout pas certains problèmes rencontrés aussi dans d'autres méthodes. Les données sont mesurées sur un espace limité, sur les bordures du duquel le lissage ne peut pas fonctionner de façon isotrope. L'étendue des marges présentant un défaut d'évaluation est directement proportionnelle à la portée. De plus, en dessous d'une certaine portée, la méthode devient imprécise ; la portée minimale peut être calculée par l'analogue du théorème de Nyquist : elle vaut deux fois la taille maximale des mailles [Nyquist 28]. Enfin, un maillage trop hétérogène conduit à faire un compromis entre les portées minimum associées à chaque classe de taille d'unité.

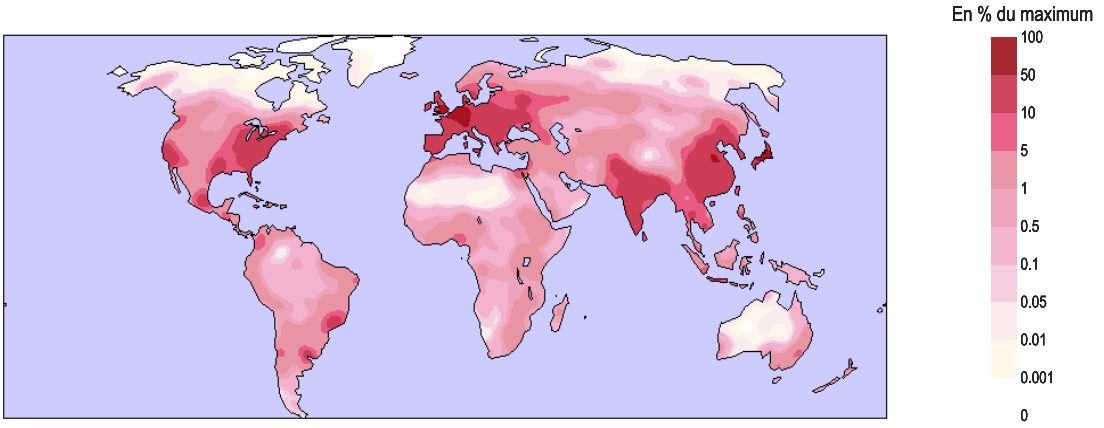


FIGURE 4.21 – Lissage par un potentiel gaussien de la population âgée de plus de 80 ans, sur une portée de 500 km. Source des données : ONU - WPP 2008. [Pison 11]

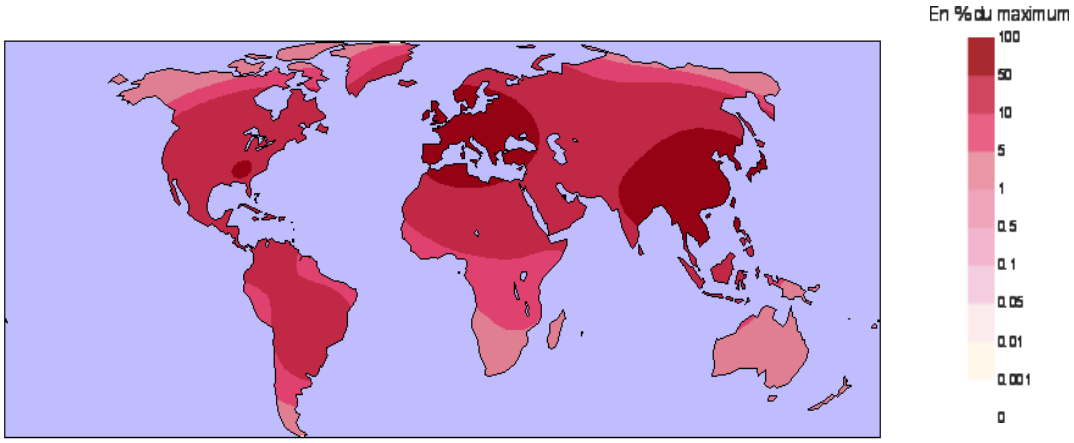


FIGURE 4.22 – Lissage par un potentiel gaussien de la population âgée de plus de 80 ans, sur une portée de 2000 km. Source des données : ONU - WPP 2008. [Pison 11]

4.2.3.4 Mise en oeuvre du transfert

Il n'existe pas à l'heure actuelle d'outils permettant de réaliser ce transfert de façon automatique. Avec la présentation du modèle de transfert que nous venons de faire, il apparaît que, rigoureusement, il ne faudrait transférer que des stocks. Ce qui signifie décomposer les ratios en leur deux composantes, numérateur et dénominateur, pour les transférer séparément. Comme un grand nombre de ratios ont pour dénominateur la population, ou la surface, il s'agit de savoir au moins évaluer ces deux variables.

Une grande majorité d'utilisateurs ont encore une approche très manuelle, basée sur des règles de proportionnalité entre surfaces échangées et donc supposent implicitement l'uniformité de la densité de répartition du caractère observé, hypothèse peu satisfaisante en général. Bien que ces approches puissent être systématisées avec l'emploi de bases de données spatio-temporelles fondées sur le paradigme du *Space-Time Composite*, (ou Plus Petit Commun Dénominateur Spatial), [Norman 03], ces approches restent peu pratiques à mettre en œuvre en raison des problèmes de mises à jour de la base de données.

Toutefois, en dehors des considérations liées à la maintenance d'une base de donnés spatio-temporelle, l'utilisation d'une grille régulière comme Plus Petit Commun Dénominateur Spatial fournit un expédient efficace pour le calcul de séries temporelles. C'est pourquoi, cette solution a rencontré la faveur du programme de l'UNEP qui diffuse les valeurs de population sur une grille de 2,5 secondes d'arc de résolution (soit environ 4,63 km de côté pour les cellules au niveau de l'équateur), calculées d'après la méthode décrite dans [Deichmann 01], par pondération surfacique simple. Par ailleurs, à l'heure actuelle, avec la diffusion massive d'images satellitaires d'occupation du sol, un nombre de plus en plus important d'utilisateurs se tournent vers l'usage de grilles de population (de 1 km2 de résolution par exemple) créés à partir de l'analyse de ces images, croisées avec d'autres sources de données (comme la luminosité nocturne des lieux géographiques) [Gallego 10]. La méthode de calcul prend en compte les usages du sol comme variable auxiliaire et fournit donc *a priori* une estimation de la population plus fiable qu'avec la pondération surfacique simple. Ces grilles qui établissent la densité de population tous les cinq ans depuis 1990, et dont le rythme de production s'accélère, servent de support pivot pour réaliser le transfert de variables entre maillages non-alignés, car la densité de population est la variable auxiliaire de nombreux indicateurs territoriaux [Nordhaus 05], [Reibel 07]. La méthode proposée est suffisamment simple à mettre en oeuvre pour tous les utilisateurs de statistiques et peut traiter un plus grand volume de données simultanément sur de grandes étendues.

Elle présente cependant plusieurs inconvénients. Premièrement, la gestion des données requiert alors de transférer l'ensemble des statistiques dans des bases de données conçues pour des formats matriciels, et dans ces cas, l'information liée à l'imbrication des zonages est très souvent perdue. Lorsqu'elle est traitée, comme dans certains SOLAP [Body 03, Miquel 03, Tchounikine 05], c'est d'une manière très peu satisfaisante (voir section 2.2.1.2 page 56 du chapitre 2), car les relations d'appartenance semblent alors figées dans le temps. De plus, il faut noter qu'elle est peu précise, et il s'agit de choisir de façon appropriée l'échelle à laquelle on souhaite utiliser les données transférées. Par exemple, plusieurs études [Lwin 09], [Plumejeaud 09b], basées sur l'usage du bâti capturé par imagerie satellitaire, montrent qu'à l'échelle des IRIS (infra-urbain), l'usage de ce type de données n'est pas suffisamment fiable car il ignore le nombre de logements vacants, les pratiques d'habitation, et d'autres aspects de la vie sociale qui ne peuvent être capturés par l'imagerie satellitaire et conduisent à mésestimer la population dans les unités infra-urbaines. Cette critique est aussi formulée pour des données concernant les exploitations agricoles [Schmit 06]. Par ailleurs, la couverture temporelle (et même spatio-temporelle) de ce type de données issues de l'imagerie satellitaire est limitée dans le passé.

4.3 Les outils pour l'évaluation de la qualité

Par rapport à la problématique que pose l'exploration et l'évaluation de la qualité, nous recensons les outils qui mettent en œuvre les méthodes de recherche de valeurs exceptionnelles précédemment exposées. La plus grande partie des outils est issue du domaine de l'analyse exploratoire des données, dont les méthodes sont de plus en plus reprises et exploitées à grande échelle par les outils de fouille de données spatiales, [Zeitouni 00], [Guo 09]. Par ailleurs, les relations hiérarchiques entre unités sont rarement exploitées, en dehors d'HyperAtlas qui est un outil du groupe de recherche HyperCarte [Grasland 05b]. L'usage de méthodes statistiques pour l'analyse de données se répand dans d'autres domaines, comme par exemple l'informatique décisionnelle, et si le focus n'est plus forcément mis sur l'analyse spatiale ou temporelle, le rôle que doit jouer un utilisateur dans l'analyse de la qualité est mieux cerné.

4.3.1 Les outils de l'ESDA

L'EDA est très liée à la représentation graphique des données, car les graphiques fournissent à l'analyste des vues nouvelles suscitant spontanément des questions [Bertin 67]. Par ailleurs, l'outil visuel emploie au mieux les capacités de l'esprit humain pour l'observation, la comparaison et la détection de patrons. Ainsi, Banos explique que *via* ce type d'outil, les « formidables capacités humaines, en termes de visualisation, d'intuition, de raisonnement par analogie et de génération d'hypothèses, sont ainsi pleinement mises à contribution, dans le cadre d'une relation homme-machine ludique et réaliste, exploitant au mieux les qualités de chacune des parties » [Banos 01]. De ce domaine fertile est née l'Analyse Exploratoire de Données Spatiales (*Exploratory Spatial Data Analysis*, ESDA) qui est une spécialisation de la discipline pour les données à références spatiales et temporelles (bien que le temps n'apparaisse pas dans sa dénomination). La promotion de l'ESDA a été largement le fait de chercheurs comme Luc Anselin [Anselin 93], Robert Haining [Haining 03], ou Gennady et Natalia Andrienko [Andrienko 06].

4.3.1.1 Fonctionalités et architecture requises

En tout état de cause, les capacités à la fois cartographiques et statistiques sont primordiales pour un outil d'analyse spatio-temporelle exploratoire. La reconnaissance, l'analyse et la mesure des formes d'association spatiale par le calcul de l'autocorrélation spatiale est une des fonctionnalités les plus classiques [Anselin 93]. Il s'agit également de disposer de méthodes pour la comparaison de différentes évolutions temporelles en vue d'identifier les différents formes d'évolution [Andrienko 01], [Andrienko 03b]. Enfin, la possibilité de détecter des associations thématiques entre différentes variables est aussi importante que les associations spatiales. En effet, ce sont ces dernières méthodes qui peuvent aider à trouver des variables auxiliaires utiles pour des manipulations plus complexes.

La méthode de gestion des données est une caractéristique importante de l'architecture d'un logiciel : lorsque que celle-ci est faite de manière adéquate, par l'emploi d'*adapteurs* à différentes sources de données, elle économise nombre de manipulations contraignantes comme la transformation des données d'un format donné à celui exigé par le logiciel. Une telle gestion des données facilite l'exploration d'un nombre conséquent de données puisqu'alors le croisement de différentes sources de données est immédiat : elles peuvent être issues de fichiers dans différents formats, ou de bases de données hétérogènes.

De même, il s'agit de proposer une structure pour les données facilitant l'analyse. Ainsi, dans le cas des statistiques spatiales, on s'aperçoit qu'elles reposent sur l'usage de la matrice de voisinage, structure de données qui peut représenter soit une distance discrète (les différents ordres de contiguïté), soit une distance continue (métrique ou de distance-temps). Si, à partir d'un modèle de données spatio-temporelles, on sait calculer les matrices de voisinage topologiques comme les matrices de distance métrique, on s'aperçoit que l'usage de distances pertinentes pour un géographe, comme la matrice de distance-temps entre unités spatiales, suivant des modes de transport différents, exige le stockage de ces matrices. A cet égard, [Lee 05] met en évidence la complémentarité qui existe entre le monde des SIG et celui de l'ESDA, où les SIG sont décrits comme des systèmes d'information dotés de capacités cartographiques. Cette proposition pour rapprocher les SIG et le monde de l'ESDA par le couplage des fonctionnalités de cartographie et d'analyse statistique est centrée sur le calcul d'indices d'association spatiale (Modèles d'Association Spatiale, MAS). [Lee 05] souligne la nécessité de fabriquer des matrices de voisinages spatiaux dans un SIG (voir figure 4.23). Cette proposition peut être étendue à l'analyse temporelle et multi-dimensionnelle, à condition que les structures de données soient capables de prendre en compte la variable temporelle.

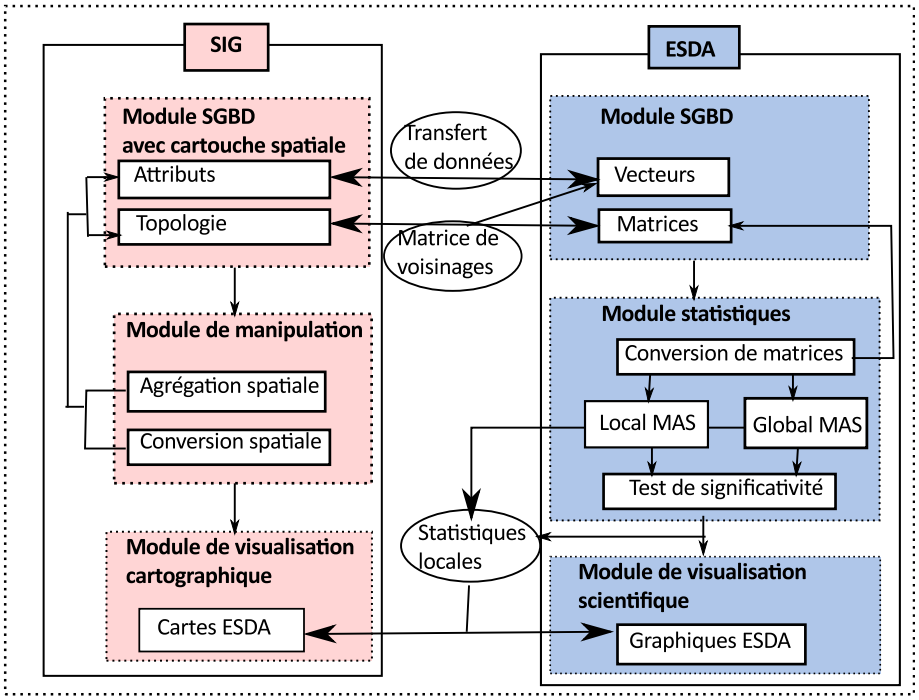


FIGURE 4.23 – Coupler les SIG et l'analyse exploratoire de données spatiales, d'après [Lee 05].

Par ailleurs, le volume des données traitées peut-être un frein à l'interactivité étant donné que les capacités de calculs requises par les méthodes d'analyse, de plus en plus complexes, sont assez élevées [Guo 09]. Les capacités de calculs sont donc également en prendre en compte et le couplage des outils avec des méthodes de calculs intensifs, et des super-calculateurs est un atout important. Par exemple, le calcul des coefficients d'autocorrélation spatiale augmente de façon exponentielle avec le nombre d'unités : il est en $O(n^2)$, si n est le nombre d'unités spatiales. Bien qu'Anselin recommande le pré-calcul de ces coefficients, plus précisément de la valeur du *spatial lag* pour chaque observation (c'est-à-dire de la valeur moyenne de la variable X considérée dans le voisinage de l'observation), [Anselin 93], cela ne nous semble pas optimal (ni même faisable) si on est en présence d'une quantité de variables X_i importante, et que, de plus, ce calcul est effectué pour tous les types de voisinage d'intérêt. Peut-être vaut-il mieux alors envisager de paralléliser le calcul de ces coefficients pour optimiser les temps de calcul, afin de répondre aux exigences de l'interactivité de la démarche exploratoire.

Enfin, ces méthodes présentent des vues différentes d'un même sous-ensemble de variables X_i : par exemple, la visualisation d'une distribution et les représentations comme le *bagplot* font partie de la vue thématique, la carte (que ce soit en représentation continue ou discrète) donne une vue spatiale des données, comme le nuage de dispersion de Moran, ou un corrélogramme, tandis que les diagrammes avec courbes d'évolution temporelles font partie de la vue temporelle. Il y a intérêt de construire les logiciels d'exploration sur le principe de synchronisation des vues, (le modèle *Model-View-Controler*) afin que lorsque un individu est repéré dans une vue, l'utilisateur puisse observer sa position dans une autre vue de son choix. Monmonier fut le premier à souligner l'intérêt de la synchronisation entre deux vues, l'une cartographique avec une carte chroplèthe, et l'autre statistique avec un diagramme de dispersion [Monmonier 89].

4.3.1.2 Critiques des outils existants

Le développement et la diffusion d'outils consacrés à l'ESDA, embarquant des méthodes d'analyse statistique, est en plein essort. La situation de 2011 est donc bien améliorée par rapport à celle de 2001 où ces outils se faisaient encore trop rares [Banos 01]. Le tableau 4.2 liste les outils les plus connus du domaine.

TABLE 4.2 – Références d'outils connus.

Libres	
Weka	[Hall 09] `http://www.cs.waikato.ac.nz/ml/weka/`
R (successeur de S-Plus)	[Bivand 08], `http://www.r-project.org`, en particulier les paquetages sp et spacetime
Protovis	`http://vis.stanford.edu/protovis/`
Grass (GIS + R)	`http://grass.itc.it/index.php`
QGis	`http://www.qgis.org`
GeoDa	[Anselin 04] `http://geodacenter.asu.edu`
SAGE	[Haining 03]
CrimeStat	`http://www.icpsr.umich.edu/icpsrweb/CRIMESTAT`
SADA	`http://www.tiem.utk.edu/sada/index.shtml`
SPACE-STAT	[Anselin 92]
SPIDER-REGARD	[Haslett 90], [Unwin 94]
XLisp-Stat	[Brunsdon 96] `http://www.stat.uiowa.edu/~luke/xls/xlsinfo/`
Commerciaux	
SAS	`http://www.sas.com/offices/europe/france/`
ESRI - The Spatial Statistics toolbox in ArcGIS 9.	`http://www.esrifrance.fr/Spatial_Analyst.asp`
SPSS	`http://www.spss.com/fr/`

Les sites suivants proposent des comparaisons et des descriptions détaillées de ces outils :
`http://fedc.wiwi.hu-berlin.de/xplore/ebooks/html/csa/node82.html`
`http://www.geovista.psu.edu/grants/VisEarth/refs1.html`
`http://www.cartomouv.parisgeo.cnrs.fr/index.php?page=accueil`
Il apparaît ainsi qu'un nombre important d'outils ont été développés dans le cadre de l'ESDA, dont certains s'attachent plus à l'aspect analyse statistique, d'autres plus à l'aspect visualisation des données. La plupart peuvent être réutilisés, soit dans leur ensemble, soit comme des composants, pour la recherche de valeurs exceptionnelles. Cependant, ces outils ne sont pas explicitement dédiés à l'analyse de la qualité des données, exception faite d'un module de SAS.

Des outils d'analyse spatiale comme SADA, Geoda, CrimeStat proposent des fonctions d'analyse statistique spatiale, couplées à des fonctions de visualisation et d'exploration de données, qui sont assez intéressantes lorsque l'on souhaite faire de l'estimation de valeur manquante : interpolation par krigeage complexe, simulation spatiale. Cependant, en dehors de Geoda, qui est open-source, le code réalisant les fonctionnalités de ces outils n'est pas réutilisable par d'autres développeurs. De plus, ces logiciels importent des fichiers décrits dans des formats propriétaires (Shapefile, DBF, Excel, etc.) sans fournir de connexion vers des bases de données spatiales, ni temporelles.

D'autres outils, comme QuantumGis (écrit avec Python), TerraLib (écrit avec C++) ou bien GRASS GIS (écrit avec C ou Python) proposent une lecture de données depuis des bases de données libres (PostgreSQL, ou MySQL), couplées à des fonctions de visualisation des données avec une interface cartographique.

Bien que pour ces derniers logiciels, la liste des fonctionnalités d'analyse spatiale disponibles immédiatement pour l'utilisateur soit plus restreinte, il faut noter qu'elle peut être étendue. En effet, ces logiciels offrent la possibilité d'intégrer des scripts pour l'analyse statistique programmés avec R. La librairie open-source R se révèle être un langage d'expression privilégié pour de nombreux statisticiens [Templ 09]. Par exemple, Geoda est devenu libre, et il s'appuie sur R : ses contributeurs enrichissent la bibliothèque R avec de nouvelles méthodes régulièrement. De même, certains services Web pour l'interpolation de données spatiales comme, par exemple, INTAMAP [Pebesma 10] sont aujourd'hui basés sur R. Si l'idée de services de calculs distribués nous semble très pertinente, l'interface d'utilisation nous apparaît comme trop sommaire. Dans [Hengl 08], l'usage de R comme outil capable de s'interfacer avec des fonctions SIG est également promu. A ce titre, bien que l'outil s'interface principalement avec le SIG commercial ArcGIS vendu par ESRI, nous pouvons citer l'initiative open-source de [Roberts 10] proposant une suite d'outils open-source pour l'analyse des écosystèmes marins, la plateforme *Marine Geospatial Ecology Tools* (MGET). Développée avec Python, cette plate-forme permet d'intégrer des scripts écrits en R pour analyser les données à références spatiales. Enfin, dans sa dernière version, le module SpatialAnalyst de ArcGIS vendu par ESRI propose d'exécuter des scripts R.

Cependant, force est de constater qu'aucun de ces outils ne fournit d'informations sur les métadonnées sous un format non textuel, par exemple, au moyen de cartes ou de représentations interactives qui permettraient à l'utilisateur de mettre facilement en relation les informations collectées sur le jeu de données qu'il analyse avec les résultats calculés. Ces logiciels ignorent tout à fait la présentation des métadonnées associées aux données, puisque le schéma d'importation des données n'intègre pas l'import des métadonnées : les données spatiales, en particulier, sont simplement réduites à l'association d'un fond de carte (l'ensemble des géométries) et d'attributs thématiques, sans aucune information sur la provenance des valeurs, ou la description des méthodes d'évaluation des données. L'affichage des fiches de métadonnées se répand dans les SIG actuels, mais encore sous une forme qui reste très primitive (un document textuel séparé), et leur consultation est séparée des données : les valeurs continuent d'apparaître dans les interfaces de façon assez dépouillée. Enfin, aucun de ces outils ne permet d'exploiter les relations d'appartenance entre les unités en vue de pratiquer une analyse de la variance qui permettrait de distinguer des valeurs exceptionnelles au regard des classes formées dans la hiérarchies des unités statistiques spatiales.

4.3.2 Prise en compte des relations d'appartenance avec HyperAtlas

Le principe d'analyse de la variance par mesure des écarts est mis en œuvre dans HyperAtlas, un logiciel issu des travaux du groupe de recherche Hypercarte [Grasland 05b]. Les analyses proposées offrent une vue d'ensemble de la position relative d'une unité territoriale et de l'information statistique associée par rapport à différents contextes à travers un atlas de cartes interactives. L'utilisateur ayant choisi interactivement le ratio de deux indicateurs qu'il souhaite étudier, par exemple le PIB par habitant, le taux de chômage, ou la part des actifs dans la population, les cartes montrent soit la position spatiale « absolue » de chaque unité dans une représentation choroplèthe classique en dégradé de couleur, soit la position relative de cette unité par rapport à trois contextes différents dans trois autres cartes choroplèthes.

Ces trois cartes, dites cartes d'écart [5], montrent pour chaque unité son écart à la moyenne définie dans un contexte qui peut-être :

- une référence globale unique : par exemple, l'union européenne des 15, des 27, ou l'aire des Pays de l'Europe Centrale et Orientale (PECO).
- une région d'appartenance de niveau supérieur à l'unité : son département, sa région, son pays.
- une région formée de ses voisins, suivant différents critères : la contiguïté d'ordre 1, ou bien des seuils de distance. Les distances sont issues de matrices de distance précalculées entre unités d'un niveau de maillage donné, pour un certain moyen de transport. Par exemple, les régions à moins d'une heure de transport en voiture forment un voisinage d'une région donnée.

Le premier contexte est dit « général », le second « territorial » et le troisième « spatial » [6]. Par ailleurs, une quatrième et dernière carte choroplèthe résume la position des unités relativement à ces trois contextes et permet ainsi de construire une synthèse des différenciations locales, territoriales ou globales (figure 4.24).

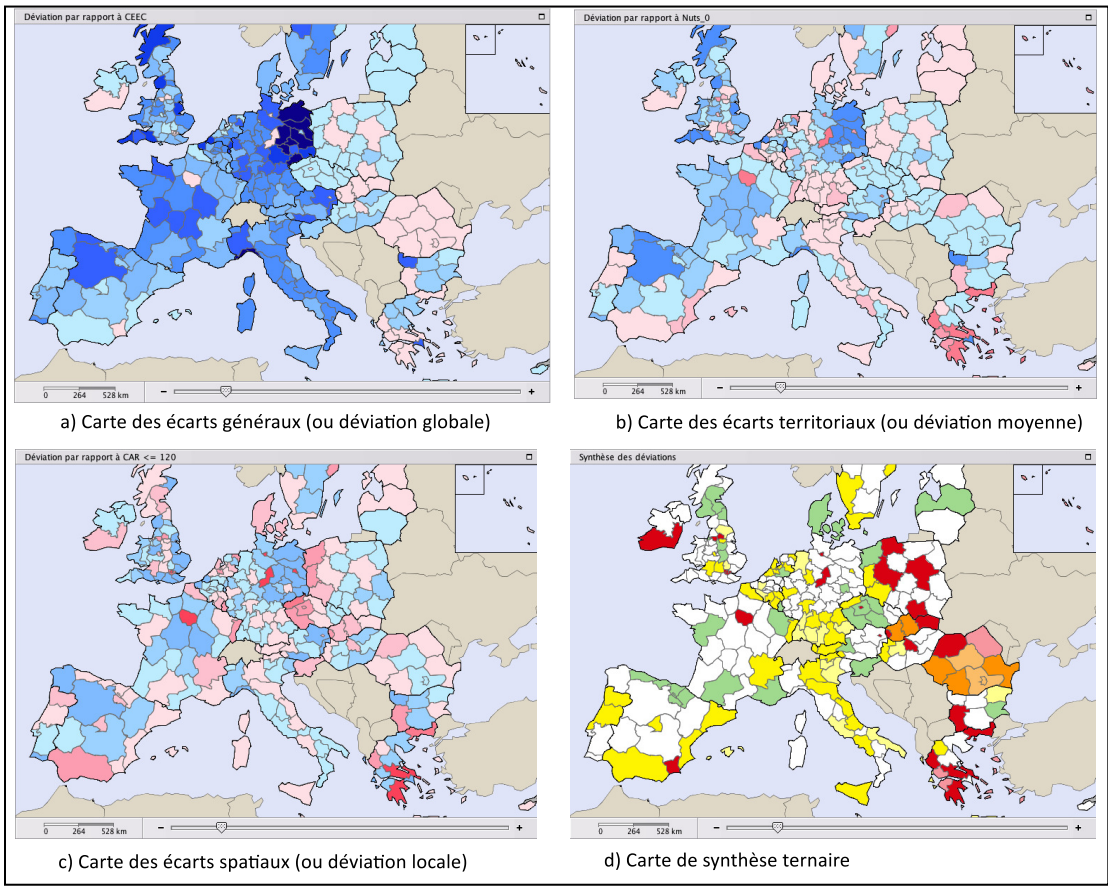


FIGURE 4.24 – Extrait des trois cartes d'écart a) général, b) territorial, c) spatial et de d) synthèse d'HyperAtlas (v1.0) pour l'étude de la part des actifs dans la population en 2030.

5. Egalement nommées cartes de déviation avant 2010.
6. Avant 2010, ces trois contextes étaient désignés respectivement « global », « médium », et « local ».

Les cartes d'écart global et local proposent des mesures de l'association spatiale somme toute assez classiques tandis que la carte d'écart territorial constitue une mesure originale du caractère exceptionnel des valeurs, en les rapportant à la moyenne des unités englobantes. La carte de synthèse, qui permet de combiner ces analyses, apparait aussi être un outil puissant, car elle permet de repérer très vite quelles sont les unités qui ont une valeur exceptionnellement haute ou basse par rapport à ces trois contextes. Selon Waniez [Waniez 10], l'avancée scientifique est réelle et permet enfin de prendre en considération l'idée selon laquelle « la réalité apparaît différente en fonction de l'échelle d'analyse ».

Un des grands avantages de cet outil réside dans sa simplicité d'utilisation. Cet outil est en effet destiné à un public d'aménageurs du territoire et il a été par exemple diffusé au Parlement Européen dans le cadre d'une étude sur la décroissance démographique en Europe [UMS 2414 RIATE 08] : il est hors de question d'exiger des utilisateurs dans ce cadre de savoir configurer des méthodes statistiques aussi complexes que celles qu'on trouve dans des outils comme GeoDa ou CrimeStat. HyperAtlas est plus simple à prendre en main que des outils munis de méthodes statistiques plus évoluées mais dont l'interprétation peut être difficile, et la configuration ardue.

Concernant l'analyse des évolutions temporelles, un menu permet de choisir des indicateurs suivant différentes dates de validité, et de calculer un taux d'évolution. Il est alors possible d'étudier la distribution spatiale de cette évolution par l'analyse de la variance. Prenons, par exemple, le rapport des valeurs de l'espérance de vie en bonne santé au niveau régional en Europe calculé entre 2030 et 2005. La carte de la figure 4.25 montre la synthèse des écarts à la moyenne européenne, la moyenne nationale et la moyenne locale (calculée par contiguïté). Il ressort de façon très évidente que, si pour toutes ces unités, l'évolution anticipée en 2030 est une hausse de l'espérance de vie, les unités faisant partie de l'ancien bloc de l'Est bénéficient plus fortement de cette hausse, ainsi que la région du Nord-Est de l'Écosse, qui apparaît ici comme une valeur exceptionnelle.

Toutefois, cet outil ne gère actuellement pas la dimension temporelle de façon pleinement satisfaisante. En effet, le maillage d'étude est obligatoirement le maillage de référence ayant servi à l'harmonisation des données. Par exemple, les cartes de la figure 4.24 montrent une projection de la part des actifs dans la population totale en 2030, avec un zonage du Danemark au niveau régional correspondant à la version de zonage de 2003.

Ainsi, pour HyperAtlas, comme pour l'ensemble des outils d'analyse spatiale et d'exploration spatio-temporelle connus à ce jour pour l'étude de données zonales, l'utilisateur se voit imposé une certaine version de zonage pour son étude. Or, il n'y a pas de raison de penser qu'une version de zonage soit plus pertinente qu'une autre *a priori*, et nous souhaiterions avoir le choix de ce maillage de référence.

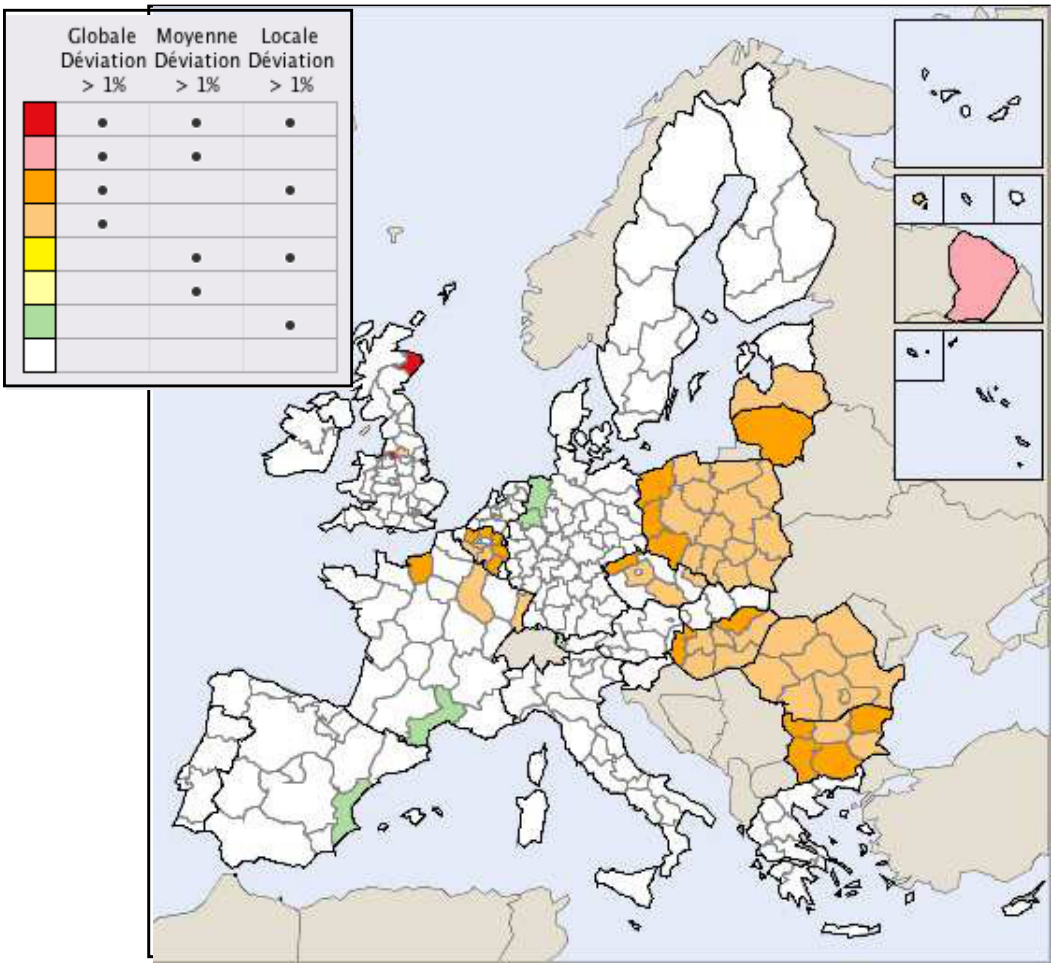


FIGURE 4.25 – Analyse spatiale du taux de variation de l'espérance de vie en bonne santé entre 2005 et 2030 (projection).

4.3.3 Prise en compte de l'utilisateur et des métadonnées dans l'évaluation de la qualité

Dans le domaine de l'informatique décisionnelle, la qualité des données possède un rôle vital. En premier lieu, [Daniel 08] comme [Chengalur-Smith 99] soulignent l'importance que revêt pour les utilisateurs la publication et la consultation *des métadonnées avec* les données, afin qu'ils puissent réajuster leurs décisions en fonction de la qualité connue des données. Des outils et des méthodes de contrôle de la qualité des entrepôts de données ont été développés, [Vaisman 07], reposant sur un travail de fond pour déterminer les critères de qualité [Wand 96].

Il est à noter que l'utilisateur doit jouer un rôle déterminant dans cette évaluation de la qualité [Daniel 08] : il doit à la fois être informé de la qualité, mais également, il peut lui même produire des avis concernant la qualité des données. Ces recherches recommandent donc la conception de systèmes interactifs d'évaluation de la qualité. Ainsi, certains modules de SAS [7], qui sont dédiés à l'analyse de la qualité de données, proposent une interface interactive pour que l'utilisateur intervienne dans le processus d'évaluation de la qualité.

7. http://www.sas.com/offices/europe/france/software/technologies/dataquality.html

Également, dans le domaine de la climatologie, une société, Météo France International, publie un outil, Clisys [8] qui propose un contrôle de cohérence des données climatiques. Pour cela, un *flag* qualité est associé à chaque élément climatique. D'après le site Web de la société, les paramètres mesurés sont les suivants :

- Tolérance : exécuté automatiquement lors du processus d'acquisition pour vérifier la validité des éléments observés.
- Cohérence interne : exécutée par des experts [9] sur le contenu de la base de données pour valider la donnée ou la considérer comme douteuse.
- Cohérence temporelle : exécutée par des experts sur le contenu de la base de données pour tester la variation d'un élément dans le temps.
- Cohérence géographique : visualisaton géographique des données climatiques avec lien direct aux fonctions de modification des données.

Clisys propose un système interactif avec visualisation des données et de leur métadonnées. Dans ce système, les métadonnées et les données sont stockées dans un même SGBD. Ce système est ouvert et permet aux experts d'intégrer leurs propres connaissances.

Ces outils ou travaux insistent donc sur l'importance de proposer des systèmes interactifs pour l'évaluation de la qualité des données, prenant en compte l'avis des experts, d'une part, et, d'autre part, de proposer un accès simultané aux métadonnées et aux données pour ces experts.

4.4 Conclusion

Dans ce chapitre, nous avons montré que l'analyse de la qualité, en particulier de la précision sémantique des données, pouvait bénéficier des méthodes de recherche de valeurs exceptionnelles par analyse statistique appliquées à toutes les dimensions de l'information, que ce soit la dimension spatiale, temporelle ou thématique. Ces méthodes ont été particulièrement développées dans le cadre de l'analyse exploratoire de données à références spatiales et temporelles (ESDA), un sujet de recherche en pleine expansion qui s'oriente aujourd'hui vers les domaines de la fouille de données spatiale et temporelle [Guo 09]. Dans ce domaine, un nombre conséquent d'outils ont été développés, basés sur une approche interactive où l'utilisateur est invité à découvrir les données et à se questionner sur les relations existant entre les différentes valeurs. Ces techniques sont un apport réel à la compréhension des données, et de plus en plus de SIG cherchent à les intégrer.

Faisant appel à des techniques de plus en plus complexes, ce domaine n'appréhende pourtant pas encore de façon absolument systématique l'analyse multi-niveau des données, bien que plusieurs auteurs aient montré le grand intérêt d'une telle approche, proposant même une démarche systématique pour l'analyse multi-niveau [Mathian 01]. Il s'agit aussi de souligner que les données à référence zonale, comme l'est l'information statistique territoriale, présentent une complexité mais aussi une richesse que de nombreux logiciels n'intègrent pas encore. La complexité vient du phénomène dit du « MAUP », qui n'est un problème comme le souligne Openshaw [Openshaw 87, Openshaw 96] que si on se satisfait du maillage imposé pour l'étude, mais pourrait également être la solution à l'analyse de nombreux phénomènes géographiques. En effet, en filtrant « la variance », l'agrégation ou la désagrégation des

8. `http://www.mfi.fr/fr/clisys-the-management-tool-for-all-climate-data-fiche-produit.php`

9. Les experts mentionnés sur ce site sont explicitement ceux qui vont utiliser le logiciel, du côté du client.

données permet d'envisager les relations entre variables à différentes échelles de raisonnement. Le logiciel HyperAtlas prend en compte ces différents niveaux d'analyse, et s'utilise pour repérer des valeurs exceptionnelles et analyser les inégalités socio-économiques à diverses échelles.

Le rôle de l'utilisateur dans l'analyse des données a aussi été mis en avant, que ce soit dans l'ESDA ou d'autres domaines utilisant les mêmes méthodes d'analyse statistique. Ces techniques de recherche de valeurs exceptionnelles font cependant appel à un niveau plutôt élevé de connaissances de modèles statistiques de la part de l'utilisateur. Le décalage entre le niveau du public utilisateur et des experts concevant ces outils peut aboutir à une utilisation malheureuse de ces outils. Ce problème, déjà mentionné par Openshaw [Openshaw 94, Openshaw 96], reste d'actualité.

Également, de plus en plus, ces méthodes font appel à une puissance de calcul élevée, alors qu'un grand nombre de logiciels d'analyse ou de SIG intégrant ces méthodes reste éloigné du domaine du calcul intensif. Toutefois, une tendance se dessine pour l'adaptation de ces logiciels au calcul intensif. Ainsi R embarque un environnement adapté pour la programmation et l'exécution en mode parallélisé de méthodes statistiques [Rossini 07].

Enfin, si la production de métadonnées s'organise de façon plus systématique au niveau des producteurs de l'information statistique territoriale, il n'existe pas de modèle systématique d'exploitation et d'enrichissement de ces connaissances dans les outils d'exploration de données actuels. En effet, si l'importance d'utiliser des métadonnées a été soulignée dans domaine de l'informatique décisionnelle, dans le domaine de l'ESDA ou de la fouille de données, les outils ne permettent de mettre en relation les analyses produites avec les informations de qualité contenues dans les métadonnées.

Deuxième partie

Proposition

Chapitre 5

Un modèle pour des hiérarchies multiples et évolutives

Ce chapitre présente un modèle destiné à organiser les données statistiques collectées sur les différents types de zonage existants, à toutes les échelles. Notre objectif est de permettre de collecter des données, issues de sources hétérogènes, qui évoluent dans le temps. Les données sont les valeurs d'indicateurs (ou variables) statistiques qui sont associées à des géométries qui forment, par assemblage, le support géographique. La conception de ce modèle est fondamentale : le modèle doit rendre compte du changement spatial, et des relations existantes entre les territoires, en particulier, l'organisation et l'histoire des territoires, quelque soit les territoires et l'organisation considérée. C'est un modèle pour des hiérarchies territoriales multiples et évolutives.

5.1 Un modèle objet indexé par des événements de changement

Nous présentons ici les deux principales motivations qui justifient notre modèle : la gestion simultanée de plusieurs nomenclatures, et l'intégration d'une ontologie historique des territoires associée aux événements de changement. Ensuite, le modèle est complètement décrit et des exemples viennent appuyer son utilité et les possibilités qu'il offre.

5.1.1 Motivations du modèle

La motivation du modèle vient de ce que nous souhaitons comparer et associer un maximum d'indicateurs statistiques, issus de bases de données constituées de façon indépendantes les unes des autres, et par là même hétérogènes. La première source d'hétérogénéité est liée au support de collecte des données, qui varie suivant les nomenclatures étudiées, et change au cours du temps.

5.1.1.1 Des supports multiples, multi-niveaux, non-alignés

Il s'agit de prendre en compte la multiplicité et l'hétérogénéité des supports spatiaux des statistiques territoriales. Nous limitons notre modélisation au support de type polygonal irrégulier, le plus fréquent dans le cas des données socio-économiques. Par exemple, le recueil de données statistiques au niveau européen est harmonisé dans la Nomenclature des Unités Territoriales Statistiques (NUTS), [Parlement européen 03]. Cette nomenclature reprend les découpages territoriaux historiques, qui sont de formes polygonales irrégulières et structure l'information en cinq niveaux, du plus fin au plus large, en terme notamment de seuils démographiques :

- LAU2 (au niveau des communes en France),
- LAU1 (au niveau des cantons en France),
- NUTS3 (au niveau des départements en France) - entre 150 000 et 800 000 d'habitants,
- NUTS2 (au niveau des régions en France) - entre 800 000 et 3 millions d'habitants,
- NUTS1 (au niveau des grandes régions en France) - entre 3 millions et 7 millions d'habitants,
- NUTS0 (au niveau des états),

Les seuils démographiques servent essentiellement à harmoniser les niveaux intermédiaires NUTS3, NUTS2 et NUTS1, mais par exemple, les états représentent une dérogation à ce principe puisque le Luxembourg est une unité de environ un demi-million d'habitants, mais placée au niveau NUTS0. L'emprise spatiale de la NUTS est celle de l'Europe constituée des 27 états membres de l'Union Européenne.

En réalité, nous visons au delà l'intégration de supports plus particuliers :

- les formes morphologiques des villes (Urban Morphological Zones, UMZ) ;
- les groupements de communes en structures de coopération intercommunale (EPCI à fiscalité propre et/ou territoire de Pays en France) ;
- l'ensemble des bassins versants d'un espace d'étude ;
- la nomenclature des États reconnus par l'ONU (les WUTS).
- etc.

Dans cette thèse, nous nous limitons à des objets vectoriels de forme polygonale, et nous ne prenons pas en compte les grilles régulières (formats raster). Par exemple, nous écartons la grille Corine Land Cover dont les pixels couvrent des zones d'un hectare, et sur lesquels sont identifiés le type majoritaire d'occupation du sol.

Ces supports de type polygones irréguliers sont définis comme des partitions spatiales de l'espace géographique qui forment alors un zonage du territoire. Un zonage est un découpage ou une partition de l'espace. Il peut servir à l'appropriation, la gestion, l'aménagement ou la connaissance de l'espace. Il existe, en effet, différents types de zonage, selon les acteurs et les fonctions : zonage politico-administratif, zonage statistique, zonage d'aménagement du territoire, etc. Un zonage peut être *couvrant* (partition totale de l'espace) par rapport à un espace d'étude déterminé, qu'on appelle l'*aire d'étude*, ou bien *non couvrant*. Dans le premier cas, on dira que le zonage est un *maillage* du territoire, [Grasland 98]. Les cas de zonages non couvrants sont fréquents, et sur ces zonages, des données socio-économiques sont aussi collectées, ou bien analysées. On peut citer par exemple, les bassins d'emplois en Europe, les formes morphologiques des villes (nomenclature UMZ), et les structures intercommunales. On peut noter que les données sont encore très rarement collectées directement au niveau des structures intercommunales, mais plutôt au niveau communal, et l'on doit agréger en fonction de la composition communale des structures intercommunales.

Très fréquemment, les nomenclatures qu'utilisent les producteurs de données définissent des niveaux hiérarchiques constitués de zonages qui s'emboîtent les uns dans les autres à partir d'un zonage élémen-

taire, le zonage le plus fin de l'espace dans la nomenclature considérée. Ainsi, les zonages forment les niveaux de cette hiérarchie, et l'emboîtement est défini par l'inclusion spatiale de toute unité de niveau inférieur dans une unité de niveau supérieure. Ces différents niveaux constituent une organisation hiérarchique du territoire dans laquelle chaque unité appartient à une ou plusieurs unités de niveau supérieur. Le niveau constitue aussi une échelle d'observation et d'analyse du territoire.

Les données associées peuvent être des variables qualitatives ou quantitatives. Les variables quantitatives sont des grandeurs numériques qui proviennent de comptages, de mesures ou de calculs effectués sur des comptages ou des mesures, comme par exemple le nombre d'habitants ou de ménages dans une unité territoriale. Parmi les variables quantitatives, les variables qui ne sont pas des ratios (pourcentage d'une variable quantitative par une autre), c'est-à-dire les variables *quantitatives absolues* (ou stocks), ont la propriété d'être additives : les valeurs de deux unités de même niveau peuvent être additionnées pour calculer la valeur associée à l'union des surfaces des deux unités.

Par rapport aux hiérarchies dites « agrégatives », un travail fondamental mené dans le cadre de la gestion de données multidimensionnelles complexes pour des entrepôts de données a donné lieu à des définitions plus précises de ces hiérarchies [Pedersen 01, Banerjee 09]. D'après ces travaux, une hiérarchie peut être *stricte*, *onto* ou *couvrante* ou au contraire *non-stricte*, *non-onto*, *non-couvrante* : « stricte » signifie que tout élément d'un niveau n'appartient qu'à un seul autre élément de niveau supérieur, « onto » correspond au fait que l'arbre représentant cette hiérarchie est équilibré et que tout élément d'un niveau non élémentaire possède un sous-élément de niveau directement inférieur, tandis que « couvrante » signifie que tout élément d'un niveau appartient à au moins un élément de niveau supérieur. Ainsi, une hiérarchie est dite stricte lorsque chaque unité possède au moins une et une seule entité supérieure, et onto lorsque toute unité de niveau non élémentaire englobe au moins une unité. Ce cas d'organisation arborescente des unités territoriales correspond à une réalité, celle par exemple de l'organisation des NUTS, figure 5.1, qui est issue de la volonté de produire des niveaux comparables en termes démographiques pour la collecte des données statistiques. Ce cas intéressant est étudié [Rigaux 95] par rapport à la dimension spatiale, car il permet l'agrégation ascendante depuis les feuilles vers la racine de l'arbre des données, et offre des possibilités de vérification avec les invariants suivants :

- (i) la géométrie d'une unité non élémentaire est l'union des géométries des unités qui la composent ;
- (ii) la valeur d'une variable quantitative absolue d'une unité non-élémentaire est constitué par l'agrégation des valeurs associées aux unités qui la composent.

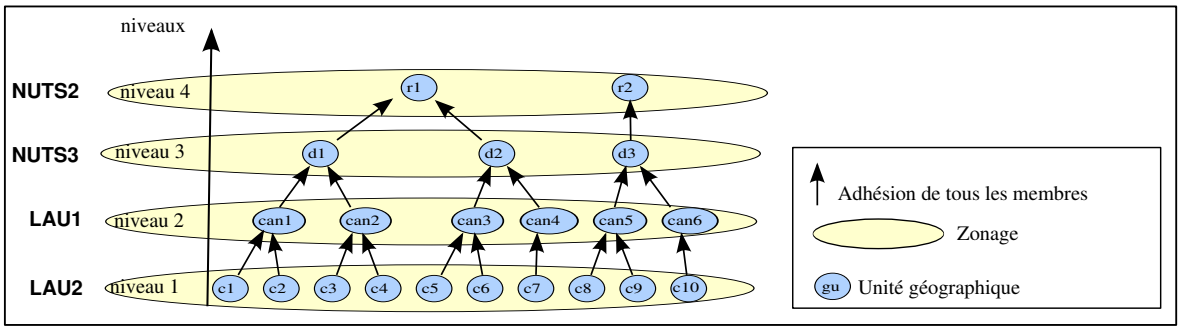


FIGURE 5.1 – Hiérarchie et niveaux dans la NUTS.

Dans le cadre de la constitution d'un réservoir de données issues de supports hétérogènes, la connaissance de ces relations d'agrégation est utile car les opérations d'agrégation vont permettre de transférer dans un zonage commun des indicateurs mesurés sur des niveaux de zonages différents. On peut éga-

lement s'appuyer sur des contraintes de désagrégation utiles : par exemple, un effectif compté sur un département se répartit sur les effectifs des communes composant ce département, et pas sur les communes du département voisin. Nous souhaitons donc nous ramener à la gestion de hiérarchies de zonage, pour lesquelles la disponibilité des variables et des géométries sur le maillage élémentaire permet de constituer, par agrégation, un ensemble d'informations multi-échelles.

Cependant, il apparaît que l'organisation hiérarchique stricte des unités territoriales n'est pas systématique. Les récentes recompositions territoriales locales en France en sont un bon exemple. Pour rationaliser la gestion territoriale, il s'agit de relancer l'intercommunalité et de construire des territoires plus vastes que la commune, d'une taille suffisante pour atteindre une masse critique en matière de population. De nouveaux Établissements Publics de Coopération Intercommunale (EPCI) sont créés, mais également ce qu'on appelle des « territoires de projet » (politique des Pays). Même si ce n'est pas explicite dans les lois, une forme de structure hiérarchique avec emboîtement spatial se dessine. Par exemple, en milieu rural, un premier niveau est constitué par les communautés de communes (un des nouveaux EPCI) ; un second par le Pays, fédérant plusieurs communautés de communes. Le Pays est le niveau de réflexion et de conceptualisation du projet de développement local ; les communautés de communes sont chargées des réalisations concrètes. Mais la création de ces zonages intercommunaux est progressive, puisque les différentes structures ne sont pas obligatoires et reposent pour partie sur l'initiative des acteurs locaux. En conséquence, il existe un certain nombre de cas ne respectant pas un emboîtement hiérarchique strict.

La figure 5.2 présente certains de ces cas, et illustre comment ce que nous avons défini comme «nomenclature des intercommunalités», constituée de trois niveaux (les communes, les EPCI et les Pays), forme une hiérarchie non stricte et non onto :

- 1. Une unité d'un certain niveau peut avoir deux unités de niveau directement supérieur. Sur l'exemple, l'EPCI e_4 est partagé entre le Pays p_3 et le Pays p_4. Dans ce cas, e_4 n'adhère pas à la fois à p_3 et p_4 ; ce sont une partie des communes constituant e_4 qui adhèrent à p_3, et l'autre partie des communes qui adhèrent à p_4. Ces cas devraient être transitoires, les EPCI devant normalement respecter, à terme, les limites des Pays, et ces derniers étant définis, en général, par leurs EPCI adhérents, et non par les communes adhérentes.
- 2. On observe des appartenances multiples à plusieurs niveaux de zonages. Sur l'exemple, la commune c_7 adhère à un EPCI e_4 mais aussi à un pays p_3 (cas possible si e_4 n'adhère pas dans son ensemble à p_3).
- 3. Une unité d'un niveau peut n'avoir aucune unité supérieure dans la nomenclature considérée. Sur l'exemple, la commune c_4 n'a pas d'unité supérieure.
- 4. Les sauts de niveau sont fréquents. Sur l'exemple, l'unité c_6 n'a pas d'unité supérieure dans les EPCI, mais appartient directement à un pays p_2. Une autre variante est la hiérarchie non complète avec manque du niveau supérieur : cas de c_5, qui appartient à e_3, mais il n'y a pas de Pays au niveau supérieur.

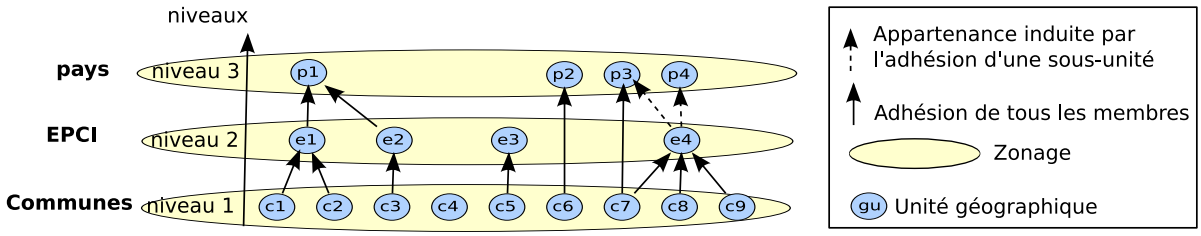


FIGURE 5.2 – Nomenclature des intercommunalités sur une région française.

Le modèle que nous proposons vise à produire de l'information agrégée sur des niveaux non élémentaires. Cela reste possible dans ces cas, à condition de définir plus précisément les relations d'agrégation dans la nomenclature. La relation d'agrégation doit lier les niveaux concernés, chaque zonage appartenant à une certaine nomenclature, et cette relation d'agrégation pourra être typée. Le typage le plus simple est l'agrégation totale : les unités considérées appartiennent entièrement à leur unité supérieure. Dans ce cas, les prédicats (i) et (ii) énoncés plus haut (page 143) s'appliquent directement pour l'évaluation des unités supérieures. Sinon, on est dans le cas d'une situation d'agrégation partielle. C'est le cas de l'EPCI e_4 qui appartient aux pays p_3 et p_4. Mais, dans ce cas, la géométrie des unités de niveau supérieure ne peut être constituée par agrégation des unités lui appartenant partiellement. Il faut donc chercher dans un zonage de niveau inférieur une relation d'agrégation totale pour constituer l'unité. Sur l'exemple, les communes c_8 et c_9 forment par agrégation totale le pays p_4.

5.1.1.2 Un support qui change

La comparaison et l'analyse des données suivant la dimension temporelle est fortement gênée par le fait qu'une même nomenclature change au cours du temps, aussi bien sur le plan de la forme des unités territoriales au niveau le plus élémentaire, que sur le plan de l'organisation (ce qui affecte les niveaux non-élémentaires). Les effets des évolutions sont visibles au niveau de chaque unité géographique.

Ce problème, identifié comme le *"split tract problem"* [Howenstine 93] a donné lieu à de nombreux travaux de recherche. De précédents travaux [Cheylan 97] soulignent que l'attribution d'une identité à une unité géographique peut permettre de retracer son évolution spatio-temporelle puisque tous ses attributs sont susceptibles de varier de façon indépendante : son empreinte spatiale, son nom, son centre, son code, son statut, son appartenance, etc. Certains auteurs proposent une ontologie du changement [Hornsby 98], basée sur la reconnaissance d'une identité pour les unités géographiques, mais celle-ci n'a pas d'implémentation immédiate parce que l'identité d'une unité géographique est subjective. En effet, cette approche reste purement abstraite car la méthode d'identification (ou reconnaissance d'une même entité à travers le temps) n'est pas abordée sur le plan opératoire. Il s'agit de déterminer un critère mesurable pour la continuation de l'identité d'une entité entre deux versions, que ce soit en comparant le nom, l'empreinte spatiale, le code ou son centre. Or, chacun de ces attributs peut changer entre deux versions, sans que l'identité de l'entité qu'ils décrivent ne soit altérée. Comme nous l'avons mis en évidence dans l'état de l'art, il apparaît qu'en réalité les modèles objets basés sur le paradigme identitaire sont très peu répandus car ils se heurtent à l'épineuse question de l'identité. Nous proposons de résoudre cette question en deux temps. Dans un premier temps, le modèle proposé est centré sur une unité géographique dont nous décrivons tous les attributs. Dans un second temps, nous proposons une méthode d'identification des unités à partir de ces attributs, et nous proposons un algorithme pour la

maintenance de ce modèle. Le modèle que nous proposons est qualifié d'objet, comme défendu par Worboys [Worboys 98]. Ce modèle permet de définir un ensemble d'unités géographiques existant dans une ou plusieurs nomenclatures, ensemble que nous nommons *le système territorial*, qui évolue au fil des versions de nomenclatures publiées. En effet, les versions de nomenclature décrivent les évolutions du système. Dans la section consacrée à la mise à jour et à la maintenance de ce modèle, nous décrivons une procédure d'identification semi-automatique des unités géographiques dans le système territorial, à partir de la comparaison de deux versions de nomenclature.

Par ailleurs, nous proposons d'indexer ce modèle par les événements du changement, ceux étant à l'origine (la cause) des modifications sur les attributs des unités géographiques. En effet, les changements ne surviennent pas sans raison. Ces raisons sont à rechercher du côté des motivations politiques, stratégiques et sociales. Cette hypothèse est soutenue par l'étude de M. Ben Rebah menée sur les recompositions territoriales des maillages administratifs en Tunisie dans le cadre de ses travaux de thèse [Ben Rebah 08]. Les stratégies à l'oeuvre sont souvent purement calculatoires comme dans le cas du *"Gerrymandering"* ou "charcutage électoral". Un exemple très connu a été décrit par Morgan [Morgan 03] : il explique que le pays de Galles, initialement divisé en deux régions nord et sud avec des valeurs moyennes de PIB par habitant par rapport aux autres régions européennes, a été plus tard remembré en deux régions est et ouest, avec une région Ouest ayant un PIB inférieur à 75% de la moyenne européenne. Ce remembrement avait pour principal objectif de faire en sorte que la partie ouest devienne éligible pour les fonds structurels européens, en accord avec l'objectif numéro un de la politique européenne de cohésion territoriale. Il ressort ici, comme dans de nombreux exemples, que de la taille et la forme des parcelles territoriales peut dépendre le nombre d'habitants recensés, les richesses décomptées ou le type majoritaire d'usage du sol enregistré. C'est pourquoi elles sont sujettes à modification, dans la cadre de stratégies politiques et économiques. Il apparaît donc fondamental que les changements dans les découpages ou l'organisation puissent être indexés par les événements décrivant les causes du changement. Dans ce cas, le modèle cesse d'être uniquement descriptif et devient alors explicatif de la dynamique du territoire. Cette approche est actuellement fortement défendue au sein de la communauté de géomatique, comme l'expliquent différents travaux [Langran 92], [Peuquet 02], [Claramunt 95], [Wachowicz 99] qui donnent lieu à des propositions de modèle orienté-objet indexé par des événements comme celui de Worboys, [Worboys 05] ou de Wachowicz, [Wachowicz 99].

De fait, décrire les causes d'un changement reste malaisé, car ces causes sont particulièrement abstraites et sujettes à débat, donc difficiles à modéliser. Sans aller jusqu'à expliquer les causes d'un changement sur le plan politique ou stratégique, nous souhaitons cependant donner aux experts les moyens de reconnaître quelles sont les unités impliquées dans un même changement, et quelle est la nature de ce changement. Ceci doit faciliter la lecture du phénomène géographique, et permettre de passer d'une lecture classique de l'espace (c'est-à-dire le découpage territorial observé suivant une succession de dates) à une lecture par type de changement : quels sont-ils ? où se produisent-ils majoritairement ? Ainsi, nous entendons faciliter l'analyse des changements survenus sur un lieu, une région en particulier puisque le modèle pourra être interrogé pour une certaine unité géographique, sur laquelle sera posée la question de la succession des changements qu'elle a éventuellement subis. Le changement se produit à un certain moment et peut impliquer une ou plusieurs unités : nous le désignons par le terme « événement ». Un événement est défini par une date, un état territorial initial et un état territorial final. Il faut noter que les événements sont datés approximativement, puisque le changement est détecté suite à l'insertion dans le système d'une nouvelle nomenclature. Cette date doit pouvoir être modifiée par un expert s'il dispose de plus d'informations sur cet événement. De plus, on imagine que des événements peuvent composer une suite d'événements qui s'inscrit dans un événement de temps long, s'étalant sur une période. Nous suggérons donc d'employer le terme d'*événement de généalogie* (à l'origine d'autres événements) pour

désigner ces événements de temps long, décrits par une période et non pas simplement une date. Un événement de généalogie sera donc constitué d'événements particuliers, qui sont eux datés.

Parmi ces événements particuliers, nous distinguons ceux qui ne s'appliquent qu'à l'unité géographique, de façon individuelle au niveau de l'identité géographique, de ceux qui associent plusieurs unités dans une transformation qui sera au minimum spatiale, c'est-à-dire lorsqu'il y a un échange de territoires entre plusieurs unités. La première catégorie d'événement est appelée « *événement de vie* » ou *LifeEvent* en anglais, alors que les seconds sont dits « *événement territoriaux* » ou *TerritorialEvent* en anglais. Ainsi, nous postulons que les événements territoriaux sont essentiellement définis par les formes et les surfaces de territoires mises en jeu, indépendamment de l'identité des unités géographiques. Ces événements peuvent avoir un impact sur les unités géographiques, du point de vue de leur identité. Ils ont donc des événements de vie comme conséquence.

Les événements de vie peuvent être :
- une apparition, qui correspond à la création dans le système territorial d'une nouvelle unité avec une identité propre ;
- une transformation, c'est-à-dire un changement de l'un des attributs de l'unité géographique, mais sans altération de son identité ;
- une disparition, qui correspond à la date de fin de vie d'une unité géographique, et à la fin de sa mise à jour, puisqu'elle n'est plus censée être modifiée dans les versions suivantes du système territorial.

Un événement de vie peut se produire de façon totalement indépendante des autres unités : par exemple, un changement de code ou de capitale pour un pays n'affecte pas les autres pays. Mais très souvent, les événements de vie sont provoqués par des interactions avec d'autres unités géographiques dans le cadre d'un événement territorial.

Nous avons besoin de décrire précisément ce que sont les événements territoriaux. C'est pourquoi nous introduisons une nomenclature du changement, adaptée des travaux de Theriault [Thériault 99] et Spery [Spery 01] [1]. La figure 5.3 présente cette nomenclature qui comporte deux niveaux : un premier niveau de classification des événements s'attache à déterminer simplement le nombre d'empreintes spatiales impliquées dans l'événement, avant et après, et s'inspire de la classification de [Thériault 99], tandis que le second niveau prend en compte la continuité de l'identité des unités géographiques participant à l'événement. Au premier niveau, il existe trois principaux événements territoriaux : la fusion (*Merge*), la division (*Split*) ou bien le remembrement (*Redistribution*). La fusion se distingue par le fait qu'à l'issue de l'événement, il ne reste qu'une unité géographique dans le système territorial, et inversement, la division ne met en jeu qu'une unité géographique avant l'événement, alors que le remembrement implique plus d'une unité géographique, avant et après l'événement. Au second niveau, la distinction entre les sous-évènements de fusion, que sont la fusion (*Fusion*) ou bien l'intégration (*Integration*), ou les sous-évènements de division, que sont la scission (*Scission*) ou bien l'extraction (*Extraction*), ou les sous-événements de remembrement, que sont la réallocation (*Reallocation*) ou bien la rectification (*Rectification*), porte sur le prolongement de l'identité des unités impliquées dans l'événement.

Il faut donc ici analyser les événements de vie associés à un événement territorial pour décider du type de spécialisation de l'événement de niveau 1. Par exemple, dans le cas d'un remembrement, il est spécialisé en rectification, si aucune des unités impliquées ne disparaît. C'est ce qui peut arriver lorsque deux voisins décident de déplacer une clôture commune : chacun garde sa propriété, qui continue

1. Dans cette adaptation, le domaine public a disparu, donc l'expropriation n'a plus lieu d'être. Nous avons aussi changé la signification initiale de intégration et extraction par rapport aux travaux de Spery.

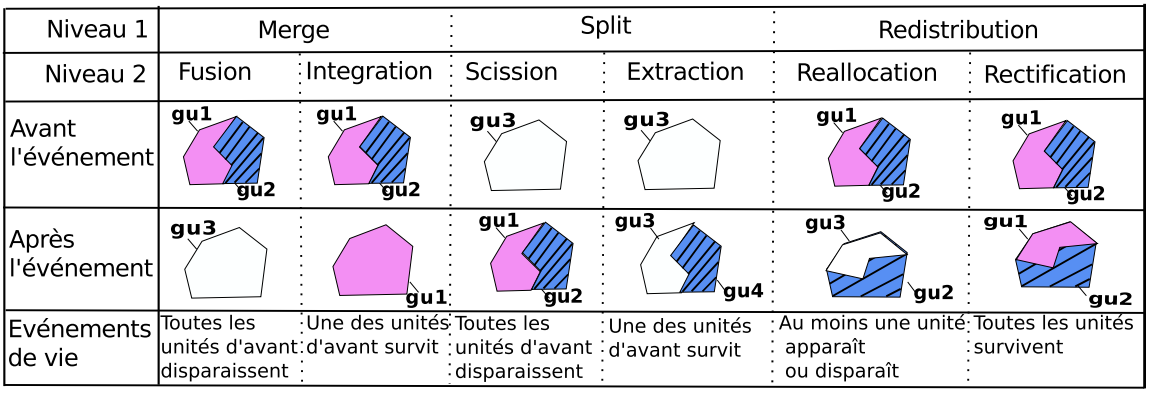


FIGURE 5.3 – Nomenclature des événements territoriaux.

d'exister, mais les empreintes spatiales sont rectifiées. Prenons comme autre exemple le cas d'une ville en pleine expansion, comme Bucarest, capitale de la Roumanie. Depuis des années, elle s'étend sur les communes alentours qu'elle absorbe et inclut dans ses limites : les communes disparaissent et perdent leur identité, du moins au niveau du découpage territorial officiel, tandis que Bucarest garde son identité et ne fait que s'agrandir. Dans ce cas, l'événement de fusion est donc spécialisé en intégration. Enfin, dernier exemple, à l'issue de la chute du Rideau de Fer en 1990, la république de Tchécoslovaquie s'est scindée en deux parties, deux États indépendants, représentant les deux identités principales qui la constituait : à l'Ouest le peuple Tchèque pour former la république Tchèque, et à l'Est le peuple slovaque pour former la république Slovaque. Cet exemple illustre un cas de changement complet d'identité, avec disparition l'unité géographique à l'origine de cet événement, qui est donc qualifié de scission au niveau 2.

Nous lions donc les changements territoriaux sur chacune des unités à des évènements qui relient et donnent un sens à ces changements. Ainsi, chaque changement est indexé par un élément qui permet de le dater et de l'identifier, mais aussi de le documenter, puisque nous pouvons alors associer un document administratif à l'évènement (des métadonnées).

Nous présentons maintenant comment notre modèle concrétise les idées proposées et répond à nos besoins.

5.1.2 Description du modèle

Le système d'information spatio-temporelle que nous proposons repose sur un modèle objet du système territorial (défini page 146), qui propose de *gérer simultanément plusieurs nomenclatures*, et qui intègre l'*unité géographique* comme objet central d'étude à l'intérieur de chaque nomenclature. En effet, chaque nomenclature définit un niveau de découpage élémentaire du territoire par la donnée d'un ensemble d'unités géographiques élémentaires. Chaque unité géographique apparaît, évolue, et disparaît, selon un critère d'identité indépendant de la nomenclature étudiée. L'historique d'une unité est enregistré, et sa durée de vie est estampillée avec un intervalle temporel (*validityPeriod*). Nous lions l'unité géographique aux différents évènements de recomposition territoriale qui peuvent engendrer son apparition, son évolution, ou sa disparition. Le modèle comprend deux parties qui peuvent être considérées indépendamment l'une de l'autre :

– la *partie identitaire* fournit un ensemble générique d'attributs qui peuvent décrire une unité géographique, pour chaque type de nomenclature.
– la *partie événementielle* décrit l'ensemble des processus de transformations survenus dans une nomenclature, et leur impact sur la généalogie et les attributs des unités.

Le modèle est décrit par un diagramme de classes dans le formalisme UML [Booch 99] dans lequel nous n'avons pas employé les notations d'agrégation ou de composition forte, afin d'éviter toute ambigüité avec les relations de composition ou d'agrégation explicitées comme des classes d'association entre les unités géographiques. De même, des contraintes spatio-temporelles entre les entités pourraient être formulées, en utilisant l'OCL comme le propose [Grumbach 01]. Afin de ne pas surcharger le modèle, les annotations OCL ne sont pas incluses. Les multiplicités indiquées sont celles définies pour un instant d'observation quelconque.

5.1.2.1 Description d'une unité géographique

Une unité géographique est complètement décrite à un instant t, à l'intérieur d'une nomenclature, par un ensemble d'attributs et de relations avec d'autres unités, comme le montre la figure 5.4. Tous les attributs ou les entités que nous proposons dans cette partie sont susceptibles d'évoluer indépendamment les uns des autres. Pour simplifier l'explication de cette partie du modèle, nous ne présentons pas la partie thématique du modèle, qui est développée dans le chapitre 6 consacré à la description des valeurs statistiques et de leur hétérogénéité, page 215. L'intervalle de validité de chaque objet du modèle représente la durée pendant laquelle il reste inchangé. La borne inférieure de cet intervalle peut être soit la date d'apparition de l'unité géographique à laquelle il est associé, soit la date de publication d'une nomenclature ayant modifié sa valeur. Quant à la borne supérieure de cet intervalle, elle peut être soit la date de disparition de l'unité géographique à laquelle il est associé, soit la date de publication d'une nomenclature ayant modifié sa valeur, soit la valeur spéciale « NOW » qui signifie que cette valeur reste valable jusqu'à la prochaine version de nomenclature, [Clifford 97].

L'unité géographique (*GeographicUnit* ou gu) possède une période d'existence (*validityInterval*) qui peut couvrir plusieurs versions de chaque nomenclature dans laquelle elle est référencée. En effet, chaque unité géographique s'inscrit dans au moins une organisation territoriale définie par une nomenclature (*Nomenclature*) composée de un à plusieurs zonages (*Zoning*). La nomenclature porte un nom (*nom*), un code (*code*), et possède une période de validité (*validityInterval*). La nomenclature est définie et révisée par un producteur (*Provider*) qui rend compte des changements territoriaux qui se sont produits en émettant une nouvelle version de nomenclature, version sur laquelle il s'appuie pour publier ses statistiques. Le modèle garde trace de la provenance de chaque version de nomenclature en décrivant le producteur par son nom (*name*) et son identifiant (*URI*), ainsi que la source (*Source*) (c'est-à-dire le support physique) qui a servi à publier la version de nomenclature à une certaine date, (*extractionDate*) en donnant le nom (*name*) et l'*URI* de cette source.

Pour chaque nomenclature, l'unité est associée à un ensemble d'attributs qui peuvent évoluer indépendamment de l'unité, et qui sont le code (*Code*), l'empreinte spatiale (*Footprint*), le centre (*Center*), et une désignation (*Designation*) dans cette nomenclature. Le code peut changer suivant les versions de nomenclature. Il est nécessaire de conserver ces codes, car les données statistiques sont fournies en lien avec une version de nomenclature dans laquelle l'unité géographique est identifiée par son code. À l'intérieur de chaque nomenclature, l'unité géographique peut posséder plusieurs désignations (*Designation*) en fonction des périodes de temps (*validityInterval*) et des langues (*lang*). L'unité géographique peut aussi posséder un (ou plusieurs) centre(s) (*Center*), définis chacun par une position dans l'espace et

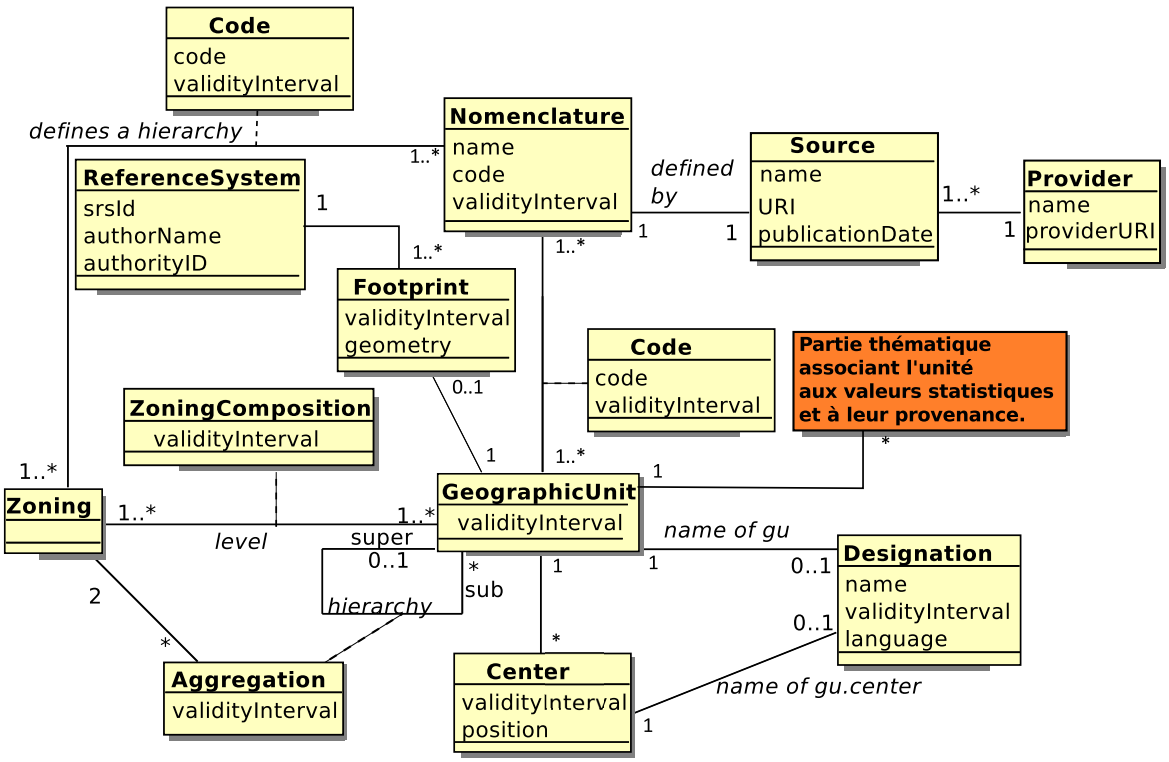


FIGURE 5.4 – Identification d'une unité géographique dans plusieurs nomenclatures.

un nom (*Designation*). Le nom du centre d'une unité n'est pas forcément le même que celui de l'unité elle-même. La multiplicité de l'association entre une unité et le(s) centre(s) est multiple, car dans la nomenclature des UMZ, certaines zones urbaines sont polycentriques, comme l'est, par exemple, l'aire urbaine formée par Bayonne-Biarritz-Anglet. La géométrie du centre dépend du type de nomenclature considéré. Ce sera, par exemple, le point le plus bas d'un bassin versant, un point représentant une capitale dans les WUTS, une adresse géographique pointant sur le siège d'un EPCI, ou bien une unité géographique de niveau inférieur dans le cas des départements.

Chaque version de nomenclature est associée à un certain fond de carte géographique, utilisant un unique système de projection (*ReferenceSystem*). Ce fond est constitué par le zonage (*Zoning*) de niveau élémentaire, et permet d'associer chaque unité à son empreinte spatiale (*Footprint*). Cette dernière peut évoluer au gré des évènements du changement, et est donc estampillée avec une période de validité (*validityInterval*). À chaque empreinte spatiale est associé le système de projection (*ReferenceSystem*) qui a été utilisé dans cette version de nomenclature, afin de pouvoir, si nécessaire, comparer des géométries issues de deux nomenclatures différentes lors d'un processus d'appariement des unités géographiques. Nous ne conservons qu'une seule empreinte spatiale par version de nomenclature, car notre modèle a pour objectif l'estimation, et non pas la représentation cartographique qui nécessite très souvent de stocker divers niveaux de généralisation de l'information géographique. Il faut noter que le producteur de données statistiques n'est pas tenu de publier les géométries des unités géographiques auxquelles il associe des informations statistiques, et il y a souvent un décalage entre la parution officielle des nouvelles frontières et celle de la publication de la Nomenclature correspondante. Le cas le plus problématique étant lorsque les géométries ne sont pas publiées du tout, et que seule une collection éparse de documents administratifs indiquent qu'il y a eu des changements de frontière, sans en donner le dessin exact. C'est la raison pour laquelle nous autorisons l'absence de géométrie.

Chaque nomenclature définit au moins un zonage (*Zoning*), couvrant ou non, pour lequel on précise la composition : c'est l'objectif de la classe d'association *ZoningComposition* entre une unité géographique (*GeographicUnit*) et un zonage (*Zoning*). L'appartenance à un zonage peut changer dans le temps si une unité change de statut : *ZoningComposition* précise par une période de validité la durée effective de la relation. Un code (*Code*) identifie le zonage dans une nomenclature, et ce code peut changer au cours du temps, il est donc estampillé par un intervalle de validité *validityInterval*. La nomenclature peut définir aussi plusieurs niveaux de zonages, strictement emboîtés ou non. Comme nous l'avons argumenté, la description des relations d'agrégation hiérarchique doit mettre en jeu les unités, mais aussi les zonages auxquels elles appartiennent, ainsi que les nomenclatures impliquées. Cette relation verticale est notée *Agregation* dans le modèle. Il s'agit d'une classe d'association entre deux unités, l'une (*super*) étant l'unité supérieure de l'autre (*sub*). Elle est datée par la période de validité (*validityInterval*) du rattachement, et elle est liée à deux niveaux de zonages différents (*Zoning*) : l'un pour le niveau inférieur de l'unité, et l'autre pour le niveau supérieur de l'unité, chaque zonage étant défini dans une version de nomenclature.

Pour illustrer l'utilisation de cette relation dans un schéma relationnel, on propose la relation suivante : `agregation (gu-inf-id, mesh-inf-id, gu-sup-id, mesh-sup-id, nomenclature-id, validityPeriod).`

Son instanciation est illustrée avec le cas de la commune c_7 représentée dans la figure 5.2 :
- *agregation (c7, communes, e4, epci, interco, [1998, now])* indique que la commune c7 fait partie de l'EPCI e4 dans la nomenclature des intercommunalités depuis 1998.
- *agregation (c7, communes, p3, pays, interco, [2000, now])* indique que la commune c7 adhère aussi au pays p3 dans la nomenclature des intercommunalités, depuis 2000.
- *agregation (c7, communes, dep38, département, NUTS, [1989, now])* indique que la commune c7 appartient au département dep38 dans la nomenclature des NUTS depuis 1989.

On remarque ainsi que certaines nomenclatures peuvent partager des niveaux géographiques communs : les intercommunalités et les NUTS partagent le même niveau communal. Définir ainsi la relation d'agrégation entre unités signifie en particulier qu'une unité ne peut pas être liée à deux unités appartenant à un même zonage. Par exemple, une commune ne peut appartenir à deux départements différents. En conséquence, chaque zonage de niveau non élémentaire doit être défini par l'ensemble des unités de niveau inférieur *adhérant de façon unique et complète* à ses unités. Par exemple, la figure 5.5 montre comment nous corrigeons le problème illustré dans la figure 5.2 : les unités du zonage « pays » devrait être composées par des unités du zonage « EPCI » si cette hiérarchie était stricte, onto et régulière. Mais comme l'unité e_4 n'adhère pas de façon complète au pays p_4 puisque qu'une de ses composantes, la commune c_7 adhère de façon indépendante au pays p_3, il faut utiliser le niveau des communes pour constituer par agrégation les pays p_3 et p_4. La relation d'agrégation autorise de définir les pays en conjuguant à la fois des adhésions d'unités du zonage « EPCI » et des adhésions d'unités du zonage « commune ». Ainsi, le modèle peut calculer de façon récursive les valeurs statistiques des unités géographiques de plus haut niveau par agrégation, si les statistiques sont additives, et aussi construire leurs empreintes spatiales par agrégation géométrique.

Enfin, ce modèle peut supporter plusieurs nomenclatures co-existantes comme illustré dans la figure suivante 5.6. Ce diagramme UML d'objets montre que l'unité géographique c_7 qui appartient au zonage des communes et qui est désignée par le nom « Balbigny » appartient à deux nomenclatures différentes :

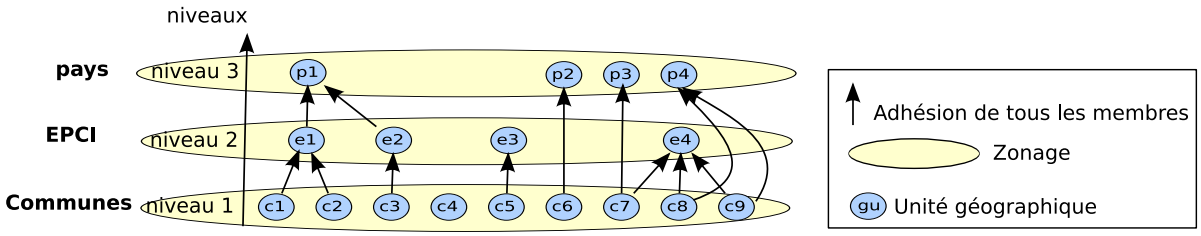


FIGURE 5.5 – Nomenclature des intercommunalités sur une région française, avec les relations de hiérarchie corrigées.

la NUTS et les « intercommunalités », qui se partagent un zonage commun de plus bas niveau : celui des communes. Ainsi cette unité possède au moins deux relations d'agrégation :

- une avec l'unité supérieure e_4 (dénommée « Territoire de Balbigny ») dans la nomenclature des « intercommunalités » au niveau EPCI.
- une avec l'unité supérieure can_1 (dénommée « Canton de Néronde ») dans la nomenclature des « NUTS » au niveau LAU1.

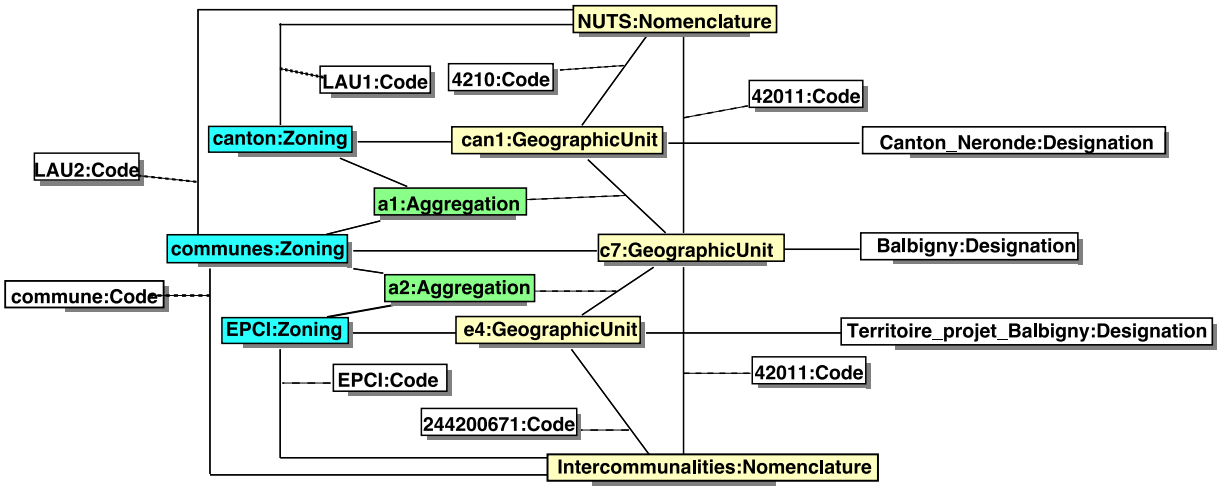


FIGURE 5.6 – Exemple de relations d'agrégation impliquant deux nomenclatures.

5.1.2.2 La généalogie et la transformation des unités géographiques

Notre modèle explicite aussi les transformations des unités causées par des événements de changement (voir le diagramme de la figure 5.7, qui est une extension du diagramme de la figure 5.4, à partir de l'objet *GeographicUnit*). Chaque événement possède un identifiant unique, et il est daté, soit par un intervalle s'il dure, soit par une date sinon. Il existe principalement deux types d'évènements :

- les événements de type « territorial » (*TerritorialEvent*) qui indexent les unités géographiques participant à un même événement. Une classe d'association (*Status*) précise le statut de chaque unité dans cet événement : le statut indique si l'unité est apparue, s'est transformée ou bien a disparu durant l'événement.
- les évènements de type « évolutif » (*LifeEvent*) ne sont associés qu'à une seule unité à la fois, et sont utilisés pour décrire l'évolution d'une unité. Cette évolution peut être liée à un événement

territorial, mais pas obligatoirement.
Dans le reste du document, nous dirons que les unités qui sont transformées ou qui disparaissent sont les prédécesseurs de l'événement, alors que les unités qui apparaissent ou se transforment suite à l'événement en sont les successeurs. Ce qui signifie qu'une unité, lorsqu'elle se transforme, est à la fois prédécesseur et successeur de l'évènement.

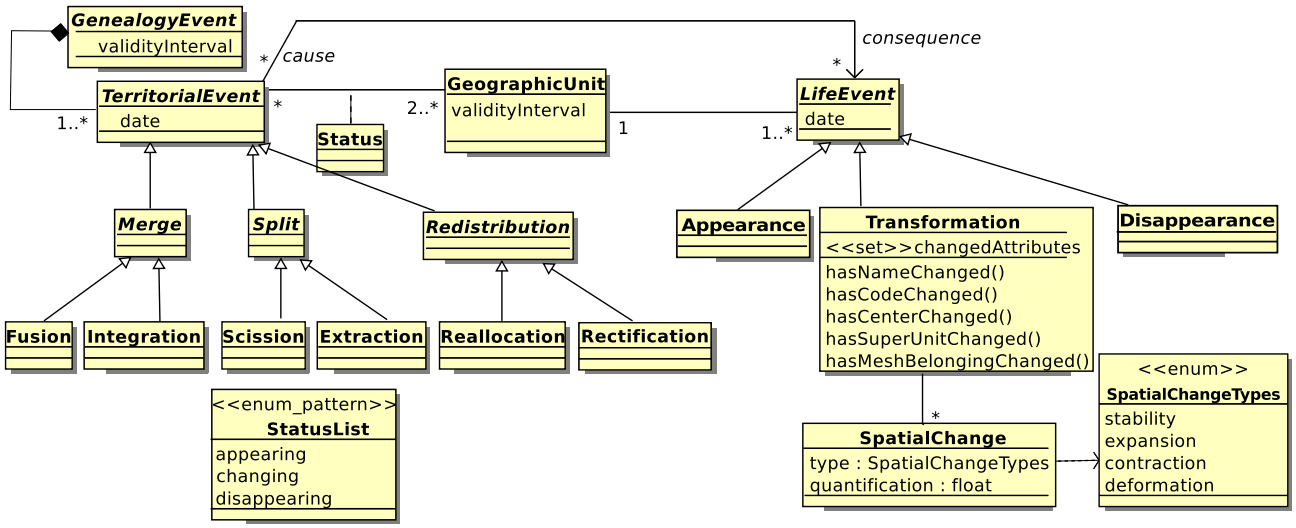


FIGURE 5.7 – Indexation des unités par les événements du changement : les causes et leurs conséquences.

Afin de modéliser plusieurs niveaux de granularité dans le déroulement des évènements, et la composition, nous introduisons l'évènement généalogique (*GenealogyEvent*) qui est composé d'évènements territoriaux, et qui a une durée non nulle. Par exemple, la restructuration d'une carte hospitalière peut procéder en plusieurs temps et lieux : le mois de mai par exemple verra la fermeture de l'hôpital de la Mûre et le rattachement du district hospitalier de la Mûre à Grenoble, et un autre mois des modifications impacteront des districts hospitaliers en Ardèche.

Un événement de type territorial permet de faire le lien entre l'état initial du système et l'état final du système. Les évènements de type territorial peuvent avoir des impacts sur l'évolution individuelle des unités impliquées dans l'événement : celles-ci peuvent apparaître (*Appearance*), ou bien se transformer sans perdre leur identité (*Transformation*) ou bien disparaître (*Disppearance*). La classe *Transformation* indique quels sont les attributs de l'unité géographique qui ont changé. En particulier, la classe *SpatialChange* précise le type de changement spatial qui a eu lieu, et dans quelle mesure (quantification donne le ratio de la surface finale sur la surface initiale). Ce ratio est celui calculé à partir des géométries associées aux unités. Si le niveau de généralisation entre les deux versions de géométrie est très différent, ce ratio risque d'être inutilisable. Cette valeur doit donc être utilisée avec précaution.

Nous avons expliqué page 148 comment les évènements territoriaux liaient les unités, et défini chacun de leur sous-type. Nous ajoutons ici un schéma 5.8 décrivant tous les cas de figure théoriques possibles. Ces six cas de figure illustrent comment nous lions les versions des unités d'une nomenclature par des évènements territoriaux. Les traits en pointillés expriment la continuation de l'identité (que nous concrétiserons par un appariement dans la suite de l'exposé) d'une unité entre deux versions successives.

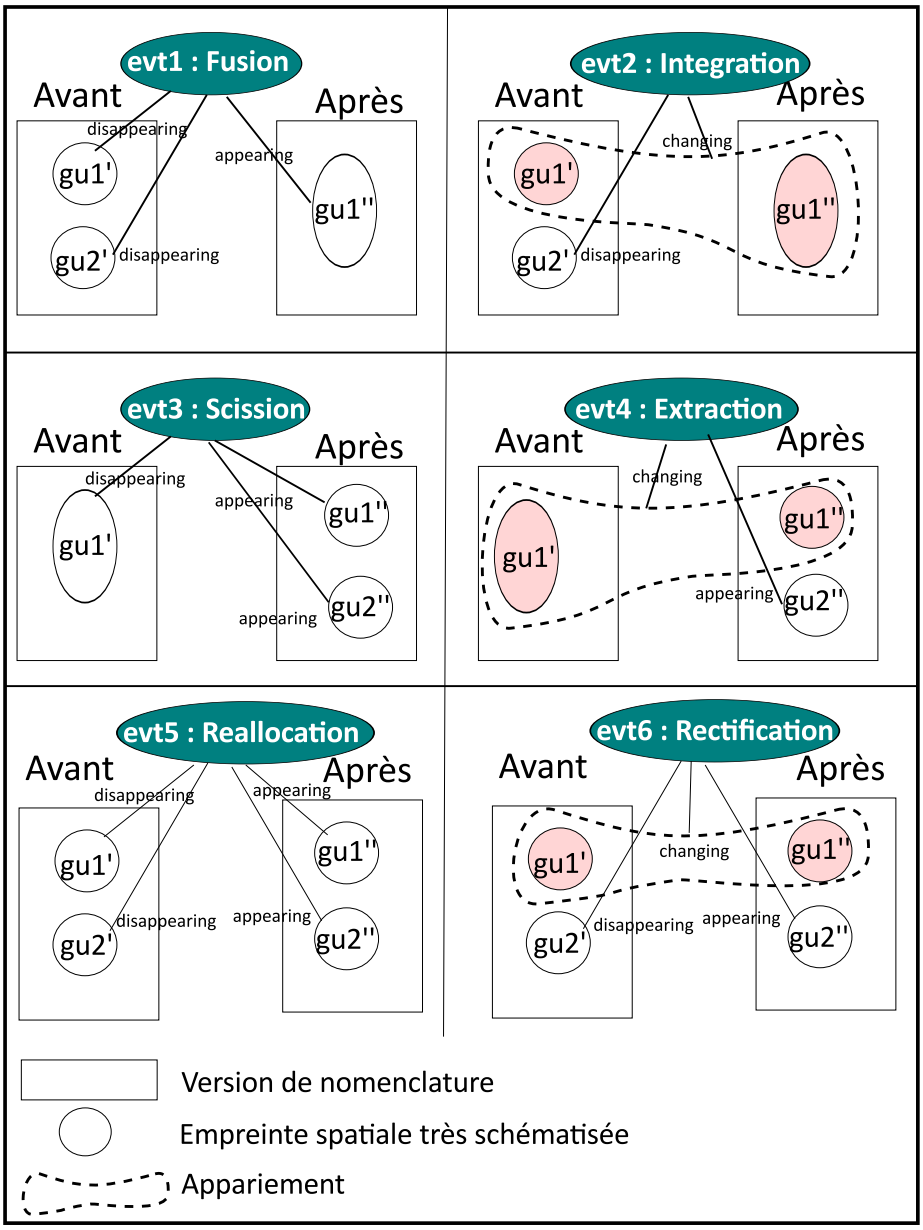


FIGURE 5.8 – Les six cas de figure possibles pour l'indexation des unités par des évènements territoriaux.

5.1.2.3 Exemples d'instanciation du modèle

Voici maintenant quelques exemples réels, de complexité croissante, montrant comment cette modélisation permet de rendre compte du changement.

Une unité peut apparaître, disparaître ou se transformer en dehors d'un événement territorial. Elle peut, par exemple, changer de nom, comme la commune de Malleval qui devient Malleval-en-Vercors en 2005. Cette *Transformation*, caractérisée par un changement de nom, est associée seulement à la commune, voir figure 5.9.

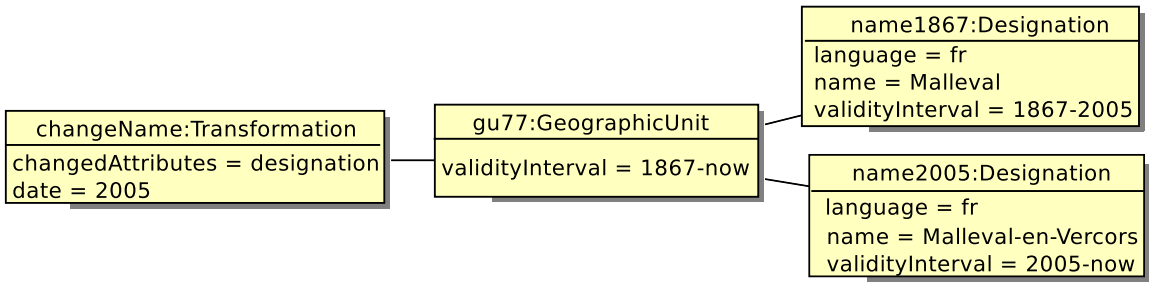


FIGURE 5.9 – Diagramme d'instances du changement de nom de Malleval (2005).

Le modèle permet d'enregistrer la modification des relations hiérarchiques à l'occasion d'une rectification. Prenons, par exemple, le changement d'appartenance de Saint-Priest, commune de l'Isère, qui passe en 1967 dans le département voisin du Rhône. C'est un événement de *Rectification* liant les trois unités, qui continuent leur existence, mais sont transformées (figure 5.10).

```
   StPriestMvnt : Rectification
Date : 1967
Changing : {Isère, Rhône, Saint-Priest}
```

FIGURE 5.10 – L'événement Rectification instancié pour le changement d'affectation de Saint-Priest (1967).

Une *Transformation* est définie pour Saint-Priest, sur son attribut d'appartenance hiérarchique. À chacun des départements de l'Isère et du Rhône, une *Transformation* de type spatiale est associée pour indiquer d'une part, la contraction de l'Isère, et, d'autre part, l'extension du Rhône (figure 5.11).

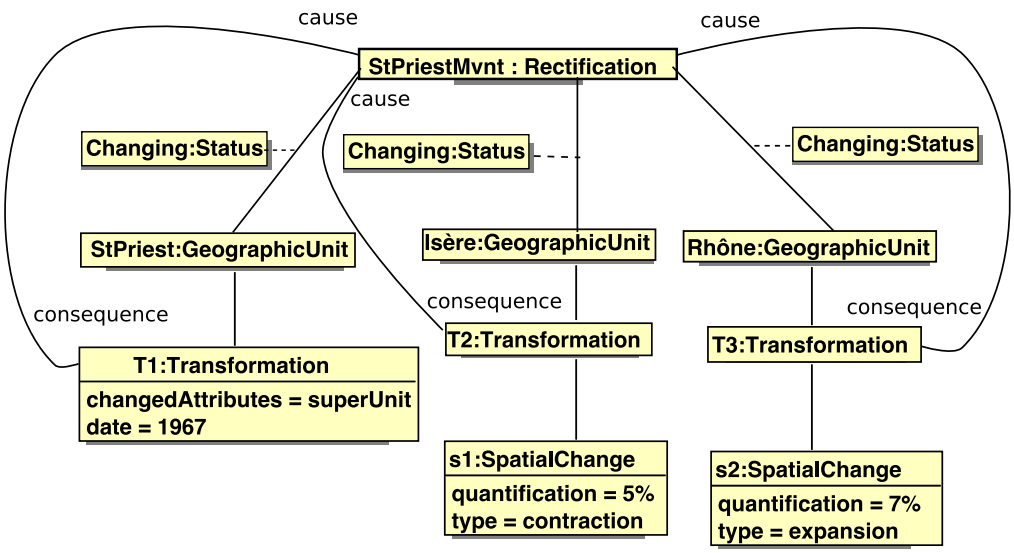


FIGURE 5.11 – Diagramme d'instances du changement d'appartenance de Saint-Priest (1967).

Le modèle permet également de décrire complètement les opérations plus complexes modifiant les limites du zonage, la plus complexe étant l'opération de réallocation. Ainsi, le remembrement de trois communes de l'Isère conduisant à la création de Chamrousse sert d'exemple pour montrer comment

l'opération de réallocation est décrite. L'opération, identifiée « ISE89 » et illustrée par la figure 5.12 montrant les deux versions de maillage communal successives, se produit en 1989 entre les trois communes iséroises de Vaulnaveys-le-Haut (gu_1), Séchilienne (gu_2), et Saint-Martin-d'Uriage (gu_3), et fait apparaître une nouvelle commune, nommée « Chamrousse » (gu_4).

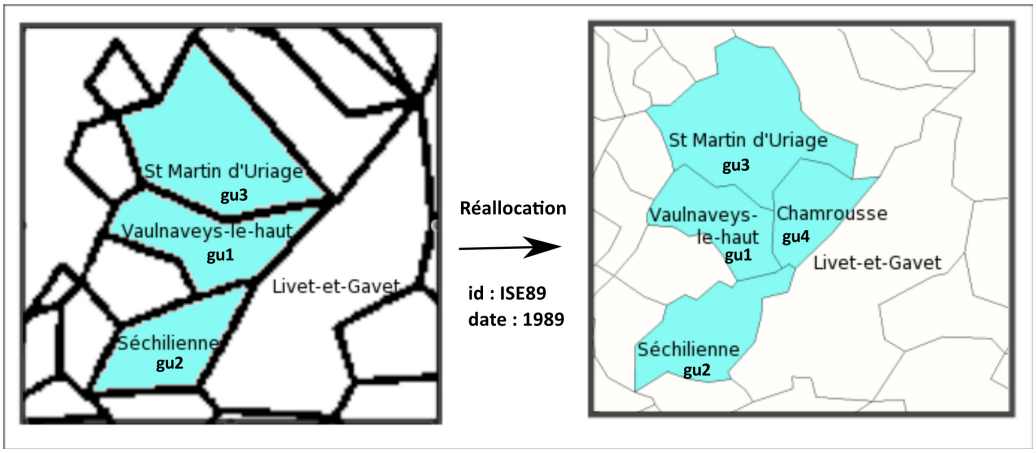


FIGURE 5.12 – Maillage communal de 1982 et celui de 1990 : création de Chamrousse.

La figure 5.13 montre l'instanciation de l'objet ISE89 correspondant.

```
   ISE89 : Reallocation
Date : 1989
Appearing : {Chamrousse}
Changing : {Vaulnaveys-le-Haut, Séchilienne,
Saint-Martin-d'Uriage}
```

FIGURE 5.13 – L'événement Reallocation instancié pour la création de la commune de Chamrousse (1989).

La commune de Chamrousse est liée à un événement de type « *Apparition* » pour cette date-là, tandis que les trois autres sont chacune liées à un événement de type « *Transformation* », où l'empreinte spatiale (*Footprint*) est indiquée comme étant l'attribut changé. Le changement est décrit par la classe *SpatialChange* qu'utilise l'événement de transformation pour décrire les conséquences de l'évènement. Le diagramme d'instances de la figure 5.14 montre comment les événements d'évolution T_1, T_2, T_3, qui sont trois instances de la classe *Transformation*, décrivent le changement induit par la réallocation ISE89. En particulier, elles indiquent que *Footprint* a changé, et utilisent une instance de *SpatialChange* caractérisant leur changement spatial.

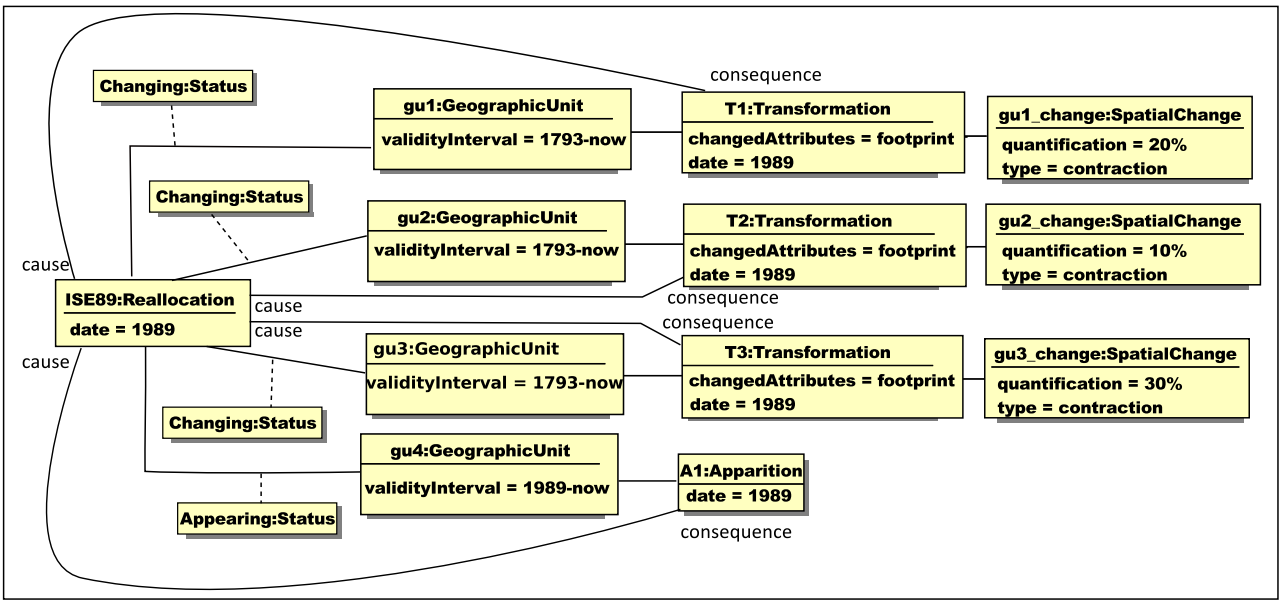


FIGURE 5.14 – Diagramme d'instances du changement pour l'apparition de Chamrousse en 1989.

À une autre échelle, la réunification de l'Allemagne en 1990 est un évènement de type *Fusion* qui fait disparaître la République Fédérale d'Allemagne (RFA) et la République Démocratique d'Allemagne (RDA), au profit d'une nouvelle unité, Allemagne, ces trois unités étant liées par l'événement dénommé *« Einigungsvertrag »* (figure 5.15). Les conséquences de cet événement *« Einigungsvertrag »* sont la *«Disparition»* de la RFA et la RDA et l'*«Apparition»* de l'Allemagne. Toutes les unités appartenant directement à la RFA à cette époque changent d'unité supérieure : elles sont aussi impliquées dans cet événement *« Einigungsvertrag »*, mais parce qu'elles changent d'unité supérieure. Ces unités de niveau 2 dans la NUTS sont désignées sous l'appellation *Länders* en allemand, et sont codées dans la NUTS comme suit : *"DE1", "DE2", "DE3", "DE5", "DE6", "DE7", "DE9", "DEA", "DEB", "DEC", "DEF"*. Leur *Transformation* porte sur la relation hiérarchique.

```
   Einigungsvertrag : Fusion
Date : 1990
Disappearing : {RFA, RDA}
Changing : {DE1, DE2, DE3, DE5, DE6, DE7, DE9, DEA, DEB, DEC,
DEF}
Appearing : {Allemagne}
```

FIGURE 5.15 – L'événement de Fusion instancié pour la réunification allemande (1990).

La carte de la figure 5.16 donne à voir le nouveau découpage de l'Allemagne réunifiée, où le territoire qui appartenait auparavant à la RFA apparaît en fond vert, et le territoire de la RDA en fond blanc.

FIGURE 5.16 – Carte des *Länders* dans l'Allemagne réunifiée (1991).

Par ailleurs, à cette occasion, l'ancienne RDA procède à un remembrement au niveau de ses unités de niveau 2 dans la NUTS, et donc cinq nouveaux *Länder* sont créés, pour remplacer les quinze anciens districts (*Bezirke* en allemand) qui existaient depuis 1952 en RDA. Cet événement qui a lieu simultanément avec la réunification de l'Allemagne correspond à la volonté de rétablir les anciennes limites des *Länders* qui existaient dans cette partie de l'Allemagne avant 1952. Lors de cet événement dénommé *« Wiederherstellung der Länder »* , (figure 5.17), les cinq nouveaux *Länders* dénommés Brandenbourg, Mecklembourg-Poméranie-Occidentale, Saxe, Saxe-Anhalt et Thüringe, désignées respectivement par leur code NUTS *"DE4", "DE8", "DED", "DEE"*, et *"DEG"* apparaissent, alors que disparaissent les quinze unités codifiées ainsi dans la NUTS : *"EN101, EN102, EN103, EN21, EN221, EN222, EN223, EN311, EN312, EN313, EN321, EN322, EN331, EN333"*. De plus, toutes les unités de niveau 3 dans la NUTS (les *Kreise* en allemand) et faisant alors partie de la RDA, changent d'unité supérieure. Par exemple, le *Kreise* Schwerin change d'unité supérieure : de l'unité codée "EN103", qui correspond au *Bezirk* de Schwerin, il passe sous l'unité supérieure codée *"DE8"*, qui correspond au *Land* Mecklemburg-Vorpommern.

```
   Wiederherstellung der Länder : Reallocation
Date : 1990
Disappearing : {EN101, EN102, EN103, EN21, EN221, EN222,
EN223, EN311, EN312, EN313, EN321, EN322, EN331, EN333}
Changing : {L'ensemble des unités de niveau 3 de l'ex-RDA}
Appearing : {DE4, DE8, DED, DEE, DEG}
```

FIGURE 5.17 – L'événement de réallocation « *Wiederherstellung der Länder* » modélisant le rétablissement des cinq anciens *Länders*.

La carte de la figure 5.18 montre une superposition entre l'ancien découpage en *Bezirke* de l'ex-RDA et le nouveau découpage en *Länders* de cette région de l'Allemagne. Sur cette carte, certaines parties ont été coloriées avec un motif en points pour montrer que les frontières des nouveaux *Länders* s'écartent de beaucoup des anciennes frontières des *Bezirke*. La carte inclut aussi dans la région du Mecklembourg-Poméranie-Occidentale le découpage de niveau NUTS3, afin de donner une idée de l'existence des sous-unités ayant changé d'unité supérieure lors de cet événement.

FIGURE 5.18 – Carte des anciens *Bezirke* et des nouveaux *Länder* dans l'ex-RDA (1991).

La première partie de ce chapitre est conclue par des exemples qui montrent que notre modèle peut refléter la réalité des territoires en perpétuelle mutation, et qu'il est capable de prendre en compte plusieurs hiérarchies évolutives simultanément. De plus, la force de l'indexation par les événements du changement a déjà aussi été démontrée dans de précédents travaux, et peut désormais donner lieu à de nouvelles propositions pour la visualisation et l'analyse du changement territorial.

Nous présentons maintenant comment la maintenance du modèle et l'identification des unités sont assurées. En effet, la plupart des modèles basés sur un paradigme identitaire rencontrent le problème de la mise à jour, et ceci freine leur diffusion et leur adoption à grande échelle. Nous proposons donc dans la deuxième partie de ce chapitre une solution à ce problème.

5.2 Mise à jour et maintenance du modèle

La mise à jour du modèle pour la partie spatiale est déclenchée par l'introduction d'une nouvelle version de nomenclature *V"*. Si la nomenclature est déjà présente pour la version notée *V'*, les unités présentes sont déjà enregistrées dans le modèle, avec un identifiant unique (*id*). À chaque unité *u'* est attaché sa liste d'identifiants dans la version considérée : noms dans différentes langues (*lang.name*), code (*code*), et géométries (*outline*) à différents niveaux de généralisation, centres (*center*) si disponibles, et si la nomenclature est constituée de plusieurs niveaux, le zonage d'appartenance de l'unité (*level*), (son niveau en l'occurrence), ainsi que l'identifiant de l'unité supérieure (*super*). L'unité se présente donc comme un objet ayant plusieurs attributs, auxquels on accède directement par la notation « . » : *u'.code* donne le code de l'unité *u'*.

Bien souvent, la nomenclature (qui contient le code des unités au minimum) est publiée séparément du fond géométrique associé, car ce ne sont pas les mêmes institutions qui travaillent sur ces données. Par exemple, en Europe, les Etats s'accordent avec Eurostat [2] pour publier une nomenclature des NUTS lorsqu'il faut valider et entériner un certain nombre de changements territoriaux qui sont intervenus depuis la publication de la version de nomenclature précédente. Mais c'est l'association Eurogeographics [3] qui est en charge de la publication des frontières dans un fond cartographique, à différents niveaux de généralisation. Nous supposons ici qu'à la fois la nomenclature et le fond géométrique associé sont disponibles, c'est-à-dire que le délai existant entre les deux publications s'est écoulé.

La mise à jour du modèle comprend deux phases : dans la première phase, le système doit apparier les unités de chacune des deux versions de nomenclature, c'est-à-dire mettre en relation les unités deux à deux lorsque c'est possible, découvrir les unités qui ont disparues, apparues ou qui se sont transformées. Il infère donc les évènements de vie, de type *LifeEvent*. Il infère également les évènements territoriaux de premier niveau, comme de second niveau, de type *TerritorialEvent*. Cette première phase se décompose en deux actions distinctes.

La première action consiste à identifier les événements territoriaux [4] qui ont eu lieu, c'est-à-dire à reconnaître les unités qui :

- ont fusionné, donc sont liées à un même évènement de type *Merge*,
- se sont divisées, donc sont liées à un même évènement de type *Split*,
- se sont redistribuées, donc sont liées à un même évènement de type *Redistribution*.

La seconde action permet d'abord d'apparier les unités, en comparant chacun de leurs attributs, et donc de reconnaître les événements de vie (apparition, transformation, disparition). Puis, en recoupant les informations sur l'appariement des unités avec la connaissance du premier niveau d'évènement territorial, le sous-type de l'événement territorial est inféré. En effet, par définition, un rattachement (*Integration*) d'une unité implique qu'au moins l'une des unités impliquées pendant la fusion se soit transformée. De même, une prise d'indépendance (*Extraction*) suppose aussi qu'au moins une des unités se soit transformée pendant la division. Et enfin, une rectification de frontières (*Rectification*) suppose que toutes les unités impliquées aient conservé leur identité en se transformant seulement.

Dans la seconde phase de la mise à jour, nous proposons qu'un expert utilise les résultats du programme pour corriger s'il le souhaite les résultats de l'appariement. Il peut en particulier modifier le statut de toute unité et modifier le type de l'évènement territorial. En effet, si l'expert change le statut

2. `http://epp.eurostat.ec.europa.eu/`
3. `http://www.eurogeographics.org/`
4. Les termes définis dans la classification page 147 sont conservés, en anglais.

d'une unité impliquée dans un évènement (par exemple, s'il la déclare transformée et non pas disparue), ceci a un impact direct sur le sous-type de l'évènement territorial auquel elle est liée. Mais également dans ce cas, et si l'algorithme d'appariement prend en compte l'unité supérieure (l'attribut *super*), l'appariement pour toutes les sous-unités de l'unité modifiée est automatiquement recalculé, comme les sous-types d'événements territoriaux. L'algorithme d'appariement que nous proposons est configuré par l'expert. Celui-ci peut demander sa ré-exécution à volonté, suivant les paramètres d'appariement qu'il définit.

5.2.1 Appariement automatique de deux versions de nomenclature

L'objectif de la première phase d'une mise à jour du modèle est d'apparier les unités d'une version avec celles de l'autre version de nomenclature. Nous proposons donc ici un algorithme d'appariement. Cet algorithme s'apparente aux algorithmes basés sur l'usage d'une fonction de croyance, comme proposé par [Raimond 07], dans le cadre de fusion de données hétérogènes non hiérarchiques. Nous sommes dans un cas relativement simple, où bien que les données puissent être incertaines et comporter des omissions ou des erreurs, leur schéma est établi et correspond au modèle proposé. Il faut définir quels sont les attributs qui sont utiles pour l'identification de l'unité géographique, et dans quelle proportion, car la croyance est relative à ce qui constitue l'identité de l'unité géographique, suivant l'expert et la nomenclature considérée. Notre algorithme cherche les unités qui sont égales pour la fonction de masse définie par l'expert, c'est-à-dire une combinaison de r critères portant sur les attributs de l'unité existants dans la nomenclature traitée, et définis comme pertinents pour la comparaison. La fonction de masse F renvoie une valeur comprise entre 0 et 1, en combinant des valeurs booléennes valuées à 0 (pour faux) et 1 (pour vrai) pondérées par des coefficients α_k attachés à chaque attribut constituant un critère C_k de comparaison. La somme des r coefficients α_k vaut 1. Ainsi, si la comparaison d'une unité u' avec une unité u'' renvoie 1, c'est que tous les attributs sont identiques deux à deux. De façon générale, la fonction de masse peut s'écrire comme une fonction qui associe à tout couple d'unités (u', u'') une valeur d'appariement p comprise entre 0 et 1 comme décrit dans l'équation 5.1.

$$\begin{aligned} &\text{Soit } F_{(\alpha_1,C_1),(\alpha_2,C_2),\ldots,(\alpha_r,C_r)} \text{ avec } \sum_r \alpha_k = 1 \\ F_{(\alpha_1,C_1),(\alpha_2,C_2),\ldots,(\alpha_r,C_r)} &: (V' \times V'') \to \mathbb{R}, \\ &(v', v'') \mapsto p,\ p \in [0,1] \end{aligned} \tag{5.1}$$

Lorsque l'algorithme d'appariement ($lifeEventResearch$) ne trouve pas d'informations pour un certain critère C_k considéré, il est ignoré dans le calcul de la fonction de masse. Par exemple, si l'une des unités du couple considéré n'a pas d'unité supérieure, ce critère est ignoré. Afin de vérifier l'invariant présenté dans l'équation 5.1, c'est-à-dire que la somme des coefficients α_k impliqués dans l'évaluation soit égale à 1, nous proposons soit de reporter le poids du coefficient du critère ignoré C_k de façon équitable entre les autres critères, ou bien d'évaluer arbitrairement le critère ignoré à 1. Il se peut aussi que bien que la fonction de masse ait renvoyée 0, c'est-à-dire que tous les attributs soient strictement différents deux à deux, mais que l'expert considère quand même que l'unité est la même. L'algorithme $lifeEventResearch$ procède toujours du haut vers le bas, c'est-à-dire en appariant d'abord les unités de plus niveau (le niveau État, par exemple) avant de chercher à apparier les unités du niveau directement inférieur. Ceci possède l'avantage de comparer en premier un nombre moindre d'unités, et de repérer très vite les unités qui sont entrées ou sorties de l'espace d'étude que couvre la nomenclature. Par exemple, si la Roumanie et la Bulgarie entrent en Europe en 2007, l'algorithme peut rapidement pointer toutes unités et les sous-unités de ces pays comme nouvelles (donc apparues). Et réciproquement, si un ou

plusieurs pays disparaissent de l'espace d'étude, leurs sous-unités sont pointées comme ayant disparu. Cette optimisation est facultative. Elle ne devient nécessaire que si un des critères d'appariement porte sur la comparaison d'unités supérieures : dès lors, il est nécessaire d'avoir apparié en premier les unités supérieures pour faciliter ce test et éviter d'entrer dans des branches non explorées. Si la nomenclature ne possède qu'un unique niveau, l'unité d'appartenance n'est plus un critère d'appariement, et l'algorithme procède alors en une seule fois, sur l'unique niveau, mais il fonctionne de la même manière.

Nous proposons avec l'équation 5.2 un exemple de définition de la fonction de masse F, qui porte sur le code ($code$), la géometrie ($outline$), la désignation ($name$), et l'unité supérieure ($super$) pour chaque unité. Concernant la langue de la désignation, l'algorithme cherche une langue où le nom a été spécifiée dans les deux versions, et dès qu'il en trouve une, il s'appuie dessus. On peut en effet envisager que parfois les unités soient nommées dans la langue native du pays et en anglais, mais il n'existe pas systématiquement une traduction en langue anglaise.

$$\begin{aligned}\text{Soit } F_{(\alpha_1,\ code),(\alpha_2,\ outline),(\alpha_3,\ name),(\alpha_4,\ super)} \text{ avec } \alpha_1 = 1/3, \alpha_2 = 1/3, \alpha_3 = 1/6, \alpha_4 = 1/6 \\ F_{(\alpha_1,\ code),(\alpha_2,\ outline),(\alpha_3,\ name),(\alpha_4,\ super)} : (V' \times V'') \rightarrow \mathbb{R}, \\ (v', v'') \mapsto 1/3 * (1 - |(u'.code - u''.code)|) \\ + 1/3 * (1 - |(u'.outline - u''.outline)|) \\ + 1/6 * (1 - |(u'.name - u''.name)|) \\ + 1/6 * (1 - |(u'.super - u''.super)|)\end{aligned} \tag{5.2}$$

Dans cet exemple, la notation mathématique $|(u'.code - u''.code)|$ indique que nous testons l'égalité (plus précisément la distance) entre l'attribut « code » des unités u' et u''. S'ils sont égaux (ou de distance nulle), la valeur absolue de leur différence renvoie 0. De même, $|(u'.super - u''.super)|$ vérifie que les unités supérieures de u' et u" sont similaires lorsqu'elles existent. La comparaison de chaînes de caractères (code et nom) ne pose pas de problème particulier. Cette comparaison de chaînes de caractères peut reposer sur la distance de Levenshtein [Levenshtein 65] par exemple, ou d'autres, qui seront appropriées en fonction du contexte (usage d'abréviations ou de casses différentes). En revanche, la comparaison des géométries peut poser problème. Nous proposons donc d'abord un test pour l'égalité de deux géométries, qui sera utilisé par l'algorithme d'appariement $lifeEventResearch$, ainsi que par l'algorithme recherchant les événements territoriaux $territorialEventResearch$.

5.2.1.1 Comparaison de deux empreintes spatiales

Il est nécessaire de comparer des géométries fournies dans un système de représentation identique. Nous utilisons pour cela les informations sur le système de projection associées à chaque empreinte spatiale pour convertir les géométries dans le même système de représentation. Par ailleurs, pour améliorer la précision des comparaisons ou bien augmenter la vitesse des comparaisons spatiales, le niveau de généralisation à employer peut être configuré par l'utilisateur. Cependant, deux versions de géométrie attachées à un territoire, ayant dans les faits conservé la même empreinte spatiale, pourront ne pas être égales suivant une test simple d'égalité (suivant la spécification de l'OGC, avec la méthode *equals*). En effet, les géométries de deux versions de nomenclatures peuvent être fournies à des niveaux de généralisation différents, même une fois converties dans le même système de projection cartographique. La figure 5.12 page 156 illustre ce problème avec le cas de la création de la commune de Chamrousse par l'évènement de Réallocation impliquant trois communes (Saint Martin d'Hères, Vaulnaveys-le-haut, et Séchillienne) en 1989 : les limites des communes avant cet évènement qui ont été portées à notre connaissance datent

de 1982, et sont très généralisées, en comparaison de celles de 1990.

Ce problème est connu dans la littérature et a donné lieu à la définition de nouveaux tests d'égalité entre les géométries :

- Devogèle utilise la distance de Fréchet [Devogèle 02] pour vérifier le niveau de ressemblance de deux contours dans le cadre d'appariement de géométries pour l'intégration de base de données hétérogènes ;
- Clementini introduit une algèbre floue pour comparer deux empreintes [Clementini 01].

Cependant, ces méthodes ont l'inconvénient majeur d'être à la fois difficiles à implémenter et coûteuses en temps de calcul.

Nous proposons donc ici un test d'égalité, simplement basé sur le calcul du rapport entre la différence de surface entre deux polygones, et la surface de la partie qu'ils ont en commun. Ce test n'est valable que si les deux polygones ont une intersection non vide. Si le rapport multiplié par 100 est inférieur à une petite valeur, notée epsilon (ε), cela signifie que l'empreinte spatiale a très peu changé et donc elle est considérée comme inchangée. Ce test correspond à la *distance surfacique*, qui est réputée robuste, [Hangouët 05], et qui a l'avantage d'être paramétrable.

La valeur epsilon (ε) doit être calibrée en fonction de la différence entre les niveaux de généralisation de deux géométries, de telle sorte que le test d'égalité soit positif pour un territoire qui n'a pas changé d'empreinte. Par exemple, pour comparer la version de NUTS 2003 et 2006, une valeur de epsilon à 2% est correcte. Cette valeur d'epsilon peut également être utilisée pour paramétrer la force du test d'égalité : un epsilon de l'ordre de 1% peut être utilisé pour un test d'égalité forte, alors qu'un epsilon de l'ordre de 40% pour un test d'égalité faible. Dans le second cas, deux empreintes sont dites identiques si elles s'intersectent, et que leur part de surface commune représente 60% de la surface totale.

Le test d'égalité (noté $\approx$) sur les empreintes spatiales (qui sont des géométries de type polygones ou multi-polygones) est le suivant pour deux géométries $g1$ et $g2$ fournies dans le même système de projection :

$$\begin{aligned} g1 \approx g2 \Leftrightarrow & (g1 \cap g2 \neq \varnothing) \wedge e \leq \varepsilon, \\ & e = (area(g1 - g2))/area(g1 \cap g2) * 100) \end{aligned} \tag{5.3}$$

L'implémentation de ce test se réalise très facilement à partir des prédicats définis dans les spécifications de l'OGC (section OGC S2.1.13) : $intersection(g1, g2)$ calcule $(g1 \cap g2)$ et $symdifference(g1, g2)$ calcule $(g1 - g2)$. Pour l'implémentation de $(g1 - g2)$, il faut utiliser la fonction symdifference telle que décrite dans les spécifications OGC S2.1.13 [5].

La figure 5.19 illustre le calcul effectué pour deux polygones $g1$ et $g2$. Si le polygone $g1$ est très significativement différent de $g2$, alors le test indique une valeur de e très grande, car la partie que les deux polygones ont en commun est très petite au regard de la surface différente. Ce test peut être utilisé pour l'appariement des géométries qui est une des étapes nécessaires de l'algorithme d'appariement que nous développons pour deux versions de nomenclatures V' et V'', où V' précède la version V''. En effet, pour inférer des types d'évènements territoriaux s'étant produits (*Merge, Split, Redistribution*), il est nécessaire de vérifier les invariants géométriques décrits dans le tableau 5.1. Le premier invariant (1) signifie que p unités g' ont fusionné en une unité g''. Le second invariant (2) est symétrique du précédent :

5. Cette fonction est commentée plus complètement sur ce site :
`http://publib.boulder.ibm.com/infocenter/idshelp/v10/index.jsp?topic=/com.ibm.geod.doc/geod156.htm`.

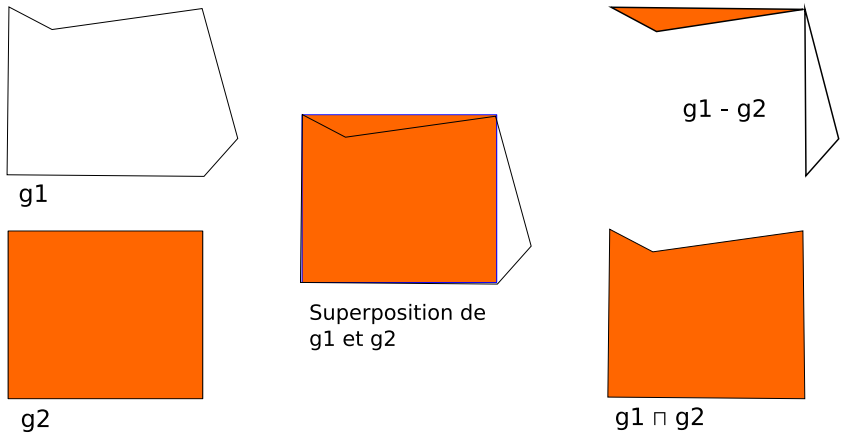


FIGURE 5.19 – Évaluation de l'égalité géométrique de deux polygones $g1$ et $g2$.

une unité g' s'est divisée en q unités g''. Lorsque $p = 1$ ou $q = 1$, l'unité g' s'apparie spatialement directement avec l'unité g''. Enfin, le troisième invariant (3) montre que p unités g' ont donné lieu à q unités g'' par une redistribution.

TABLE 5.1 – Invariants géométriques attachés aux différents types d'évènements territoriaux.

Type de l'évènement territorial	Invariant géométrique	
Fusion (*Merge*)	$\exists g''_j \in V'' / \forall g'_i \in V', i \in \{1..p\}, \bigcup_{i=1}^{p} g'_i \approx g''_j$	(1)
Scission (*Split*)	$\exists g'_i \in V' / \forall g''_j \in V'', j \in \{1..q\}, \bigcup_{j=1}^{q} g''_j \approx g'_i$	(2)
Redistribution (*Redistribution*)	$\forall g'_i \in V', i \in \{1..p\} \wedge \forall g''_j \in V'', j \in \{1..q\}, \bigcup_{i=1}^{p} g'_i \approx \bigcup_{j=1}^{q} g''_j$	(3)

5.2.1.2 Algorithme de détection des évènements territoriaux

Nous présentons en premier lieu un algorithme pour la détection des **événements territoriaux** ($TerritorialEvent$) s'étant produits entre deux versions de découpages territoriaux V' et V'', basé sur le test d'appariement géométrique précédemment décrit. Appelé $territorialEventResearch$, l'algorithme organise d'abord les unités de chaque version V' et V'' en deux listes distinctes :

- BEFORE contient la liste des unités u' de V', cette liste est de taille n
- AFTER contient la liste des unités u'' de V'', cette liste est de taille m

L'objectif est de constituer une matrice d'appariement spatial, nommée SPATIALMATCH, dont les entrées sont les index des unités de chaque version (les lignes correspondent aux unités u' de la première version V' de nomenclature, les colonnes correspondent aux unités u'' de la seconde version V'' de nomenclature). Les valeurs de la matrice d'appariement spatial sont des entiers, dont les valeurs correspondent aux codes détaillés et expliqués dans le tableau 5.2. : g'_i est la notation de la géométrie d'une unité u'_i de V', et g''_j est la notation de la géométrie d'une unité u''_j de V''.

TABLE 5.2 – Codes définis pour l'appariement spatial des unités.

Valeur du code	Nom du code	Invariant vérifié
-1	INTERSECTION	$g'_i \cap g''_j \neq \varnothing$
0	VIDE	$g'_i \cap g''_j = \varnothing$
1	EQUAL	$g'_i \approx g''_j$
2	MERGE	$g'_i \subseteq g''_j$
3	SPLIT	$g'_i \supseteq g''_j$
4	REDISTRIBUTION	$\exists X / \bigcup_{p \neq i} g'_p \approx X \ \wedge \ \exists Y / \bigcup_{q \neq j} g''_q \approx Y$ $\wedge \ g'_i \cup X \approx g''_j \cup Y$
5	TRANSFORMATION	$g'_i \cap g''_j \neq \varnothing \wedge$ $\nexists X, \nexists Y / (\bigcup_{p \neq i} g'_p \approx X \ \wedge \ \bigcup_{q \neq j} g''_q \approx Y \wedge \ g'_i \cup X \approx g''_j \cup Y)$

Si, pour toutes les unités u''_j de V'', SPATIALMATCH [x][j] vaut VIDE, alors l'unité u'_x a disparu. Réciproquement, si pour toutes les unités u'_i de V', SPATIALMATCH [i][y] vaut VIDE, alors l'unité u''_y est nouvelle. Cette matrice permet d'accéder rapidement aux unités impliquées dans un même évènement :

- si SPATIALMATCH [x][y] vaut MERGE, alors toutes les unités u'_i telles que SPATIALMATCH [i][y] vaut MERGE sont aussi impliquées dans une fusion territoriale, avec u''_y comme unité résultant de cette fusion.
- si SPATIALMATCH [x][y] vaut SPLIT, alors toutes les unités u''_j telles que SPATIALMATCH [x][j] vaut SPLIT sont aussi impliquées dans une division territoriale, avec u'_x comme unité source de la division.
- si SPATIALMATCH [x][y] vaut REDISTRIBUTION, alors toutes les unités u''_j telles que SPATIALMATCH [x][j] vaut REDISTRIBUTION sont le fruit d'un même remembrement territorial, alors que toutes les unités u'_i telles que SPATIALMATCH [i][y] vaut REDISTRIBUTION sont à la source du remembrement.
- si SPATIALMATCH [i][j] vaut EQUAL, alors les unités g'_i et g''_j ont une géométrie identique, inchangée.
- enfin, il peut exister des cas de transformation territoriale, mettant en jeu une à plusieurs unités u'_i de la version V', avec des unités u''_j de la version V'', mais dont l'union des géométries g'_i n'est pas égale à l'union des géométries g''_j. Ce cas se produit lorsque l'espace d'étude couvert par la seconde nomenclature V'' se réduit ou s'étend sur la mer par exemple, ou bien sur des territoires non inclus dans l'espace d'étude couvert par la nomenclature V''. Ces cas sont repérés par le code TRANSFORMATION. Par exemple, une inondation côtière durable constituerait une amputation du territoire d'une unité, sans impliquer de transformations avec d'autres unités.

Nous illustrons par un exemple comment la matrice est construite et devrait être remplie. L'exemple est une région du Danemark, comparée entre la version 2003 et la version 2006, dont la figure 5.20 présente les découpages successifs.

Pour ces deux découpages, la liste des unités pour chaque version est ainsi constituée :

- BEFORE $\leftarrow$ [DK001, DK002, DK003, DK004, DK005, DK006], de taille $n = 6$
- AFTER $\leftarrow$ [DK011, DK012, DK013, DK021, DK022], de taille $m = 5$

La matrice SPATIALMATCH est de taille $n \times m$, donc 6×5.

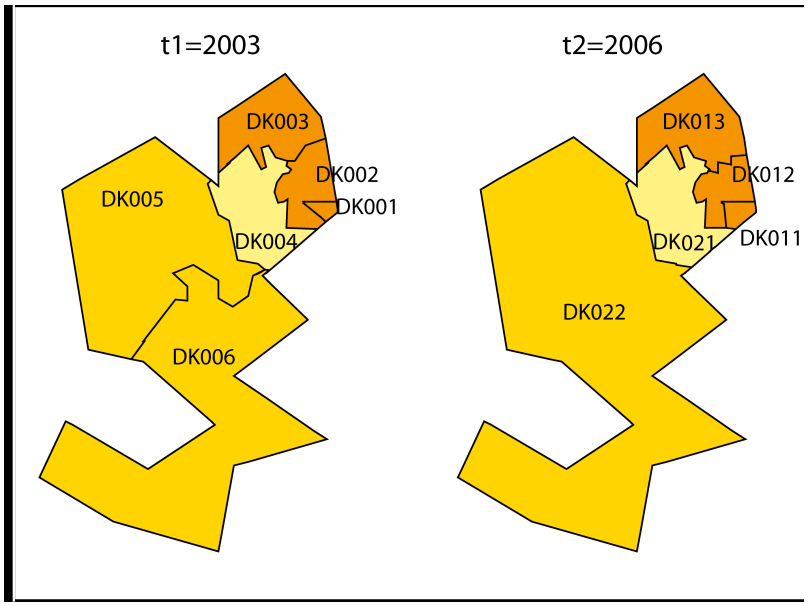


FIGURE 5.20 – Découpages du Danemark entre les versions de NUTS 2003 et 2006.

TABLE 5.3 – Instanciation de SPATIALMATCH pour le cas du Danemark.

code	id	DK011	DK012	DK013	DK021	DK022
		u''_1	u''_2	u''_3	u''_4	u''_5
DK001	u'_1	4	4	4	0	0
DK002	u'_2	4	4	4	0	0
DK003	u'_3	4	4	4	0	0
DK004	u'_4	0	0	0	1	0
DK005	u'_5	0	0	0	0	2
DK006	u'_6	0	0	0	0	2

Nous présentons ensuite comment remplir cette matrice en deux parcours croisés successifs des listes BEFORE et AFTER par l'algorithme $territorialEventResearch$, qui réalise une suite de tests d'intersection et d'égalité entre les unités de chaque liste. Le premier parcours compare chaque unité de V' à V'' en testant d'abord l'intersection (ligne 5 de l'encadré 5.4), puis, si celle-ci existe, l'égalité (ligne 7 de l'encadré 5.4). Ce parcours est donc en coût $O(n \times m)$ pour le nombre de comparaisons spatiales. Le second parcours recherche les unités impliquées dans un évènement territorial, quelque soit son type (MERGE, SPLIT ou REDISTRIBUTION), et son coût est de $O(n^2 \times m^2)$. Au départ, la matrice SPATIALMATCH est initialisée avec toutes ses valeurs à VIDE. Lorsqu'une intersection est trouvée entre une unité u'_i de V' et une unité u''_j de V'', sans que l'égalité des géométries soit directement vérifiée, l'algorithme affecte la valeur INTERSECTION à l'entrée SPATIALMATCH[i][j] (ligne 6 de l'encadré 5.4). Lorsque l'algorithme $territorialEventResearch$ s'achève, tous les cas d'intersection ont été identifiés, et il ne doit rester aucune entrée ayant le code INTERSECTION dans la matrice. Les tests d'intersection sont tous réalisés lors du premier parcours et sauvegardés dans SPATIALMATCH, et les parcours suivants économisent des tests d'intersection en consultant la matrice SPATIALMATCH.

Nous présentons les opérations effectuées lors des deux parcours, à l'aide d'un pseudo code qui réutilise les notations précédemment introduites. Les variables A, B sont de type géométrie et $Found$ est un booléen. L'opération d'affectation est représentée par "$\leftarrow$". Par exemple, «$B \leftarrow g''j$» signifie que l'on affecte la valeur de la géométrie $g''j$ à la variable B.

Premier parcours

Le premier parcours (voir encadré 5.4) compare chaque unité de V' à V'' en testant d'abord l'intersection puis si celle-ci existe l'égalité.

```
1   Pour p allant de 1 à n
2       A ← g'_p
3       Pour q allant de 1 à m
4           B ← g''_q
5           Si (A ∩ B ≠ ∅)
6               SPATIALMATCH[p][q] ← INTERSECTION
7               Si (A ≈ B)
8                   SPATIALMATCH[p][q] ← EQUAL
9               Finsi
10          Finsi
11      Finpour
12  Finpour
```
(5.4)

Second parcours

Le second parcours (voir encadrés 5.5, 5.6 et 5.7) est basé sur une observation simple : toute redistribution territoriale, implique au moins une unité de chaque version, et parmi ces deux unités, l'une au minimum possède une intersection non vide avec une troisième qui fait partie de la redistribution. Ainsi, il est possible de reconstituer l'ensemble des unités impliquées dans une même redistribution territoriale (notée redistributionK) en constituant deux listes sans doublons : la liste BEFOREKLISTE des unités précédant l'évènement redistributionK, et la liste AFTERKLISTE des unités succédant à l'évènement redistributionK. Chacune des listes peut être constituée progressivement à partir d'un parcours de la liste des unités de la première version V', puis de la seconde version V'', puis à nouveau de la première version V', en ajoutant à chaque tour les unités qui s'intersectent dans leur listes respectives AFTERKLISTE et BEFOREKLISTE, et ainsi de suite jusqu'à ce que les deux listes restent identiques entre deux itérations. Lorsque la liste BEFOREKLISTE est de taille un, il s'agit d'une division (SPLIT) ; lorsque la liste AFTERKLISTE est de taille un, il s'agit d'une fusion (MERGE). Sinon, il s'agit d'une redistribution (REDISTRIBUTION), sauf si la contrainte d'égalité entre les unions de géométries précédant et succèdant à l'évènement n'est pas respectée (c'est alors une TRANSFORMATION).

La figure 5.21 illustre le fonctionnement de l'algorithme $territorialEventResearch$ sur un espace d'étude où se sont produites une fusion, une division, trois redistributions et une transformation. On remarque que l'algorithme s'arrête dès qu'une des deux listes n'est plus modifiée.

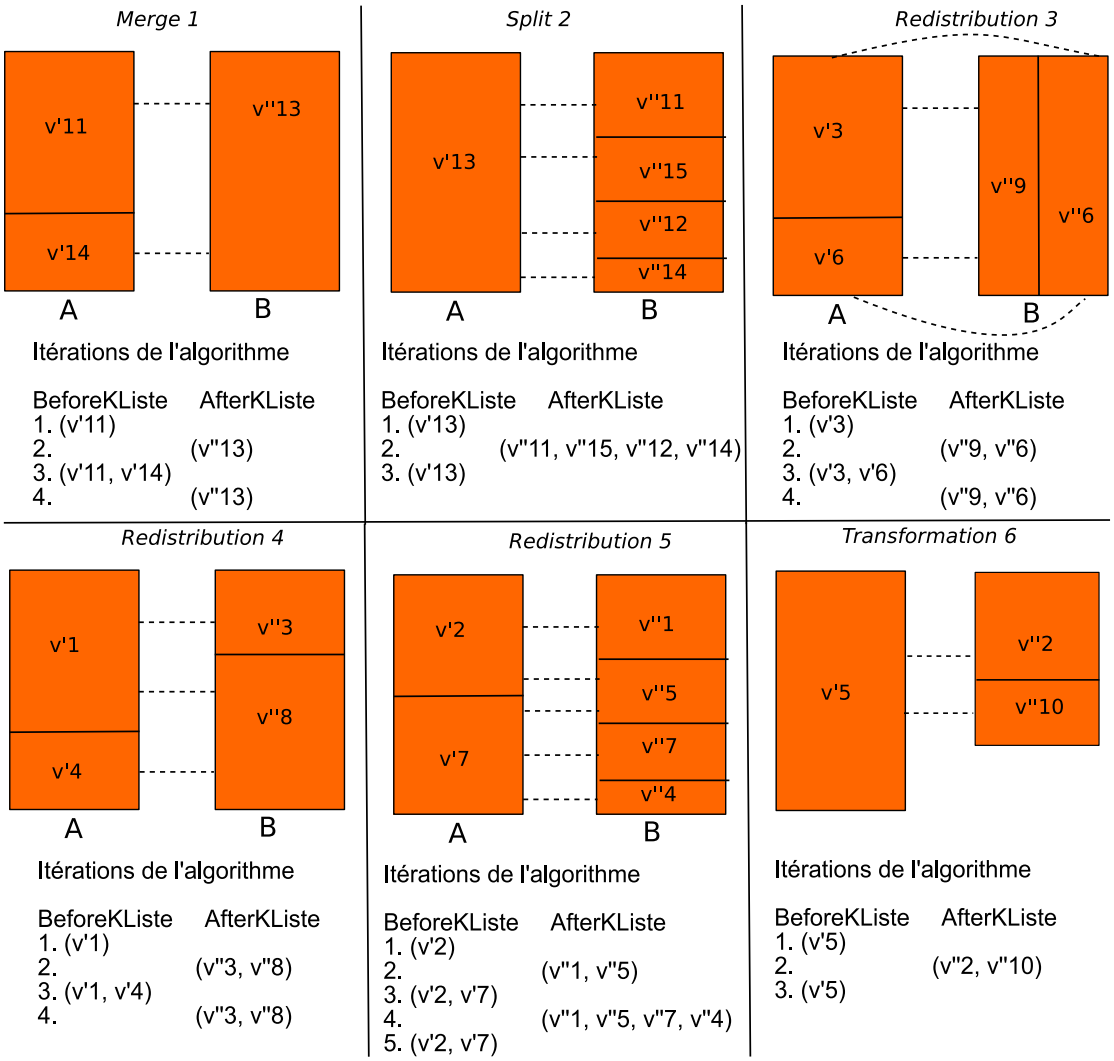


FIGURE 5.21 – Exemples de changements territoriaux, entre les unités v' dont l'union forme A et les unités v'' dont l'union forme B.

Nous donnons une formalisation de la seconde partie de l'algorithme $territorialEventResearch$, (voir encadrés 5.5, 5.6 et 5.7), basé sur l'usage de deux tableaux de booléens, BEFOREK, AFTERK, de tailles respectives n et m :

- BEFOREK[i] vaut vrai si u'_i est impliquée dans une redistribution K,
- AFTERK[j] vaut vrai si u''_j est impliquée dans la même redistribution K.

Ces tableaux remplacent l'usage des listes BEFOREKLISTE et AFTERKLISTE décrit précédemment. Également, le compteur *nbNewInK* est utilisé pour dénombrer le nombre de nouvelles unités ajoutées dans les listes à chaque itération. Lorsque ce compteur vaut zéro - ligne 28 de l'encadré 5.6 -, l'algorithme s'arrête et traite le cas trouvé (soit une redistribution, soit une transformation territoriale) - encadré 5.7 -, puis réinitialise les tableaux BEFOREK et AFTERK pour effectuer une nouvelle recherche de redistribution - lignes 3 à 8 de l'encadré 5.5. De même, deux compteurs *nbAfter* et *nbBefore* sont incrémentés chaque fois qu'un nouvel élément est ajouté au tableau AFTERK - ligne 13 de l'encadré 5.6 -, ou au tableau BEFOREK - ligne 24 de l'encadré 5.6.

L'algorithme $territorialEventResearch$ initialise le tableau BEFOREK avec la première unité $v'i$ rencontrée dans la version V' ayant une intersection non vide avec une unité $v''j$ de la seconde version, et qui n'est impliquée dans aucune fusion ou bien division, ni n'est déjà été appareillée à une autre unité : ce critère correspond à SPATIALMATCH[i][j] qui vaut INTERSECTION - lignes 12 à 17 de l'encadré 5.5. À chaque itération, pour un des tableaux donnés, BEFOREK (ou AFTERK), l'algorithme sélectionne toutes les unités qui possèdent une intersection non nulle avec une unité de la seconde version V'' - lignes 8 et 9 de l'encadré 5.6 - (ou respectivement de la première version V' - lignes 19 et 20 de l'encadré 5.6), et qui ne sont impliquées dans aucune fusion ou bien division, ni ne sont déjà été appareillées à une autre unité. Il les ajoute ensuite, si elles ne sont pas déjà présentes, dans le tableau AFTERK - lignes 10 et 12 de l'encadré 5.6 - (ou respectivement BEFOREK - lignes 21 et 23 de l'encadré 5.6).

Lorsque que les deux tableaux AFTERK et BEFOREK restent stables entre deux itérations pour le nombre de valeurs à vrai - lignes 28, 29 et 30 de l'encadré 5.6, l'algorithme cesse d'itérer et teste si l'union géométrique des unités à vrai du premier tableau BEFOREK est égale à l'union géométrique des unités à vrai du second tableau AFTERK - ligne 33 de l'encadré 5.7. Si ce test est positif, alors une redistribution a été trouvée, et l'algorithme $territorialEventResearch$ parcourt la matrice SPATIALMATCH pour mettre à jour la valeur des unités u'_i ou u''_j appartenant à cette redistribution (SPATIALMATCH[i][j] vaut alors REDISTRIBUTION, MERGE ou SPLIT, suivant la taille des tableaux BEFOREK et AFTERK). Puis les tableaux BEFOREK et AFTERK sont réinitialisés à faux (ce qui correspond au vidage d'une liste). Si le test est négatif - ligne 49 de l'encadré 5.7, c'est un cas de transformation, où par exemple u'_i intersecte u''_j, mais u'_i s'est étendue ou s'est réduite en empiétant sur un territoire en dehors de l'espace d'étude (la mer par exemple). Dans ce cas, l'algorithme met à jour l'entrée SPATIALMATCH[i][j] correspondante et lui affecte la valeur de TRANSFORMATION, et il réinitialise également les tableaux BEFOREK et AFTERK à faux pour toutes les unités.

L'algorithme $territorialEventResearch$ recommence alors le second parcours - lignes 25 et 2 de l'encadré 5.5, afin de trouver une unité u'_i de la première version V' ayant une intersection non vide avec une unité u''_j de la seconde version V'' (telle que SPATIALMATCH[i][j] vaille INTERSECTION), puis retrouve toutes les unités incluses dans cette nouvelle redistribution. Si l'algorithme n'en trouve pas, c'est que tous les cas possibles ont été élucidés : appariement, fusion, division, redistribution, transformation. L'algorithme $territorialEventResearch$ de détection de changements territoriaux se termine ici.

```
1   i ← 1
2   Tant que (i ≤ n)
3       Pour p allant de 1 à n
4           BEFOREK[p] ← false
5       Finpour
6       Pour q allant de 1 à m
7           AFTERK[q] ← false
8       Finpour
9       j ← 1
10      Found ← false
11      Tant que(j ≤ m et !Found)
12          Si(SPATIALMATCH[i][j] = INTERSECTION)
13              A ← g'_i
14              B ← g''_j
15              BEFOREK[i] ← true
16              AFTERK[j] ← true
17              Found ← true
18          Sinon
19              j ← j + 1
20          Finsi
21      Fintantque
22      Si (Found)
23          (suite du traitement dans l'encadré 5.6 page 172)
24      Finsi
25      i ← i + 1
26  Fintantque
```

(5.5)

```
1   nbBefore ← 1
2   nbAfter ← 1
3   redistributionComplete ← false
4   Tant que (!redistributionComplete)
5       nbNewInK ← 0
6       Pour p allant de 1 à n
7           Pour q allant de 1à m
8               Si(SPATIALMATCH[p][q] = INTERSECTION
9                       et BEFOREK[p] = true et AFTERK[q] = false)
10                  AFTERK[q] ← true
11                  nbNewInK ← 1 + nbNewInK
12                  B ← B ∪ g''q
13                  nbAfter ← 1 + nbAfter
14              Finsi
15          Finpour
16      Finpour
17      Pour p allant de 1 à n
18          Pour q allant de 1 à m
19              Si(SPATIALMATCH[p][q] = INTERSECTION
20                      et AFTERK[q] = true et BEFOREK[p] = false)
21                  BEFOREK[p] ← true
22                  nbNewInK ← 1 + nbNewInK
23                  A ← A ∪ g'p
24                  nbBefore ← 1 + nbBefore
25              Finsi
26          Finpour
27      Finpour
28      Si (nbNewInK = 0)
29          redistributionComplete ← true
30      Finsi
31  Fintantque
32  (suite du traitement dans l'encadré 5.7 page 173)
```

(5.6)

```
33  Si(A ≈ B)
34      Pour p allant de 1 à n
35          Pour q allant de 1 à m
36              Si(BEFOREK[p] et afterK[q])
37                  Si(nbBefore = 1)
38                      spatialMatch[p][q] = SPLIT
39                  Sinon
40                      Si(nbAfter = 1)
41                          spatialMatch[p][q] = MERGE
42                      Sinon
43                          spatialMatch[p][q] = REDISTRIBUTION
44                      Finsi
45                  Finsi
46              Finsi
47          FinPour
48      FinPour
49  Sinon
50      Pour p allant de 1 à n
51          Pour q allant de 1 à m
52              Si(BEFOREK[p] et afterK[q])
53                  spatialMatch[p][q] = TRANSFORMATION
54              Finsi
55          FinPour
56      FinPour
57  Finsi
```

(5.7)

5.2.1.3 Construction d'une matrice d'appariement global

Nous proposons ici un autre algorithme, $lifeEventResearch$, qui complète l'algorithme *territorialEventResearch* de détection des changements territoriaux, qui permet d'apparier les unités de deux versions V' et V'' afin d'inférer des **évènements de vie** ($LifeEvent$) : apparition, transformation ou disparition. $lifeEventResearch$ propose des hypothèses concernant l'appariement deux unités de chaque version sur tous les critères C_k utiles désignés par l'utilisateur, et pondère le résultat global de l'appariement en fonction du poids α_k accordé par l'utilisateur à chaque critère C_k, à l'aide d'une fonction de masse F.

Concernant l'empreinte spatiale, l'utilisateur peut proposer un critère de comparaison géométrique basé sur le test d'égalité précédemment proposé, mais avec un seuil epsilon ε d'acceptation plus haut (de l'ordre de 50% par exemple), signifiant que deux unités ont la même empreinte géométrique si leur

intersection est non nulle, et qu'elles partagent plus de (100-ε)% de surface (ici, 50% par exemple). Ainsi, deux unités possédant une intersection non nulle et partageant plus de 50% en commun peuvent être considérées comme égales.

La fonction de masse F entre une unité u'_i de V' et une unité u''_j de V'' vaut 1 lorsque l'appariement est total (tous les critères de comparaison sont égaux deux à deux). La fonction de masse vaut 0 lorsqu'aucun des critères de comparaison de l'unité u'_i n'est égal aux critères de comparaison d'une autre unité u''_j dans la version V''. L'algorithme $lifeEventResearch$ peut utiliser un seuil β à partir duquel une unité u'_i est considérée comme appariée avec u''_j dans une transformation. Le seuil β est un paramètre, compris entre 0 et 1, qui peut être configuré.

Il est possible de constituer une matrice d'appariement global MATCH, dont la structure est similaire à SPATIALMATCH : la matrice d'appariement global MATCH se construit à partir de deux listes ordonnées, celles des unités u' de la version enregistrée V', et celle des unités u'' de la nouvelle version de nomenclature à enregistrer V''. Dans un premier temps, pour chaque critère C_k de comparaison défini par l'utilisateur, une matrice d'appariement CKMATCH pour ce critère est construite : les valeurs valent 1 si l'égalité est complète entre deux unités u'_i et u''_j pour le critère C_k considéré, 0 sinon. Dans un second temps, MATCH agrège les résultats des r matrices CKMATCH en effectuant pour toute entrée (i,j) de MATCH la somme pondérée des entrées (i,j) des r matrices CKMATCH, voir équation 5.8

$$\text{MATCH}[i][j] = \sum_{k=1}^{r} \alpha_k \text{CKMATCH}[i][j] \tag{5.8}$$

Cette proposition possède l'avantage de la souplesse : la fonction de masse peut moduler un certain nombre r de critères C_k, pondérés par des coefficients α_k, tous choisis par l'utilisateur. Un même attribut de l'unité géographique peut être choisi plusieurs fois pour servir à constituer un critère C_k, avec des méthodes de mesure d'égalité variées pour cet attribut. Par exemple, deux géométries sont égales si leur intersection est non vide et que la part de surface partagée est de 1%, ou bien de 30%, suivant la façon de paramétrer le seuil ε du test d'égalité spatiale. Il est envisageable ici d'utiliser des méthodes de mesure de la distance entre géométries plus perfectionnées, comme celles décrites par [Bel Adj Ali 01].

L'algorithme $lifeEventResearch$ est extensible car la fonction de masse est modulable. En effet, nous avons proposé une fonction d'agrégation par somme pondérée. Elle peut être remplacée par une fonction d'agrégation renvoyant le maximum ou bien le minimum des critères calculés, par exemple. Si la fonction de masse F renvoie le maximum des critères calculés, alors deux unités sont égales si au moins un des critères d'appariement est supérieur ou égal au seuil β d'appariement : F est plus permissive. Alors que si la fonction de masse F renvoie le minimum des critères calculés, il faut que tous les critères calculés soient supérieurs ou égaux au seuil d'appariement β pour que deux unités soient appariées, et, dans ce cas, F est moins permissive.

5.2.1.4 Calcul des hypothèses d'appariement

À partir de la lecture de la matrice d'appariement global MATCH, il est possible de construire des hypothèses sur le type d'évènement de vie (*LifeEvent*) qui doit être associé à chaque unité de chaque version, et qui éventuellement relie deux unités de deux versions différentes dans le cas de Transformation. Les entrées de MATCH sont les index des unités de chaque version (les lignes correspondent aux unités u' de la première version V' de nomenclature, les colonnes correspondent aux unités u'' de la seconde

version V'' de la nomenclature), et chaque cellule contient la valeur de fonction de masse comparant une unité u' et une unité u''. Les hypothèses émises en étudiant la matrice d'appariement global MATCH sont les suivantes :

- si il existe u'_i, u''_j telle que $F(u'_i, u''_j) = 1$, alors u'_i et u''_j sont la même unité ;
- si il existe u'_i, u''_j telle que $1 > F(u'_i, u''_j) > \beta$, alors u'_i est liée à u''_j par un évènement de transformation ;
- si pour tout u''_j, $\beta \geq F(u'_i, u''_j) \geq 0$, alors u'_i a disparu ;
- si pour tout u'_i, $\beta \geq F(u'_i, u''_j) \geq 0$, alors u''_j est apparue.

Les hypothèses sur les évènements de vie sont indépendantes des hypothèses sur les évènements territoriaux qui, elles, sont construites à partir de la lecture de la matrice SPATIALMATCH. Il est intéressant et possible de configurer des seuils epsilon de test d'égalité géométrique différents pour constituer les deux matrices. Pour obtenir un résultat exploitable, le test d'égalité géométrique utilisé pour détecter des évènements territoriaux et construire SPATIALMATCH devrait être plus fort que le test d'égalité géométrique utilisé pour apparier deux unités et construire MATCH. En effet, il est envisageable qu'à l'intérieur d'une redistribution, des unités soient considérées comme appariées entre deux versions, car leurs frontières ne sont pas trop modifiées. Ceci est rendu possible par l'usage d'un test d'égalité faible pour l'appariement dans MATCH, conjointement avec un test d'égalité géométrique fort pour la détection des évènements territoriaux dans SPATIALMATCH.

Il est possible de conserver dans la base de données ces hypothèses sans les modifier, mais il est important que l'expert puisse contrôler ces hypothèses afin de les modifier si besoin, en particulier s'il considère qu'une unité ne s'est pas transformée, mais a disparu, ou inversement, qu'une unité qui semble ne pas s'apparier avec aucune autre est en réalité appariée à une unité existant dans la version précédente. En effet, deux unités complètement différentes (par leur code, leur désignation, et leur géométrie) peuvent cependant êtres considérées comme étant les mêmes pour un expert. Par exemple, dans l'exemple du Danemark introduit précédemment et illustré dans la Figure 2, les unités DK003 et DK013 peuvent être considérées comme égales par l'expert, bien qu'ayant à la fois changé d'empreinte spatiale, de désignation et de code. L'introduction d'un critère de comparaison spatial basé sur un test d'égalité faible (fort epsilon ε), et affecté d'un poids relativement fort, peut affiner le résultat des tests d'appariement. Cependant, suivant les régions d'étude, il faut que l'expert puisse vérifier au cas par cas les hypothèses. Par exemple, dans un cas, deux unités partageant 50% de surface commune seront considérées comme égales, et dans un autre, deux unités partageant 65% de surface commune ne seront pas considérées comme égales par l'expert.

Enfin, il est envisageable d'étendre les critères d'appariement à d'autres attributs avec référence spatiale et temporelle, qui ne soient pas stockés directement dans le modèle. Par exemple, une grille de population de résolution fine, de type un kilomètre par un kilomètre, pourrait permettre d'apparier les unités non pas sur la quantité de surface spatiale partagée, mais sur la quantité de surface géographique peuplée partagée. Cette hypothèse possède une forte résonance avec la problématique du géographe qui s'intéresse aux territoires et à leur peuplement. En effet, l'identité d'un territoire est fortement liée à celle de la population (histoire, culture) qui l'habite.

Voici un scénario qui illustre ce que pourrait apporter le croisement avec un attribut de peuplement. Imaginons comme dans la figure 5.22 qu'une unité g_1 soit constituée essentiellement de désert, avec une oasis A (représentée par un palmier) où se concentre la population. Dans une autre version du zonage, le géographe étudie maintenant g_2, qui a quasiment la même emprise spatiale, mais qui inclut l'oasis B, et non pas l'oasis A. On peut imaginer que les deux oasis A et B sont habitées par des populations de cultures différentes, et que pour l'expert, ces deux unités ne sont donc pas identiques. Cependant, le test

basé uniquement sur la surface spatiale partagée dirait que les deux unités doivent être appariées, bien que finalement, elles ne partagent en commun qu'un seul immense désert. Au contraire, un test basé sur la surface de territoire peuplé partagée dira que g_1 et g_2 ne devraient pas être appariées.

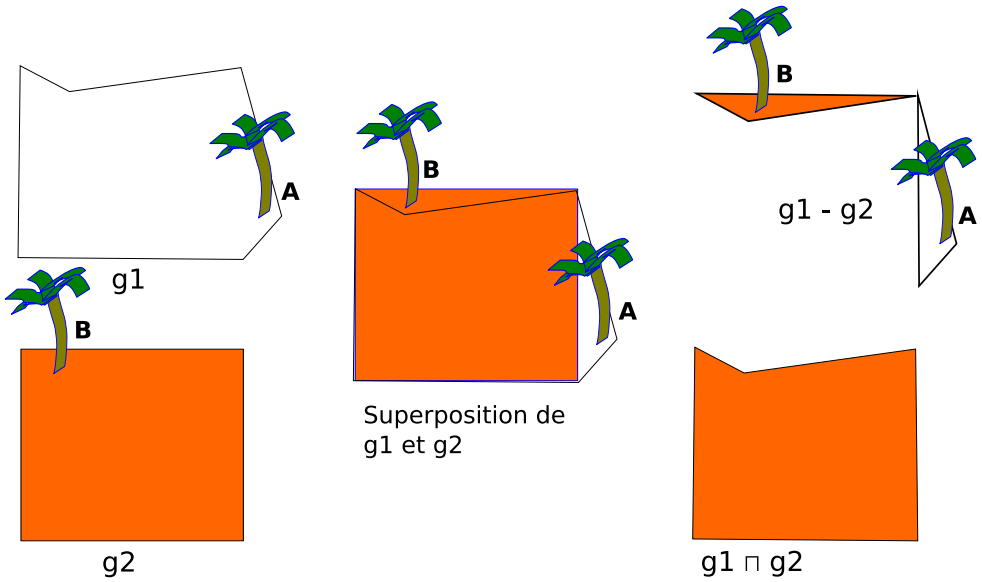


FIGURE 5.22 – Exemple de l'intérêt de considérer la densité de peuplement pour l'appariement des unités géographiques.

5.2.1.5 Validation et expérimentations

Nous avons implémenté ces deux algorithmes, $lifeEventResearch$ et *territorialEventResearch* en Java au dessus de la base de données *ESPON 2013 database*.

Pour la réalisation du test de distance surfacique, nous nous sommes servis des primitives implémentées dans PostGIS, voir les deux extraits de code 5.1 et 5.2. Concrètement, les tests d'intersection doivent aussi être paramétrés car des unités qui ne s'intersectent pas (comme la Belgique et le Danemark ou bien la République Tchèque et l'Autriche) sont dites intersectées d'après le test d'intersection de PostGIS, qui est conforme au modèle des 9i défini par [Egenhofer 91].

```
public DBInputMatchData() throws Exception{
  createDBConnection();
  //stIntersection = db.prepareStatement("SELECT intersects(?, ?)");
      // Surpringly, this intersects return true for (BE, DK) and (
     CZ, AT)
  stIntersection = db.prepareStatement("SELECT␣(area(intersection(?,
     ␣?))/area(geomunion(?,␣?))>0.02)");
  stEquals = db.prepareStatement("SELECT␣area(symdifference(?,␣?))
     *1.0␣/␣area(intersection(?,␣?))␣");
  stUnion = db.prepareStatement("SELECT␣geomunion(?,?)"); //throw
     some exception sometimes
}
```

Listing 5.1 – Extrait de code Java réalisant le test de distance surfacique.

```
/*
 * Test using postgis function is there is an intersection between
    the two outlines, and compute the ratio of the non-shared area
    by the shared area.
 * @return true if the ratio is inferior to epsilon.
 * When one of the tests fails, or the query on DB fails, we return
    false.
 * @see util.IInputMatchData#isEquals(org.postgis.Geometry, org.
    postgis.Geometry, double)
 */
public boolean isEquals(Geometry outline1, Geometry outline2, double
    epsilon) throws Exception {
  boolean isEquals = false;
  try {
    stEquals.setObject(1, new PGgeometry(outline1));
    stEquals.setObject(2, new PGgeometry(outline2));
    stEquals.setObject(3, new PGgeometry(outline1));
    stEquals.setObject(4, new PGgeometry(outline2));

    ResultSet rs1 = stEquals.executeQuery();
    while (rs1.next()) {
      double difference =  rs1.getDouble(1);

      if (difference < epsilon) {
        isEquals = true;
      }
    }
  } catch (SQLException e) {
    throw new Exception (e.getMessage());
  }
  return isEquals;
}
```

Listing 5.2 – Extrait de code Java réalisant le test de distance surfacique.

Le test d'égalité de nom a été réalisé en utilisant la distance de Levenshtein [Levenshtein 65], avec un petit ajustement sur les noms courts : en dessous de 6 caractères, l'écart toléré pour la différence vaut le quart de la longueur de la chaîne testée. Cette distance a l'avantage de nous permettre de gérer des modifications mineures de chaînes de caractères, et de contourner les coquilles qui se glissent à l'acquisition des informations dans le système. Par exemple, des unités DK00F (niveau 3 de la NUTS version 2003) et DK050 (niveau 3 de la NUTS version 2006) ont un nom qui diffère très peu : DK00F s'appelle « Nordjyllands Amt » puis perd son suffixe 'Amt' et s'appelle alors « Nordjylland ». La distance de Levenshstein trouve ces deux noms égaux. Comme leurs deux géométries sont égales à 10 % près, on les trouve appariées.

La base de données *ESPON 2013 database* a servi à valider notre algorithme, à la fois en termes de rapidité comme en termes de complétude et d'exactitude des résultats. Étant donné que le coût principal de l'algorithme réside dans des comparaisons de géométries, nous donnons les caractéristiques des géométries des unités géographiques qui ont été insérées dans la base de données *ESPON 2013 database*. Cette base contient les unités géographiques de la nomenclature NUTS, à tous les niveaux (du niveau départemental au niveau des états) pour les versions suivantes : 1995, 1999, 2003, 2006. Les géométries

de ces unités ont été établies par le projet *ESPON 2013 database* à un niveau de généralisation qui sert à la réalisation des cartes des projets Européens travaillant pour ESPON [Grasland 10c]. Le tableau 5.4 en résume les caractéristiques importantes pour le calcul de test d'intersection en donnant le nombre de points (minimum, maximum, moyen), et l'écart-type sur cette mesure.

TABLE 5.4 – Caractéristique des géométries utilisées pour la validation : nombre de points des contours des unités géographiques.

niveau	moyenne	minimum	maximum	écartType
NUTS0	128	8	473	119
NUTS1	74	7	268	46
NUTS2	39	5	166	21
NUTS3	21	4	99	10

Comparativement au temps qu'il a fallu aux équipes de recherche pour réaliser ce dictionnaire du changement des unités à la main (plus d'un an, [Ben Rebah 11]), les résultats de notre algorithme sont un gain considérable : le tableau 5.5 donne les temps d'exécution pour l'appariement de la NUTS par comparaison des versions 1999 et 2003 puis 2003 et 2006. En une demi-heure, notre programme établit la généalogie des unités entre deux versions d'une nomenclature complète, comprenant 1907 unités. La question du passage à l'échelle (par exemple, sur l'ensemble des 36 000 communes françaises) pourrait être résolue par la parallélisation du calcul sur une grappe de processeurs partageant un accès en mémoire à la matrice SPATIALMATCH. En effet, le calcul de la matrice SPATIALMATCH lors du premier parcours comme expliqué page 168 dans l'encadré 5.4 est la partie la plus couteuse en temps (99% du temps total d'exécution). Or, le calcul des intersections des unités dans cette matrice se répartit directement entre plusieurs processeurs, car c'est une tâche qui n'a pas de dépendance avec d'autres.

TABLE 5.5 – Mesures de temps de calculs pour l'appariement, niveau par niveau, version par version.

Niveau	secondes	Nombres d'unités en 1999	Nombres d'unités en 2003	Nombres d'unités en 2006
NUTS0	7	42	42	-
NUTS1	21	92	110	-
NUTS2	104	280	314	-
NUTS3	1576	1441	1441	-
NUTS0	7	-	42	43
NUTS1	26	-	110	115
NUTS2	124	-	314	317
NUTS3	1872	-	1441	1590

Il s'agit également de vérifier la validité des résultats. L'algorithme complet ($lifeEventResearch$ et *territorialEventResearch*) est exécuté avec les paramètres décrits dans le tableau 5.6.

TABLE 5.6 – Paramètres du test d'appariement pour la NUTS entre les versions 2003 et 2006.

Paramètre	Valeur
Seuil β d'appariement global	1/2
Seuil ε_1 pour le test de distance surfacique testant la présence d'évènements	0.01
Seuil ε_2 pour le test de distance surfacique testant l'égalité des géométries	0.2
Poids du critère "géométrie"	1/3
Poids du critère "code"	1/3
Poids du critère "unité supérieure"	1/6
Poids du critère "nom"	1/6

Le tableau 5.7 indique le nombre d'évènements territoriaux détectés, classés suivant leur type, ainsi que le nombre d'unités correctement appariées sur la comparaison des versions 2003 et 2006. Nous les avons comparés au travail manuel réalisé dans l'étude du projet ESPON, et n'avons trouvé qu'un petit nombre d'évènements territoriaux non reconnus. La cause vient de ce que certaines géométries posent des problèmes de comparaison et supportent mal les opérations d'union géométrique, entraînant un abandon forcé pour l'algorithme dans le second parcours (décrit dans les encadrés 5.5, 5.6 et 5.7 page 171). Par ailleurs, toutes les transformations (de code, de nom, de géométrie) sont trouvées.

TABLE 5.7 – Résultat du programme de construction de la généalogie des unités de la NUTS entre 2003 et 2006.

Niveau	Fusion	Scission	Redistribution	Apparition	Transformation	Disparition
NUTS0	0	0	0	1	42	0
NUTS1	0	2	1	8	107	3
NUTS2	1	2	2	13	304	10
NUTS3	5	9	11	101	1489	61

Sur l'ensemble des évènements territoriaux reconnus (25 pour $\varepsilon_1 = 0.01$), seuls 4 manquent au niveau NUTS3, que nous listons ci-dessous. Dans ces cas, l'algorithme a simplement indiqué que les unités précédant chaque évènement (les unités à gauche des équations de cette liste) ont disparu, et que les unités leur succédant (les unités à droite) sont apparues.

– $DK00A + DK00B + DK00C + DK00D + DK00E \rightarrow DK041 + DK042 + DK032$
– $ITG22 + ITG24 \rightarrow ITG2A + ITG2B + ITG2C + ITG26 + ITG27$
– $PL321 + PL322 \rightarrow PL323 + PL324 + PL325 + PL326$
– $PL341 + PL342 \rightarrow PL343 + PL344 + PL345$

Il est aisé de repérer les erreurs car, pour ces paramètres, nous n'avons relevé aucun faux appariement, mais plutôt des absences d'évènements territoriaux, qui se traduisent par des Apparitions ou des Disparitions qui ne sont pas liées par des évènements. Le programme restreint donc de façon utile la liste des résultats à traiter manuellement.

En changeant les paramètres, avec par exemple $\varepsilon_1 = 0.1$ et $\varepsilon_2 = 0.2$, plus d'évènements territoriaux sont identifiés (car les paramètres du test spatial sont plus lâches). Ainsi, les évènements précédemment listés sont identifiés, mais également l'indépendance du Monténégro par rapport à la Serbie au niveau NUTS0, marquée par un évènement d'Extraction. Cependant, des erreurs peuvent se glisser (par exemple

un évènement entre LT004 et LT007 est inféré, alors qu'il n'existe pas). D'autres évènements sont manqués, comme par exemple, la redistribution se produisant au Danemark entre les unités DK001, DK002, et DK003 de la version 2003 et les unités DK011, DK012, et DK013 de la version 2006.

Ainsi, malgré son efficacité manifeste, cet algorithme présente une certaine variabilité dans ses résultats, qui dépend à la fois des paramètres des tests de distance et de la pondération des critères. Cette incertitude sur les résultats, que nous proposons d'explorer avec d'autres outils dans nos perspectives, justifie la proposition d'une interface pour aider à leur analyse et leur correction par l'expert.

5.2.2 Pilotage de l'appariement par un expert

Un aspect important de cette recherche concerne l'exploitation de ces résultats. Du fait de l'incertitude existant sur les résultats de l'appariement, il est nécessaire de faire intervenir la validation d'un expert. Nous proposons donc ici une interface graphique permettant à l'expert d'interagir avec le système pour réaliser différentes tâches en vue d'une exploration interactive des résultats, pour leur contrôle ou leur analyse, voire leur modification. La première tâche consiste à configurer la fonction de masse et lancer l'exécution de l'algorithme. La seconde tâche lui permet de vérifier et corriger les hypothèses de vie calculées par notre algorithme. Durant la troisième tâche, l'expert peut également visualiser les évènements territoriaux calculés, modifier les dates des évènements, et enrichir la connaissance sur ces évènements, via quelques métadonnées, et éventuellement un ou plusieurs documents décrivant plus en substance l'évènement.

Comme le montre la figure 5.23, les trois tâches s'enchaînent dans un cycle d'analyse, qui peut être réitéré par l'expert jusqu'à satisfaction. Nous proposons un canevas pour effectuer ces trois tâches, en exploitant au mieux les informations collectées dans les matrices d'appariement spatial et global.

5.2.2.1 Description des tâches

Configuration de la fonction de masse

Cette tâche est en amont ou en aval des deux autres tâches : elle permet de configurer l'algorithme d'appariement. L'utilisateur peut sélectionner une nomenclature, ainsi que les versions à comparer, et définir les paramètres de l'algorithme d'appariement. Les critères d'appariement sont sélectionnés à ce moment-là, et pondérés suivant les préférences de l'utilisateur. Également, sont définis le seuil d'appariement, et la finesse du test d'égalité géométrique pour l'appariement spatial ou pour la détection d'évènements territoriaux. Lorsque l'utilisateur lance l'exécution, les matrices d'appariement spatial et global sont calculées.

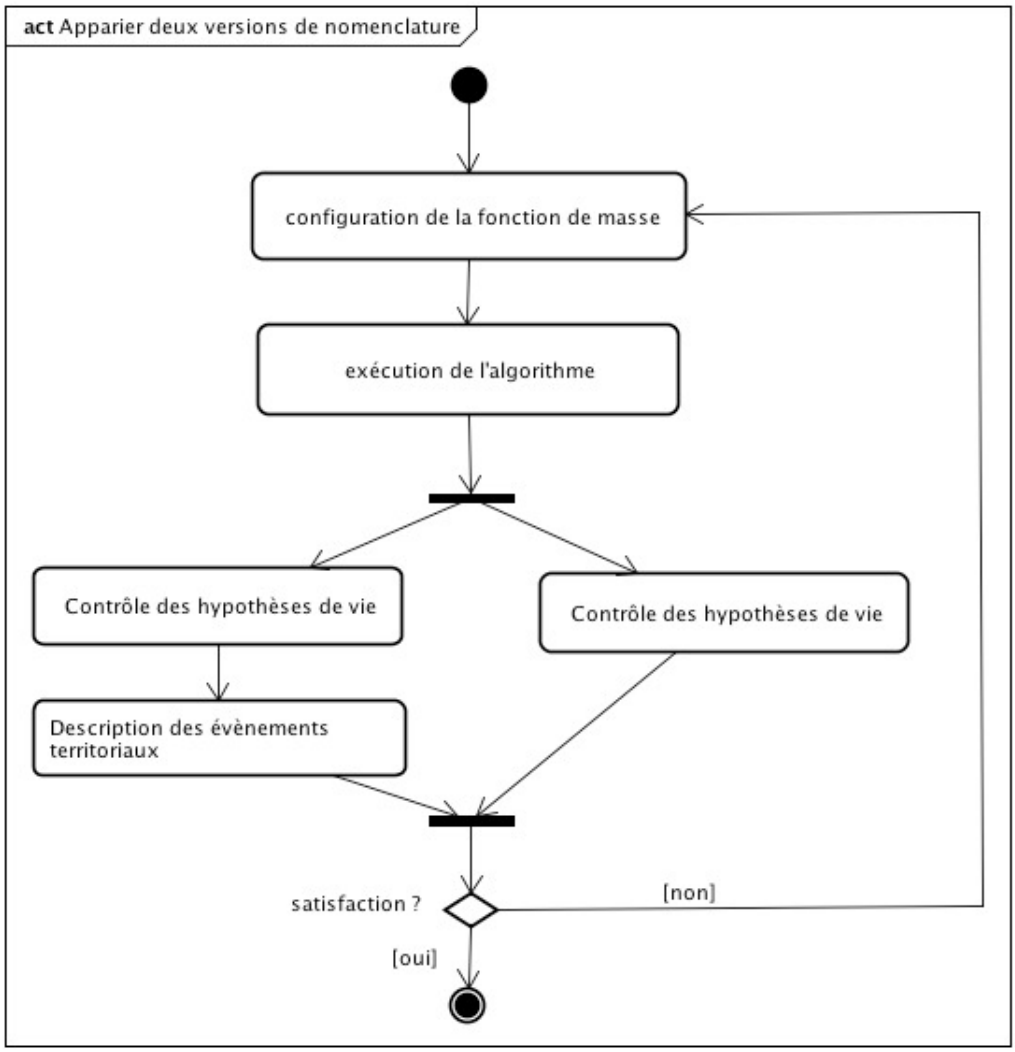


FIGURE 5.23 – Cycle d'analyse comparative de deux versions de nomenclature en mode semi-automatique.

Contrôle des hypothèses de vie

La seconde tâche concerne le contrôle des hypothèses de vie (apparition, transformation ou disparition), mais implique un trop grand nombre d'unités *a priori* pour l'expert. Ceci est du à un phénomène bien connu des psychologues cognitiens (le « change blindness » en anglais), [Rensik 02] : le défilement des images devant l'oeil humain ne lui permet pas de détecter les parties modifiées lorsque trop d'objets sont soumis à son attention. En revanche, le nombre d'évènements territoriaux est, de façon générale, plus restreint : il est plus facile de travailler à partir de la liste des évènements territoriaux, et de traiter chacun successivement, que de gérer un par un les évènements de vie. L'idée est qu'en traitant un évènement territorial, toutes les unités qui sont concernées par l'évènement le seront aussi. Par exemple, dans le cas d'une redistribution territoriale, l'expert pourra vérifier les hypothèses de vie pour chacune des unités impliquées. Pour le faire de façon convenable, il doit connaître d'un seul coup d'œil les unités impliquées dans l'évènement en question, avant et après, et accéder rapidement à leurs attributs : position spatiale, nom (dans toutes les langues), code, centre, unité supérieure. Il faut aussi qu'il puisse

voir comment la fonction de masse a été évaluée pour chacune des unités, et la valeur de chacun des critères d'appariement, afin d'ajuster éventuellement la fonction de masse dans un autre cycle d'analyse. Étant donné que des unités peuvent s'être transformées indépendamment de tout évènement territorial, l'interface doit aussi proposer de sélectionner les évènements de vie, triés en fonction de leur type (*Apparition, Disparition, Transformation*), qu'ils soient liés ou pas à un évènement territorial. Concernant les évènements de type *Transformation*, il est aussi intéressant de pouvoir les trier en fonction du critère de transformation.

Description des évènements territoriaux

La troisième tâche est en lien étroit avec la seconde tâche car, en modifiant le statut d'une unité impliquée dans un évènement territorial, l'expert peut subséquemment modifier le sous-type de l'évènement territorial. Si, par exemple, dans une union (évènement de type *Merge*), l'expert corrige le statut d'une unité en indiquant qu'elle s'est transformée, ceci change automatiquement le sous-type de l'évènement qui, de *Fusion*, devient *Integration*.

Afin de garantir la cohérence du modèle et la vérification des contraintes d'appariement géométrique pour chaque type d'évènement, l'utilisateur ne peut pas changer le type primaire des évènements territoriaux, ni ajouter ou supprimer des unités impliquées dans chaque évènement. Ainsi, l'évènement territorial qui a été inféré par le test d'égalité fort n'est pas impacté dans sa composition par le changement d'appariement des unités. Simplement, la spécialisation de son type est éventuellement impactée. C'est aussi pour cette raison que l'utilisateur ne peut apparier deux unités qu'à la condition qu'elles soient impliquées dans le même évènement territorial, ou bien associées à aucun évènement.

Il peut également, au niveau de chaque évènement, éditer des informations descriptives, comme la date de l'évènement, ou bien les métadonnées (nom de l'expert, date d'édition des données, et description textuelle de l'évènement), et lier l'évènement à un document décrivant les raisons du changement. Concernant la date de l'évènement, il est supposé s'être produit à la date de publication de la nouvelle nomenclature. L'expert peut modifier cette hypothèse, mais la date saisie (qui est précisée au format international YYYY-MM-DD, c'est-à-dire en année, mois, et jour) devra toujours être comprise entre la date de publication de la version V' et la date de publication de la version V''. Les métadonnées sont fournies dans un format semi-structuré, le format XML, qui respecte le schéma de la norme ISO 19115 pour des données géographiques. L'utilisateur doit seulement mentionner la date de saisie des informations descriptives, ainsi que l'auteur, mais il peut aussi :

- citer la source de ses informations, de façon précise ;
- décrire l'évènement en question en y associant des thèmes, des mots-clés, ainsi qu'un texte descriptif.

Nous ne détaillons pas ici comment construire un tel formulaire XML pour ces métadonnées. Nous supposons pour l'instant qu'il insère directement sa description dans un champ-texte libre.

5.2.2.2 Proposition d'interface

Par rapport à ces tâches, il convient de proposer une interface graphique qui facilite l'appropriation du programme et de ses résultats par un non-informaticien. Le but de cette interface est de permettre la réalisation des tâches citées, dans l'optique de faciliter l'exploration des changements survenus entre deux versions, et d'autoriser l'utilisateur a contrôler les résultats produits par l'algorithme. La configuration de la fonction de masse est une tâche suffisamment indépendante des deux autres pour être présentée

dans un panneau de configuration séparé, dans lequel l'utilisateur pourra consulter les paramètres de l'algorithme, les modifier et relancer l'exécution si il le souhaite.

Panneau de configuration de l'algorithme

Le panneau de paramétrage de l'algorithme doit d'abord permettre de sélectionner les données à comparer : la nomenclature, les versions, et le niveau de généralisation des géométries à utiliser pour chacune des versions de nomenclature. Il faut si possible choisir des niveaux proches ou identiques. Lorsque ce choix est réalisé, il est validé. Ensuite l'utilisateur peut définir la fonction de masse associée à ces données.

La fonction de masse est définie par le type d'agrégation utilisée (somme, minimum, ou maximum) et le seuil d'appariement β des unités. Par ailleurs, les critères C_k de comparaison possibles pour la nomenclature choisie sont énumérés dans une liste : l'utilisateur peut en ajouter (en cochant le critère), en enlever (en le décochant) ou éditer le critère sélectionné. Il doit exister au minimum un critère de comparaison associé à la fonction de masse. Un critère de comparaison C_k est un attribut a_k choisi parmi les attributs existants pour les unités de la nomenclature considérée. Par exemple, le code, le nom dans une langue, l'unité supérieure et la géométrie sont les seuls attributs disponibles dans la nomenclature des NUTS, auquel est associé un coefficient α_k de pondération. Pour l'attribut géométrie ($outline$), il faut spécifier également le seuil global epsilon ε qui définit le niveau de tolérance du test d'égalité : un epsilon à 1% définit un test d'égalité forte, et lorsque ce seuil augmente, le test d'égalité devient plus lâche.

Dans ce panneau, le test d'égalité géométrique forte pour la détection d'évènements territoriaux doit être configuré séparément. L'utilisateur doit donc choisir un autre seuil epsilon, pour ce test, avec si possible une valeur plus faible que celle choisie pour l'appariement des unités. La figure 5.24 donne un exemple de ce que pourrait être ce panneau.

L'exécution se lance en poussant le bouton intitulé « $validate$ ». En poussant le bouton intitulé « $cancel$ », les modifications apportées sont ignorées. Ce panneau sert aussi à consulter la dernière configuration de l'algorithme.

Panneau de traitement des résultats

Le panneau de traitement des résultats doit permettre d'effectuer les tâches de contrôle des hypothèses de vie, simultanément avec celles d'étude des évènements territoriaux. En effet, certains évènements de vie seront liés à des évènements territoriaux, et d'autres non. Si l'utilisateur sélectionne un évènement territorial, les évènements de vie associés doivent apparaître, et *vice-versa*. Si un évènement de vie sélectionné ne correspond à aucun évènement territorial, la partie réservée à l'édition d'évènement territoriaux devrait être vidée de toute information pré-calculée. La figure 5.25 donne un aperçu de ce panneau. Les parties en grisé ne sont pas éditables et affichent une information lue dans le modèle ou les matrices d'appariement, alors que celles en blanc sont éditables : ce sont soit des radio boutons, des cases à cocher, ou des champs à texte libre. Ce panneau est composé de trois parties :

- dans la première partie, en haut à gauche, l'utilisateur sélectionne des évènements ;
- dans la seconde partie en bas à gauche, les informations concernant l'évènement territorial étudié apparaissent ;
- dans la troisième partie, à droite, l'utilisateur visualise et édite les informations liées à un évènement de vie.

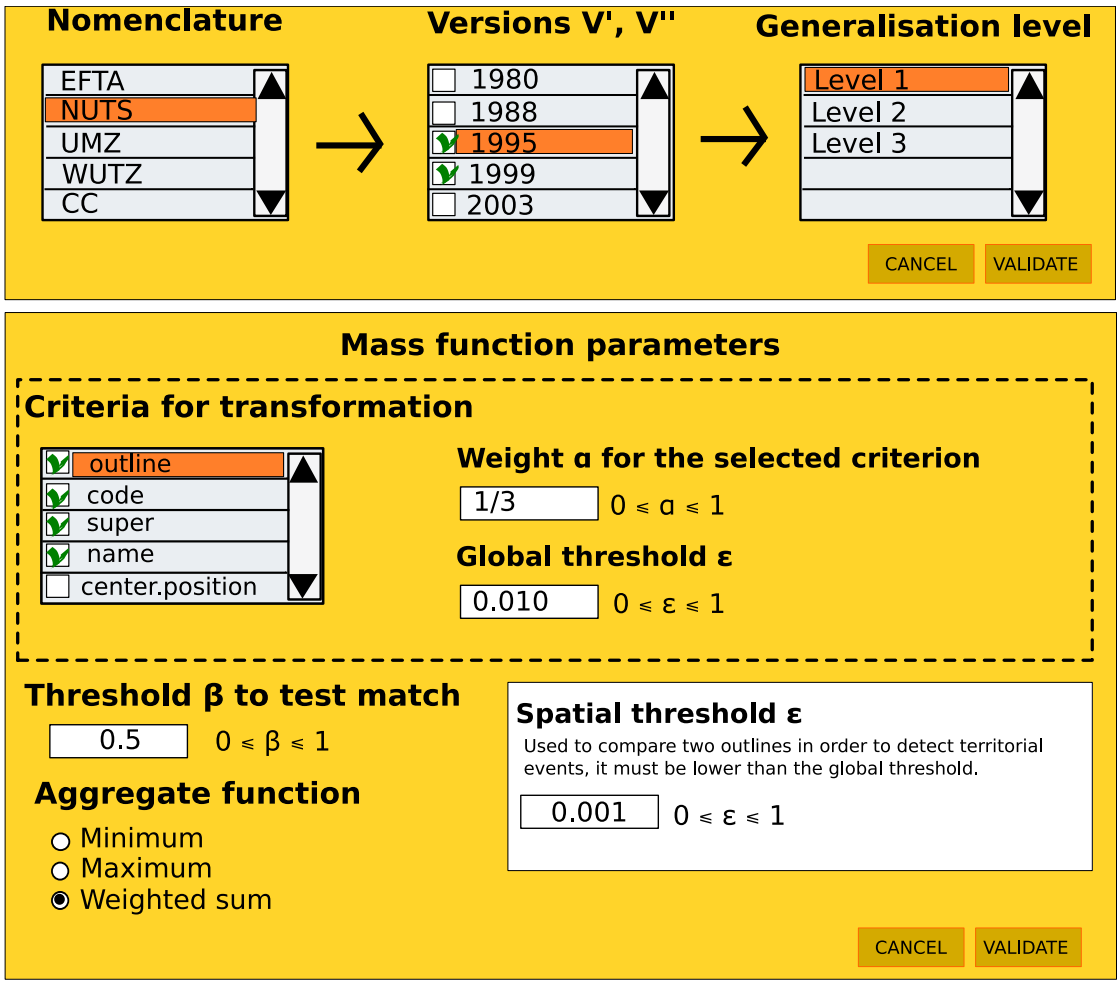


FIGURE 5.24 – Interface de configuration de l'algorithme d'appariement entre deux versions de nomenclature.

Dans la première partie, les évènements territoriaux sont proposés dans une liste (*TerritorialList*), à travers laquelle l'utilisateur peut sélectionner un évènement à contrôler. Cette partie propose également une entrée par évènement de vie, avec une seconde liste (*LifeList*). La sélection d'un évènement territorial met à jour la liste des évènements de vie. Et réciproquement, la sélection d'un évènement de vie positionne la sélection sur l'évènement territorial correspondant dans la première liste, si l'évènement de vie est lié avec un évènement territorial. Les deux listes peuvent être filtrées suivant les types d'évènements : fusion (*Merge*), division (*Split*), ou remembrement (*Redistribution*)) pour la liste *TerritorialList*, apparition (*Appearance*), transformation (*Transformation*), disparition (*Disappearance*) pour la liste *LifeList*. La seconde liste *LifeList* peut aussi être filtrée suivant les différents critères C_k de transformation définis dans la fonction de masse, par changement du code et/ou changement de la géométrie par exemple. Il est donc possible dans ce panneau de choisir un évènement territorial (*TerritorialEvent*), et/ou un évènement de vie (*LifeEvent*). Lorsque l'expert a vérifié un évènement, il peut le valider et il apparaît alors en gras dans la liste d'évènements, et le décompte d'évènements analysés s'incrémente.

En sélectionnant un évènement de vie, les informations associées aux unités impliquées dans l'évènement apparaissent dans la troisième partie à droite. Celle-ci se compose à la fois d'une zone pour la visualisation des géométries dans les deux versions, et d'une zone d'affichage des informations concer-

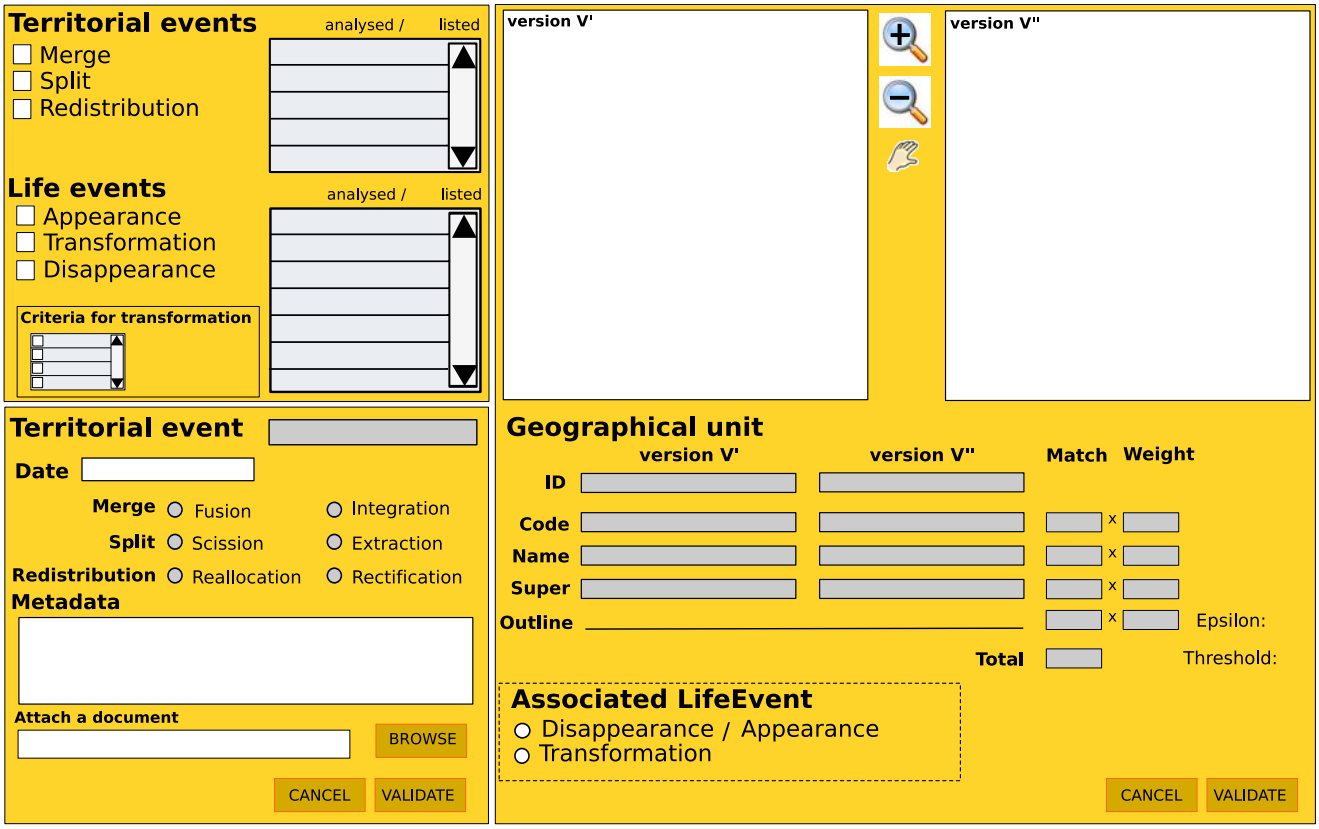


FIGURE 5.25 – Panneau de sélection, analyse et édition des évènements.

nant l'évènement de vie sélectionné. La zone cartographique est composée de deux cartes vectorielles, une pour la version V', et l'autre pour la version V''. L'utilisateur peut zoomer et dézoomer simultanément sur les deux cartes, et se déplacer également dans les deux cartes (elles sont synchronisées). Par défaut, en sélectionnant un évènement territorial, les cartes sont centrées sur l'enveloppe des géométries impliquées. Il est possible de modifier les évènements sélectionnés en cliquant dans la zone cartographique et en choisissant une autre unité. Si celle-ci est impliquée dans un autre évènement territorial, l'évènement est sélectionné et les cartes sont re-centrées en conséquence. Les informations associées à l'unité sont dans tous les cas affichées sous la version correspondante. Si elle est apparue ou disparue, les informations apparaissent sous la version V'' ou V', respectivement. Si elle s'est transformée, les informations apparaissent sous les deux versions V' et V''. En face de chaque critère d'appariement sont affichées la valeur de l'attribut et la valeur du test d'égalité avec l'unité appariée si l'unité est appariée à une autre, ainsi que le coefficient de pondération associé à ce critère. La valeur de la fonction de masse est aussi affichée, ainsi que le seuil d'appariement. Si l'unité n'est pas appariée, elle apparaît seulement dans la version dans laquelle elle existe.

Lorsque l'utilisateur souhaite modifier l'appariement calculé, il doit sélectionner l'évènement de vie associé ou bien l'une des deux unités impliquées dans la carte. Ainsi, pour un couple d'unités (u', u'') ou un singleton sélectionné (u' ou u''), l'utilisateur peut spécifier soit :

- qu'il y a eu transformation, de l'unité u' vers l'unité u''
- qu'il y a eu disparition pour l'unité u' et apparition pour l'unité u''.

Si deux unités sont déjà appariées, il doit d'abord dissocier le couple d'unités (u', u'') sélectionné avant de pouvoir apparier u' avec une autre unité u''.

Lorsque l'utilisateur est satisfait de son édition, il la valide, et l'évènement associé apparaît alors en gras dans la liste d'évènements. Le décompte d'évènements de vie analysés s'incrémente alors.

Les deux figures 5.26 et 5.27 montrent un exemple d'analyse sur un cas de redistribution territoriale.

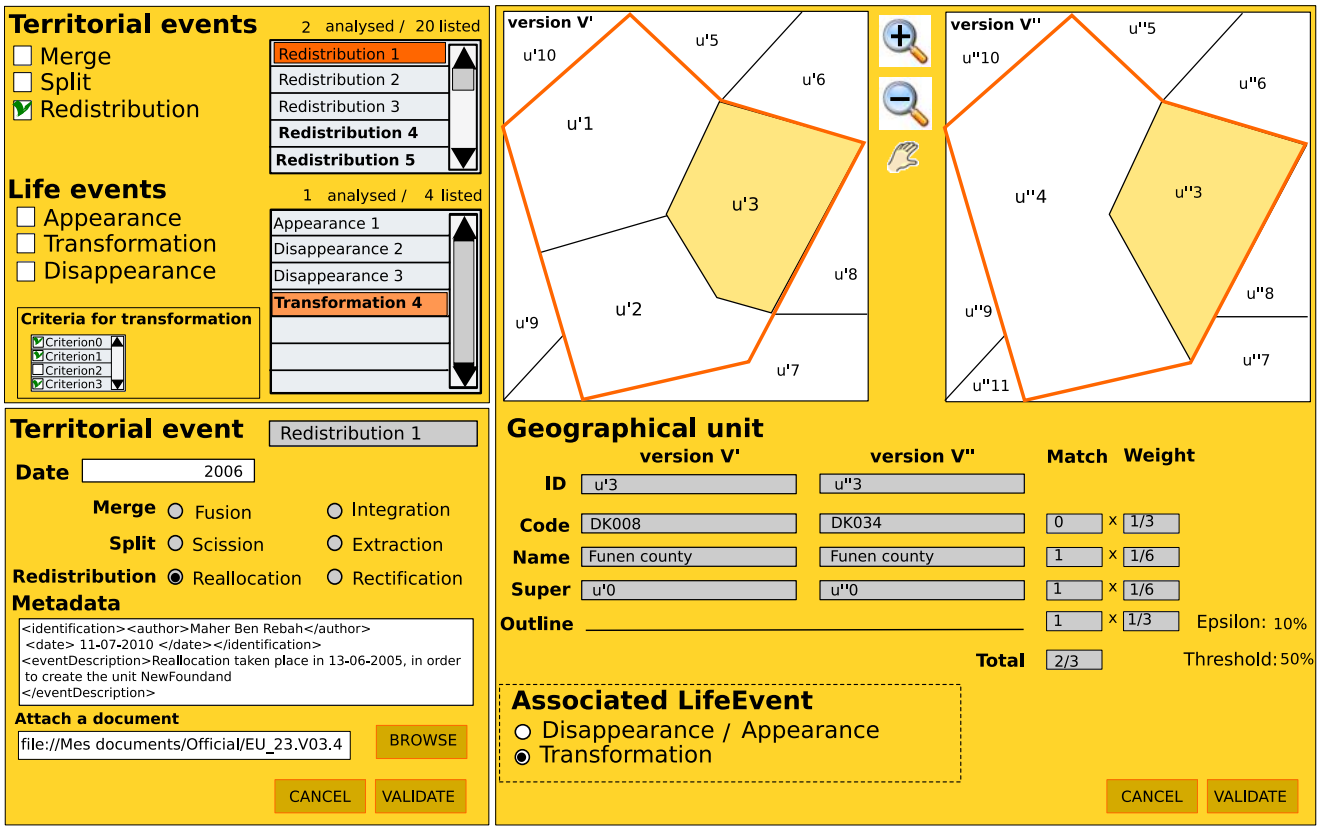


FIGURE 5.26 – Exemple d'édition d'une transformation associée à une redistribution territoriale.

Dans la figure 5.26, l'expert a déjà analysé certains évènements, ils apparaissent en gras. Il a choisi de ne visualiser que les évènements de type *Redistribution*. La $Redistribution1$ est sélectionnée pour l'édition : elle implique les trois unités $u'1$, $u'2$ et $u'3$ de la version V' et les deux unités $u''4$ et $u''3$ de la version V''. D'après les hypothèses calculées par l'algorithme, $u'1$ et $u'2$ ont disparu, $u'3$ s'est transformée en $u''3$ car seule sa géométrie a changé (pour moins de 10% de sa surface), et $u''4$ est apparue. V'' a été publiée en 2006, donc la date proposée pour l'évènement est 2006. Comme deux unités ont disparu dans cette redistribution, le sous-type inféré de l'évènement est une reallocation (*Reallocation*). L'expert a édité quelques informations descriptives concernant l'évènement territorial $Redistribution1$. Enfin, au niveau de l'évènement de vie concernant les unités $u'3$ et $u''3$, il peut encore décider qu'en réalité, ce n'est pas une transformation, et qu'$u'3$ a disparu, et $u''3$ est apparue. Lorsqu'il valide son édition, l'évènement apparaît comme validé (en gras) dans la liste des évènements de vie en haut à gauche, comme on peut le voir dans la 5.27. Cette figure 5.27 illustre l'édition de l'évènement d'apparition de l'unité $u''4$. L'unité n'est appariée à aucune autre dans la version V', et donc ses valeurs d'appariement sont nulles. Si l'utilisateur choisissait une des unités de la version V' non appariées et dans la redistribution considérée (comme $u'2$ ou $u'1$), les valeurs d'appariement seraient affichées, et vaudraient moins que le seuil d'appariement.

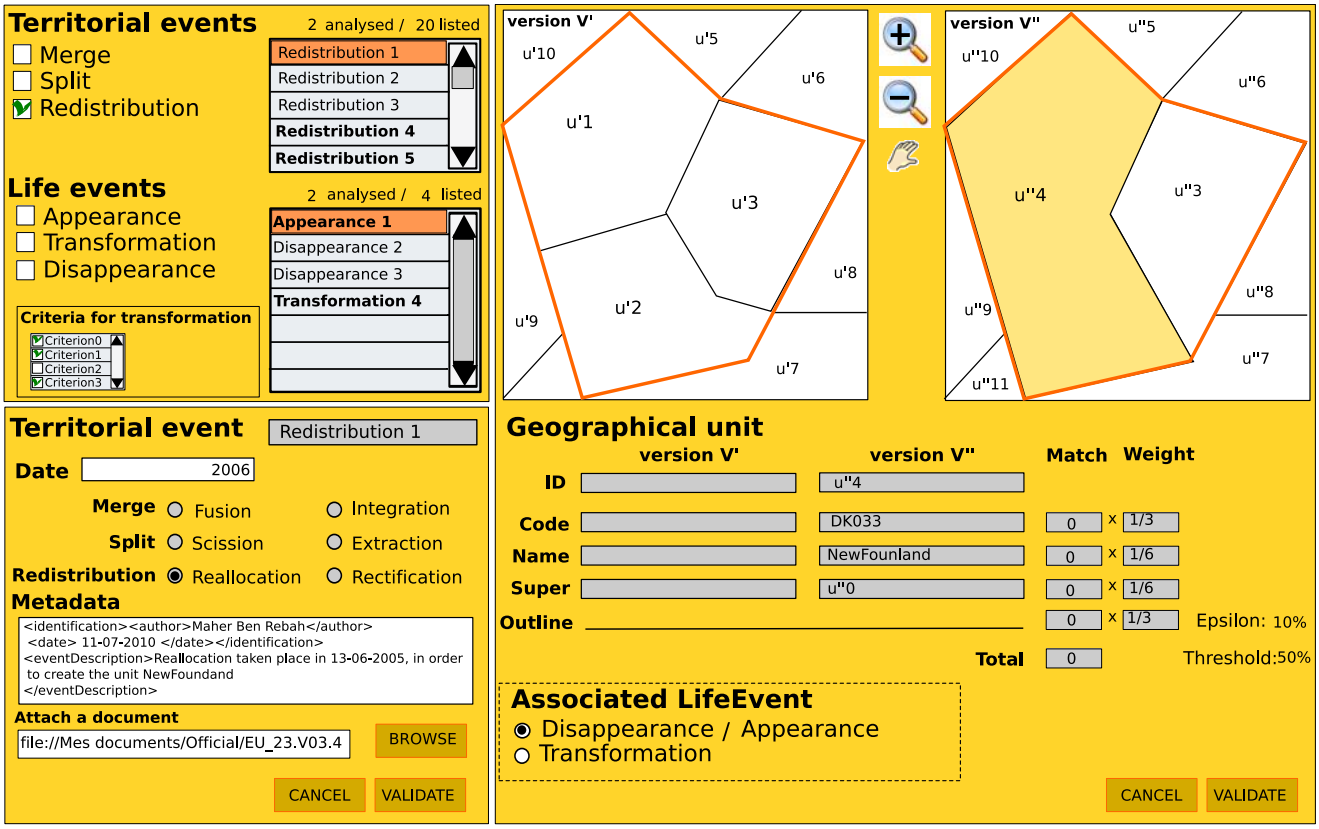


FIGURE 5.27 – Exemple d'édition d'une apparition associée à une redistribution territoriale.

5.3 Exploitation du modèle pour l'exploration interactive du changement

Dans cette section, nous proposons une analyse interactive à l'aide d'une carte de densité du changement, qui ne s'intéresse qu'aux changements territoriaux, c'est-à-dire aux changements de frontières, et d'identité des unités qui composent le zonage étudié. Cette analyse s'effectue sans disposer d'aucune information statistique territoriale, et ne nécessite que la connaissance de la généalogie des unités territoriales (les modalités d'appariement) dont le calcul a été proposé dans la section précédente, page 161.

5.3.1 Conception de la carte de densité de changement

Il s'agit de dessiner une carte illustrant la stabilité de l'organisation territoriale sur une période d'étude définie par l'utilisateur. L'objectif visé est de proposer une représentation qui puisse permettre d'identifier les zones sujettes à de nombreux changements ou au contraire stables. En effet, les représentations usuelles de cartes avec les zonages, qui sont soient juxtaposées ou qui défilent dans un mode dynamique avec un curseur temporel ne permettent pas à l'utilisateur de percevoir d'un seul coup d'oeil les changements, car ceux-ci sont trop nombreux, de nature diverses, et dispersés dans l'espace géographique. Ce problème, connu sous le nom de « cécité au changement » ou « *change blindness* », a été largement exploré par des cognitiens [Rensik 02]. La méthode manuelle qui consiste à vérifier ces changements (limités aux changements de frontières) par superposition des géométries des zonages de deux versions d'une nomenclature se révèle laborieuse. Il existe donc un besoin pour un outil focalisant

l'attention des utilisateurs sur les changements territoriaux, et de leur conséquences en terme de transformation des unités impliquées. Or, la connaissance de la généalogie des unités géographiques et de l'ensemble des évènements territoriaux peut être utilisée pour concentrer l'attention des utilisateurs sur les zones géographiques ayant changé.

Le processus de construction de la « carte de densité des évènements du changement » consiste à calculer le nombre d'évènements territoriaux dans lesquels chaque unité géographique a pu être impliquée durant la période d'étude, puis à en donner une représentation sous la forme d'une carte choroplèthe dont les limites sont celles d'une version de zonage choisie. Précisément, à partir d'une période d'étude définie par l'utilisateur, définie par une date de début (*startDate*), et une date de fin (*endDate*), et en se plaçant dans une des versions de zonage existant dans cette période, sur une date de référence (*reference*), il s'agit de décompter le nombre d'évènements e_i dans lesquels chaque unité gu_j a été impliquée, elle, ou une de ses prédécesseurs, ou une de ses successeurs, à tous les degrés, durant cette période. Les prédécesseurs de degré 1 d'une unité sont les unités impliquées dans un évènement qui provoque son apparition. Les successeurs de degré 1 d'une unité sont les unités impliquées dans un évènement qui provoque sa disparition. Les prédécesseurs de degré supérieur sont des ancêtres ou des descendants de ces unités prédecesseurs ou successeurs. Le degré indique le nombre de générations entre chaque unité. Par exemple, étudions une période couvrant 4 versions de zonages A, B, C et D constitués des unités {A_1, A_2, .., A_6} pour A, {B_1, B_2, .., B_7} pour B, {C_1, C_2, .., C_7} pour C, et {D_1, D_2, .., D_6} pour D. Une partie des unités s'apparient et se transforment en dehors de tout évènement territorial pour la majorité d'entre elles. Ainsi, A_1 est appariée avec B_1, qui s'apparie avec C_1 avec simplement un changement de code, et C_1 à son tour s'apparie avec D_1. Mais on observe également des évènements territoriaux :

- À l'instant t_1, une division : $A_2 \rightarrow B_2 + B_3$
- À l'instant t_2, une redistribution : $B_3 + B_5 \rightarrow C_4 + C_5$
- À l'instant t_3, une fusion : $C_5 + C_7 \rightarrow D_6$

C_4 et C_5 ont pour prédécesseurs de degré 1 B_3 et B_5, et A_2 comme prédécesseur de degré 2. C_5 a pour successeur D_6. Le nombre d'évènements dans lesquels C_5 est impliquée sur cette période est 3, tandis que ce n'est que 2 pour C_4 et 0 pour C_1.

À partir de ce calcul, si la carte dessinée pour l'utilisateur correspond à la version C, par exemple, les unités C_i sont colorées en fonction du nombre d'évènements décomptés, couleur qui fonce avec le nombre d'évènements. Par exemple, si la palette graphique est (blanc, jaune, orange, rouge) pour 0, 1, 2 ou 3 évènements respectivement, alors C_1 est en blanc, C_5 est en orange et C_6 est en rouge. la figure 5.28 montre les 4 cartes de densité qui peuvent être établies en fonction de la version A, B, C ou D de zonage que l'utilisateur choisit de regarder. Chacune de ces cartes, qui diffèrent suivant la version qu'il observe, permet à l'utilisateur d'avoir une connaissance immédiate de la distribution spatiale du changement territorial.

Le fait que les cartes diffèrent parait surprenant, et on s'attendrait pour une période d'étude donnée à une unique carte du changement, qui serait construite comme le *PPCD_spatial* par intersection de ces versions de zonages. Cependant, les morceaux de territoires visibles n'auraient alors probablement plus d'identité. L'intérêt de notre approche réside dans le fait que les cartes montrées à l'utilisateur correspondent à des frontières et des territoires qui existent ou ont existé.

Chaque carte suscite l'attention de l'utilisateur sur les zones qui ont changées, et nous proposons de la coupler avec une coupe temporelle permettant à l'utilisateur d'obtenir une visualisation symbolisée de la généalogie de l'unité qu'il sélectionne. Cette visualisation sous forme de graphe, dit « graphe de gé-

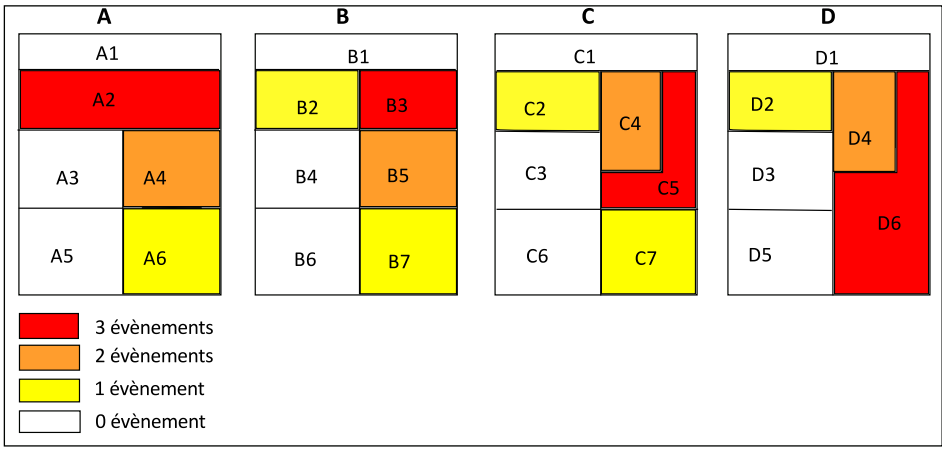


FIGURE 5.28 – Cartes de densité du changement.

néalogie », apparaît lorsque l'utilisateur clique sur une unité, et permet de détailler les transformations de cette unité au cours du temps lors des évènements territoriaux ou de vie. Cette représentation correspond à la vision du mouvement historique, définie par [Andrienko 03a]. Les évènements de vie sont d'une couleur différente des évènements territoriaux. Les surfaces allouées à la représentation de ces unités sont toujours proportionnelles à la surface réelle des unités, mais la forme reste rectangulaire, avec une largeur stable au cours du temps. En cliquant sur, par exemple, l'unité C_1, dont le graphe est linéaire, nous montrons l'ensemble des évènements de vie associés, qui eux aussi peuvent être sélectionnés, en vue d'approfondir la connaissance sur le changement indiqué. Ainsi, l'utilisateur apprend que l'unité C_1 a changé de code entre la version B et C de zonage, voir figure 5.29. Et pour l'unité C_4, il découvre qu'elle est issue d'une redistribution entre B_3 et B_5, et que B_3 était le produit d'une division de A_2, voir figure 5.30.

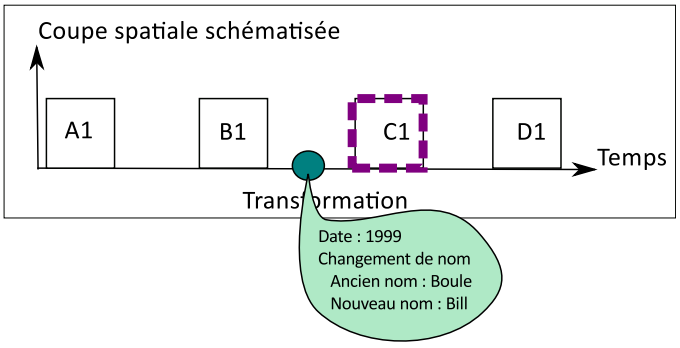


FIGURE 5.29 – Graphe de généalogie pour C_1.

Enfin, dans l'interaction inverse, cliquer sur un des évènements (par exemple figure 5.30) fait apparaitre une superposition des deux versions de zonages impliqués, celui précédant et suivant l'évènement. L'image est alors centrée sur l'unité d'intérêt et le niveau de zoom est ajusté pour ne montrer que cette unité avec ses voisines impliquées dans des évènements. Également, une description complète de l'événement apparaît, fournissant des informations descriptives à l'utilisateur : date et nature de l'événement, et conséquences pour toutes les unités. Ceci permet à l'utilisateur de comprendre en détail le changement, ce qui peut lui être utile pour l'analyse thématique du changement.

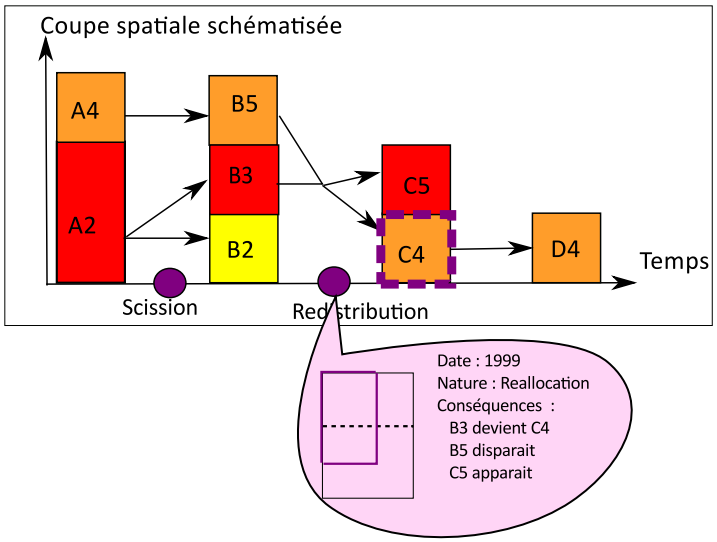


FIGURE 5.30 – Graphe de généalogie pour C_4.

5.3.2 Illustration avec l'exemple du Danemark

Après la présentation théorique de la construction d'une carte de densité du changement, et des graphes associés, un exemple complet issu d'un cas réel de changement territoriaux étant intervenus et enregistrés dans la NUTS illustre son fonctionnement. Cet exemple se situe au Danemark, que nous étudions entre 1980 et 2010, au niveau NUTS3. Les cartes de la figure 5.31 montrent comment les unités de la zone centrale du Danemark ont évolué entre 1980 et 2010, sur deux de leurs attributs identitaires : les frontières, et le code.

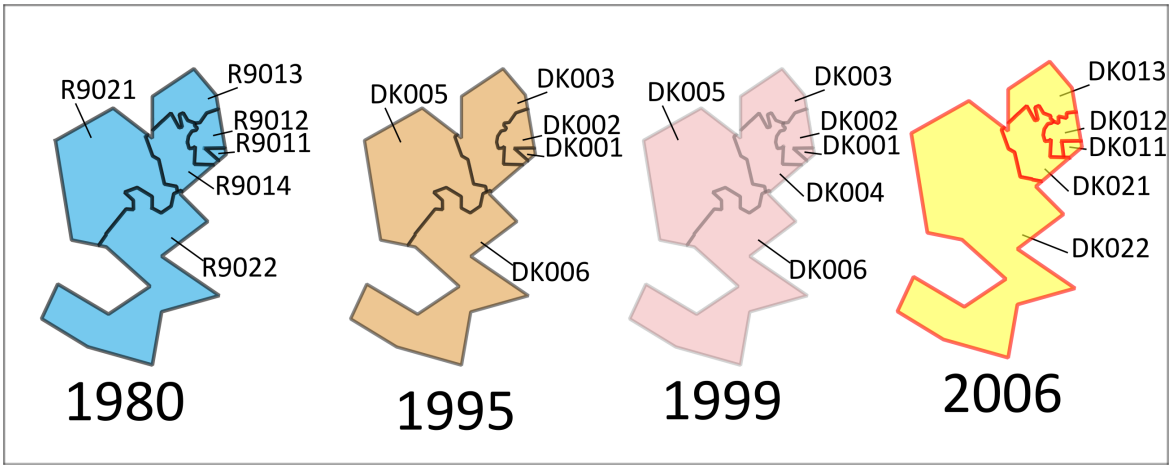


FIGURE 5.31 – Quatre versions de zonages au Danemark, recensées dans la NUTS.

Les évènements, au nombre total de quatre, qui se sont produits durant cette période dans cette région sont listés dans le tableau 5.8.

– 'DK003' est impliquées dans trois de ces évènements,
– 'DK004' (et son ancêtre 'R9014') est impliquée dans deux de ces quatre évènements,
– 'DK001' et 'DK002' sont impliquées chacune dans un même évènement de rectification,
– 'DK005' and 'DK006' sont impliquées chacune dans un même évènement de fusion.

TABLE 5.8 – Liste des évènements (territoriaux et de transformation) s'étant produits dans l'espace de la région centrale du Danemark depuis 1980.

Evènement	Date	Unités précédent l'évènement	Unités succédant à l'évènement
Fusion	1995	R9013, R9014	DK003
Scission	1999	DK003	DK003, DK004
Fusion	2006	DK005, DK006	DK022
Rectification	2006	DK003, DK002, DK001	DK013, DK012, DK011

Nous nous concentrons sur une des transitions, entre 2003 et 2006, car une transformation majeure de la nomenclature géographique sur cette zone a réduit le nombre de « *amters* » (équivalent des unités départementales en France) de 15 à 11, par l'entremise d'opérations de fusion et redistribution. Également, cinq nouvelles régions furent aussi créées à cette occasion, afin de constituer un niveau de zonage régional qui n'existait pas auparavant. En effet, avant 2006, au Danemark, seuls deux niveaux de zonage existaient, le niveau départemental, les niveaux 2, 1, et 0 étant tous représentés par l'état du Danemark dans la NUTS.

Dans la zone que nous étudions avec cette transition, le premier évènement concerne trois unités ayant pour code 'DK001', 'DK002' et 'DK003' en 2003. Il est considéré comme un évènement de *Rectification*, car selon les experts, ces unités se sont simplement transformées, bien que leur code, leur géométrie et leur nom aient changé. Ainsi, 'DK001' nommée 'Koebenhavn Og Frederiksberg' en 2003 a été étendue en empiétant sur le territoire de l'unité 'DK002'. Par suite, elle change de code et devient 'DK011' et s'appelle 'Byen København'. De façon similaire, 'DK003', nommée 'Frederiksborg' a été étendue, en changeant de code (DK013) et de nom ('Nordsjælland'). Enfin, 'DK002' a perdu 35% de sa surface, a possède 'Københavns omegn' et 'DK012' comme nouveau nom et code .

Toujours dans cette zone, le second évènement est une *Fusion* entre les unités codées 'DK005' et 'DK006' pour former l'unité 'DK022'. Cet évènement a comme conséquences trois évènements de vie, qui sont l'*Apparition* de DK022 nommée 'Vest-og Sydsjælland' et la *Disparition* de 'DK005' et 'DK006'.

Le dernier évènement concerne l'unité codée 'DK004', nommée 'Roskilde', qui se transforme en changeant de code ('DK021') et de nom (Østsjælland). Du fait de la création d'un niveau régional, cette unité a également une nouvelle unité supérieure, codée 'DK02'. Il est modélisé comme un évènement de transformation, qui n'est lié à aucun évènement territorial, et dont le diagramme UML de la figure 5.32 donne une représentation.

La figure 5.33 montre la carte de densité du changement correspondant à la période 1980-2010, synchronisée avec le graphe de généalogie de l'unité DK003 et les informations associées à l'évènement de redistribution s'étant produit entre 2003 et 2006. Dans la carte de densité de la figure 5.33, DK003 est affichée en rouge car elle est impliquée dans trois des quatre évènements territoriaux de la période considérée et l'aire considérée, l'unité 'DK004' est affichée en orange pour deux des quatre évènements

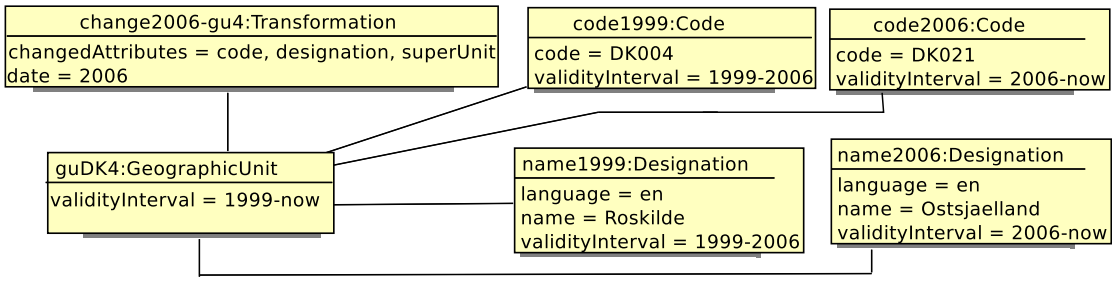


FIGURE 5.32 – Interface pour l'exploration interactive du changement territorial

territoriaux, et les autres unités sont en jaune car elles ne sont liées qu'à un seul évènement territorial.

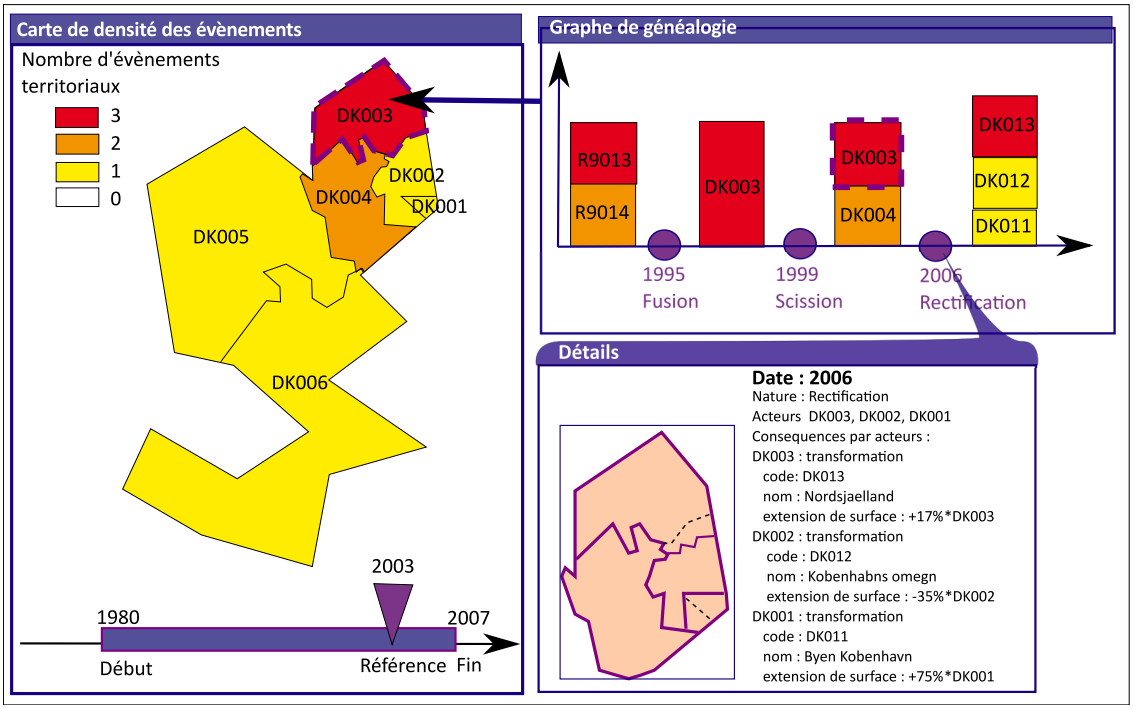


FIGURE 5.33 – Interface pour l'exploration interactive du changement territorial

Ainsi, le concept d'exploration interactive du changement à l'aide d'outils comme la carte de densité des évènements et le graphe de généalogie est utile à la compréhension du changement. Ce concept d'analyse interactive, qui exploite la connaissance des évènements capturés dans notre modèle évolutif identitaire, s'implémente de façon directe dès lors que l'appariement entre les versions de nomenclature a été calculé comme proposé dans la section précédente.

5.4 Conclusion

Dans la première partie de ce chapitre, nous avons introduit un modèle spatio-temporel objet, indexé par des évènements. Tout en soulignant son intérêt, nous avons remarqué que le principal frein au succès d'un tel modèle vient de la difficulté inhérente à l'opération d'identification : elle implique de définir des critères pour la continuation ou non de l'identité d'une unité géographique à travers le temps. Lorsque cette identification n'est pas clairement spécifiée, la mise à jour du modèle s'avère trop complexe, voire impossible.

Dans la seconde partie, nous avons proposé des critères objectifs pour comparer deux versions de nomenclature et identifier les évènements territoriaux et de vie. Pour le critère géométrique, qui peut poser problème, nous proposons d'utiliser la distance surfacique, qui a l'avantage d'être un test d'égalité paramétrable en fonction du contexte de l'étude. Également, nous avons élaboré un algorithme permettant d'apparier deux versions de nomenclature, basé sur une fonction de masse paramétrable. Cet algorithme, qui a été implémenté et validé sur les différentes versions de la NUTS entre 1995 et 2006, se montre efficace. Toutefois, ses résultats varient en fonction du paramètrage, en particulier de celui des tests de distance entre géométries. Cette incertitude nous a amené à proposer une interface en vue de piloter l'appariement des versions de nomenclature, et d'évaluer les hypothèses d'appariement automatiquement calculées. Ainsi, dans cette seconde partie, nous montrons qu'il est possible de réaliser et maintenir un modèle spatio-temporel objet, indexé par des évènements.

Dans la troisième partie, nous avons souhaité mettre en valeur le potentiel de ce modèle pour l'analyse du changement, avec un outil d'exploration interactive de ces changements. Il n'existe pas à l'heure actuelle de système équivalent permettant d'analyser les modifications territoriales. Ce type d'outil facilite l'analyse des changements sur le plan historique et géographique, car l'interrogation par la carte de densité des évènements permet de répondre immédiatement à des questions de type : où a lieu le plus fréquemment le changement, à quelle date, comment impacte-t-il les unités ? Avec cet outil, l'expert peut détecter des motifs dans le changement territorial, (plus fréquent ici que là, à telle époque, concomitant avec tel gouvernement), et en donner une interprétation politique, comme le faisait [Ben Rebah 08], mais par l'analyse manuelle et laborieuse des données. L'étape suivante sera l'analyse des valeurs statistiques associées à ces territoires en constante évolution.

Un tel modèle permet de construire un graphe spatio-temporel entre les unités de chaque version de nomenclature, qui sont liées par différentes relations (des relations de généalogie et de hiérarchie). Sur les nœuds de ce graphe se greffent les valeurs des variables statistiques associées à chaque unité géographique.

Chapitre 6

Définition et utilisation d'un profil de métadonnées pour l'information statistique territoriale

Ce chapitre propose d'adapter la norme ISO 19115 pour l'information statistique : les motifs, ainsi que les modalités de cette adaptation sont exposés dans la première section. Puis, un cadre pour la collecte puis la rediffusion des données et des métadonnées est proposé dans la seconde section de ce chapitre, en vue de faciliter l'acquisition des métadonnées et leur traitement ultérieur en mode automatisé.

6.1 Définition d'un profil de métadonnées pour l'information statistique territoriale

6.1.1 Motivations

Cette recherche a été menée dans le cadre d'un projet Européen, *ESPON 2013 database*, pour l'aménagement du territoire, visant à constituer non pas simplement une base de données, mais un système d'information documentant la qualité des données et offrant des moyens de contrôle de cohérence sur les données. Ce projet, financé par ESPON, l'observatoire européen pour l'aménagement du territoire, doit fonctionner comme un point central de collecte de données hétérogènes (des statistiques territoriales principalement, collectées à différents niveaux géographiques) issues de sources officielles (Eurostat, INSEE, DG-REGIO), mais aussi de groupes de recherche produisant de nouveaux indicateurs statistiques.

Le chapitre 3 a montré l'importance de documenter les données via des métadonnées. En effet, les métadonnées sont des « données qui renseignent sur certaines données et qui permettent leur utilisation pertinente » [Bergeron 92]. Les métadonnées donnent donc à l'utilisateur des éléments pour comprendre si les données sont en adéquation avec ses besoins, en décrivant à la fois le contenu de l'information, sa fiabilité, et sa disponibilité. Elle permettent ainsi d'établir la qualité des données au sens le plus large du terme [Servigne 05]. Ces métadonnées doivent reposer sur un format structuré et partagé, pour autoriser un usage interopérable entre les différents systèmes d'information géographique. Le projet

ESPON 2013 database doit donc choisir un standard pour les métadonnées. Il s'agit également de tenir compte d'impératifs portant à la fois sur les délais de mise en place de l'acquisition des métadonnées, et des contraintes liés aux utilisateurs. En effet, le projet a démarré en même temps que d'autres projets qui ont vocation a produire de nouveaux indicateurs, et qui devront renseigner des métadonnées. Il faut donc que le standard choisi soit compris des utilisateurs et rapidement mis en place.

Dans le domaine de la statistique, le besoin de produire des métadonnées a été éprouvé très tôt [McCarthy 82] et réitéré plusieurs fois : [UN/ECE 95], [Dean 96], [Kent 97]. Pourtant, le seul standard existant pour l'information statistique, le standard SDMX, est encore balbutiant, confère la section 3.4.2 page 92. Loin d'être adopté par tous les usagers de statistiques, il est très critiqué pour les difficultés techniques que son usage soulève.

A contrario, la norme ISO 19115 a été largement étudiée et adoptée par les différents producteurs et usagers de données à références spatiales tels que l'Agence Européenne de l'Environnement [1] et le *Joint Research Center* [2]. Parce qu'elles se présentent le plus souvent comme des ensembles datés de nombres associés à des unités territoriales, les statistiques socio-économiques sont des informations à références spatiales et temporelles qu'il serait légitime de décrire par des métadonnées au standard ISO 19115. L'emploi de cette norme est, de plus, une obligation légale dans le cadre européen, avec la directive INSPIRE [3], pour toutes les données à références spatiales et temporelles. Or, le projet ESPON 2013 database doit également se conformer à la directive européenne INSPIRE, ce qui fait pencher la balance en faveur de la norme ISO 19115.

Cependant, les acteurs du système de collecte de données socio-économiques se tiennent à l'écart de cette norme. Preuve en est l'absence de citation de cette norme dans les comptes-rendus des dernières conférences internationales [4] sur les systèmes de gestion de l'information statistique.

Une des raisons est sans doute le manque d'adéquation de la norme ISO 19115 vis à vis des spécificités de l'information statistique. Parmi elles, la structuration particulière, composite et hétérogène, de l'information statistique est à souligner. Il existe donc une réelle complexité dans le fait d'associer à une telle structure des métadonnées expressives et exploitables.

6.1.2 Structure de l'information statistique territoriale

L'information statistique circule généralement sous la forme de jeux de données, qui regroupent chacun une collection d'indicateurs différents, mesurés chacun de façon spécifique sur un ensemble d'unités territoriales, à différentes dates. Un jeu de données présente de manière schématique trois niveaux d'information différents :

Premier niveau Ce niveau d'information concerne le jeu de données : il renseigne sur son nom, son responsable, son créateur, ses modalités de distribution, et sa maintenance (régularité, fréquence).

Deuxième niveau Le deuxième niveau d'information décrit chaque indicateur. En effet, un indicateur est désigné dans le jeu de données par un code, qui ne suffit pas à sa compréhension : il est aussi nécessaire de lui associer un nom, une description textuelle, l'unité de mesure employée, et une

1. http://www.eea.europa.eu/fr
2. http://ec.europa.eu/dgs/jrc/index.cfm
3. http://inspire.jrc.ec.europa.eu/
4. http://www.unece.org/stats/documents/2007.05.msis.htm, http://www.unece.org/stats/documents/2009.05.msis.htm

classification thématique. La sémantique est aussi donnée par la méthode de mesure (méthodologie) qui est spécifique à chaque indicateur. Un indicateur comme la consommation d'eau par ménage peut-être recueilli par un sondage ou une enquête auprès d'un échantillon représentatif des ménages, puis estimé à partir de cet échantillon, alors que le nombre de naissances par commune est un chiffre enregistré au niveau des mairies, sans estimation.

Troisième niveau Le troisième niveau d'information décrit les valeurs des indicateurs, sur chacune des unités statistiques. Pour un même indicateur, on constate que sa sémantique varie souvent d'un producteur de données à l'autre, et d'une époque à l'autre. Nous illustrons ce problème avec l'indicateur « chômage ». En dépit d'une tentative d'harmonisation européenne symbolisée par le partage d'une définition commune définie par l'Organisation Internationale du Travail, l'Institut National de la Statistique et des Etudes Economiques (INSEE) et Eurostat publient des chiffres différents pour la même unité (la France) : ainsi, le taux de chômage publié par l'INSEE en février 2008 (8,4 %) diffère de celui estimé par Eurostat (8,8 %). Pour les deux instituts, un chômeur est une personne qui n'a pas eu d'activité rémunérée supérieure à une heure pendant une semaine, et qui peut prouver sa recherche d'emploi. Cependant, les méthodes de calcul, de pondération et de correction des chiffres à partir de l'enquête emploi trimestrielle diffèrent entre l'INSEE [5], et Eurostat [6]. L'exemple du chômage illustre aussi l'évolution des méthodes de calcul et de mesure dont sont l'objet les indicateurs. Par exemple, l'INSEE fait évoluer régulièrement sa méthodologie de calcul du chômage [Goux 03], en le justifiant dans des documents accessibles en ligne [7]. Il est donc nécessaire d'accompagner les valeurs des indicateurs d'une information de provenance (le lignage) indiquant la source, la date et la méthode de collecte et de traitement des données, afin de pouvoir les comparer et les interpréter plus justement. Cette information est souvent complexe : elle comporte à la fois des formules de calcul, des définitions, et est le plus souvent compilée dans des documents externes au jeu de données. Le guide de l'OCDE sur la construction des indicateurs composites [OCDE 08] montre la difficulté de résumer ces manipulations (pondérations, réajustements, etc.) en une simple formule. En effet, dès qu'elles impliquent plusieurs indicateurs, il faut s'assurer que les composants sont désignés par un code connu et documenté quelque part, et que cette documentation est accessible à tous.

Enfin, suivant les sources utilisées, le niveau de restriction d'usage sur les données peut varier. Par exemple, des données spécifiques sur l'emploi en Pologne peuvent avoir été collectées par des organismes de recherche qui ne souhaitent pas les diffuser au grand public, mais seulement à un nombre limité d'utilisateurs, alors que généralement ces statistiques sur le reste du territoire européen sont disponibles librement. C'est pourquoi ces contraintes doivent être également associées au niveau des valeurs dans le jeu de données.

Il faut noter également que le lignage des valeurs peut être commun à un ensemble d'indicateurs, en particulier dans le cas de tableaux de contingences. Un tableau de contingence correspond à la désagrégation d'un indicateur (la population par exemple) en fonction de catégories (classes d'âge, catégories d'emploi, etc.) pour former de nouveaux indicateurs. Ces indicateurs (population active en milliers par tranche d'âge et sexe, par exemple, voir tableau 2 page 17, publié par l'INSEE [8]) partagent alors le même lignage. Le code de l'indicateur est alors un pointeur vers une cellule du tableau de contingence, par exemple « actifs_15_64_m » pour population active âgée entre 15 et 65 ans, de sexe masculin (voir

5. http://www.insee.fr/fr/methodes/sources/pdf/eeencontinu.pdf
6. http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-29012010-AP/FR/3-29012010-AP-FR.PDF
7. http://www.insee.fr/fr/methodes/sources/pdf/estimations_chomageBIT_enquete_emploi.pdf
8. http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATCCF03170

tableau 6.1).

Age	Hommes	Femmes	Ensemble
15 ans ou plus	actifs_15_m	actifs_15_f	actifs_15
15-64 ans	actifs_15_64_m	actifs_15_64_f	actifs_15_64
15-24 ans	actifs_15_24_m	actifs_15_24_f	actifs_15_24
25-49 ans	actifs_25_49_m	actifs_25_49_f	actifs_25_49
50-64 ans	actifs_50_64_m	actifs_50_64_f	actifs_50_64
55-64 ans	actifs_55_64_m	actifs_55_64_f	actifs_55_64
65 ans ou plus	actifs_65_m	actifs_65_f	actifs_65

TABLE 6.1 – Exemple de codes d'indicateurs créés pour une table de contingence extraite du site INSEE, « Population active en miliers selon le sexe et l'âge en 2008. »

6.1.3 Etude de la compatibilité avec la norme ISO 19115

Comme expliqué dans la section 3.3.1, la norme ISO 19115 décrit un ensemble d'informations (obligatoires ou facultatives) qui peuvent être associées aux données dites « brutes » et propose une grammaire XML pour la structuration de ces informations. Les informations s'organisent dans différentes rubriques. Elles apparaissent dans le schéma simplifié de la norme, que nous reproduisons dans la figure 6.1. La norme offre la possibilité d'adapter le niveau de détails et de richesse des informations aux besoins des utilisateurs, à condition de conserver les éléments du noyau, qui apparaissent en orange dans la figure 6.1.

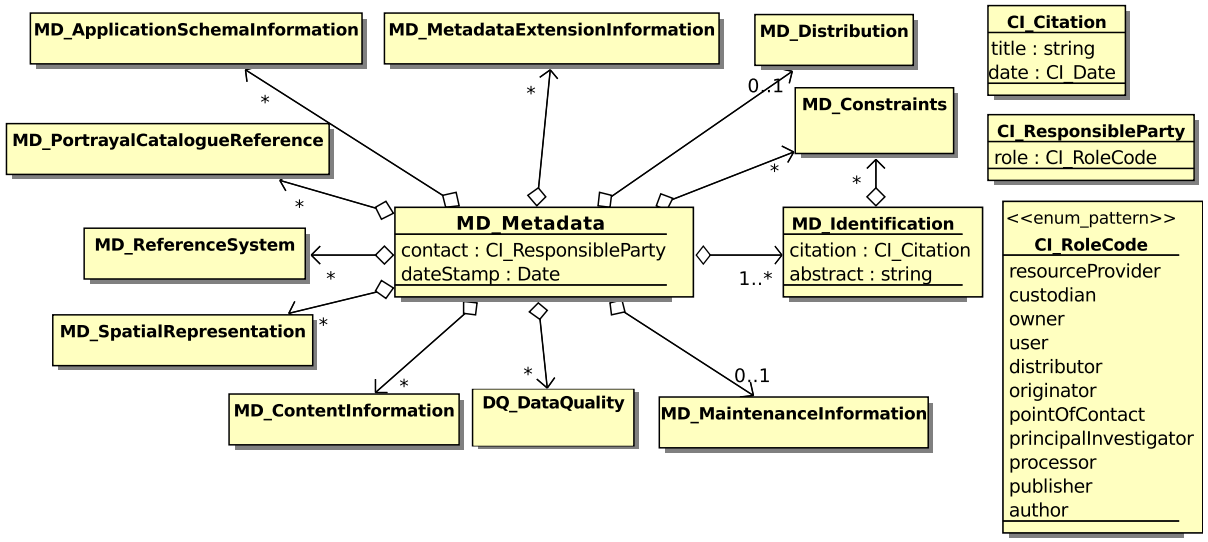


FIGURE 6.1 – Les différentes rubriques de la norme ISO 19115, modélisées d'après le schéma publié sur `http://www.isotc211.org/2005/gmd/`.

Les attributs obligatoires sont ceux concernant la fiche de métadonnées elle-même (MD_Metadata) et l'identification du contenu (MD_Identification). Pour l'identification des données, il faut inclure au minimum un titre ou un pointeur permettant de citer la ressource avec la date de publication (CI_Citation) et un résumé (abstract). Pour décrire la fiche de métadonnées, la norme exige de fournir au moins le responsable (par son rôle) et la date de rédaction de ces métadonnées. Cet ensemble de données peut

être ensuite étendu à volonté dans un *profil*, qui est une extension des éléments de base (par ajout de nouveaux éléments, et/ou spécialisation des éléments existants), mais il doit forcément inclure le noyau. Par rapport à la création d'un profil adapté à l'information statistique, il s'agit avant tout de s'assurer qu'il est possible de spécifier les informations jusqu'au niveau de détail requis.

La norme est le plus souvent utilisée au niveau de MD_Metadata pour décrire des données thématiques rattachées à un support spatial de type quelconque. Ceci s'accompagne d'une simplification extrême de la description thématique, qui présuppose que les données thématiques ont une description commune, et la même provenance, comme l'illustre l'exemple suivant. Les données du *Corine Land Cover*, décrivant la nature d'occupation des sols, disponibles en ligne sur le site de l'Agence Européenne de l'Environnement, sont associées à une fiche de métadonnées MD_Metadata, conforme à la norme ISO 19115, mais qui ne comporte aucune description détaillée des 45 classes d'usage du sol que contient cette base de données. Ceci est une observation qui se vérifie quel que soit le format (vectoriel [9] ou raster [10]) de ce jeu de données. Même si les différentes classes d'usage du sol partagent le même lignage, étant donné qu'elles ont été déduites à partir de traitements de la même information : l'image satellitaire, il serait tout de même nécessaire de décrire ce que signifie chaque classe. En effet, ces catégories thématiques d'occupation du sol varient suivant les sources et les époques. Il est important d'en produire une description suffisamment précise pour permettre de comparer les résultats de ce jeu de données à d'autres jeux de données. C'est une conclusion que tirent les travaux de Comber et Wadsworth relatifs à l'alignement des catégories d'usage du sol [Comber 05, Comber 10, Wadsworth 06]. Enfin, il faut remarquer que cette pratique de simplification de l'information thématique se vérifie pour d'autres jeux de données. Même lorsqu'ils contiennent plusieurs indicateurs ayant des sources différentes, la partie consacrée à l'identification ne détaille pas chaque indicateur mesuré et encore moins la provenance particulière de chaque valeur (suivant la date, la localisation, ou l'indicateur).

Or, pour décrire l'information statistique, il est nécessaire de produire une information thématique détaillée du jeu de données. Cette description est nécessairement portée par l'élément MD_Identification qui, dans la norme, contient aussi les informations de classification. MD_Identification pourrait donc permettre de décrire un indicateur. Cependant, MD_Identification ne porte ni les informations sur l'usage (MD_Constraint), ni les informations de qualité (DQ_DataQuality), qui sont rattachées à MD_Metadata. La description de l'information statistique nécessite de rattacher ces informations à la description de l'indicateur, et, de plus, l'élément MD_Identification de MD_Metadata doit être unique. Pour décrire plusieurs indicateurs, la logique exige donc de multiplier le nombre de fiches MD_Metadata qui contiendront les informations pour un indicateur donné.

Ainsi que la section 3.3.1 page 78 l'a déjà exposé, il apparait que la rubrique MD_ApplicationSchemaInformation, bien que peu utilisée et souvent mal comprise [Barde 05], permet de spécifier des niveaux d'agrégation récursifs (voir figure 6.2). Il est possible de proposer une fiche de métadonnées (MD_Metadata) décrivant un jeu de données (DS_Dataset) qui lui même regroupe plusieurs fiches de métadonnées (MD_Metadata), une pour chaque indicateur. Il est ainsi possible de spécifier des informations d'identification pour chaque indicateur.

L'élément DQ_Quality possède un élément LI_Lineage qui définit les informations de lignage (la traçabilité) permettant de retranscrire les méthodes de mesure des valeurs et leur provenance. L'élément DQ_Quality peut contenir un nombre illimité d'éléments LI_Lineage. Ainsi, il est possible de spécifier autant de lignages que nécessaire : un pour chaque valeur au maximum, ou moins, si on imagine un

9. http ://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2000-clc2000-seamless-vector-database-1
10. http ://www.eea.europa.eu/data-and-maps/data/corine-land-cover-clc1990-100-m-version-12-2009

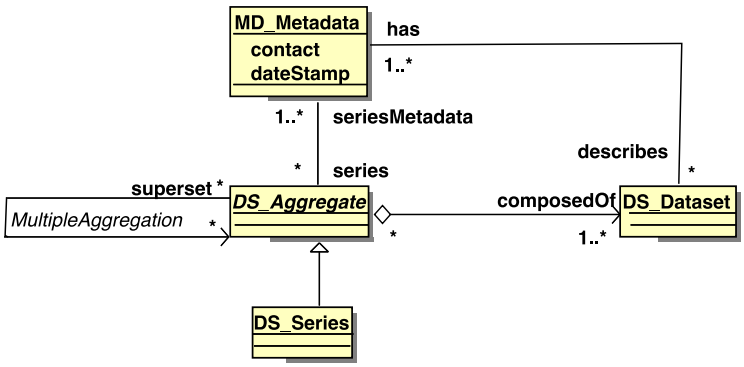


FIGURE 6.2 – Composition simplifiée de la rubrique MD_ApplicationSchemaInformation de la norme ISO 19115, modélisé d'après le schéma publié sur `http://www.isotc211.org/2005/gmd/`.

mécanisme pour regrouper des valeurs ayant la même provenance. Un mécanisme pour définir la portée spatio-temporelle d'un lignage a été prévu dans la norme, à travers l'élément EX_Extent, qui peut être réutilisé. Il est donc possible de préciser la portée spatio-temporelle du lignage défini :

- dans le pire des cas (un élément de lignage différent pour chaque valeur), la portée sera définie par l'unité spatiale considérée et par la date de validité de l'indicateur ;
- sinon, la portée définit une certaine aire géographique et une certaine plage temporelle.

Ainsi, l'étude technique de la norme ISO19115 montre qu'il est possible d'adapter cette norme à l'information statistique qui présente la caractéristique de présenter trois niveaux d'information différents pour un jeu de données. Il s'avère donc que la norme n'est pas souvent utilisée de façon optimale et ses capacités descriptives sont sous-employées (peut-être méconnues aussi).

6.1.4 Création d'un profil de la norme ISO 19115

Le profil *esponMD* [11] que nous proposons définit l'extension de la norme ISO 19115 pour l'information statistique terrritoriale. Il doit tenir compte des spécificités de l'information statistique mais a aussi pour objectif d'être *opérationnel*. Par opérationnel, nous entendons qu'il facilite la saisie des métadonnées autant que possible pour les producteurs de données, en limitant le plus possible le nombre d'éléments à renseigner. Pour les éléments qui ne sont pas renseignés, nous montrons dans la section suivante comment ils sont déduits.

6.1.4.1 Gestion des différents niveaux d'information

Notre profil utilise le mécanisme d'agrégation de fiches décrit dans la figure 6.2, car l'élément DS_Dataset permet de disposer de plusieurs fiches de MD_Metadata, une par indicateur. Comme il est aussi nécessaire de représenter une information commune à l'ensemble du DS_Dataset, celle qui décrit les auteurs des métadonnées (CI_ResponsibleParty), la description générale du jeu de données (MD_Identification), la distribution (MD_Distribution) et la maintenance (MD_Maintenance) éventuelle des données, nous utilisons la racine MD_Metadata.

Ainsi, MD_Metadata décrit le jeu de données, et possède un attribut *describes* de type DS_Dataset.

11. comme *ESPON MetaData*, du nom de l'organisme finançant le projet *ESPON 2013 database*.

L'élément DS_Dataset contient (*has*) plusieurs fiches MD_Metadata décrivant séparément les indicateurs.

6.1.4.2 Adaptation de l'élément MD_Identification pour un indicateur

L'identification d'un indicateur se fait avec ***IndicatorInformation*** qui étend *AbstractMD_Identification*, comme l'illustre la figure 6.3 [12].

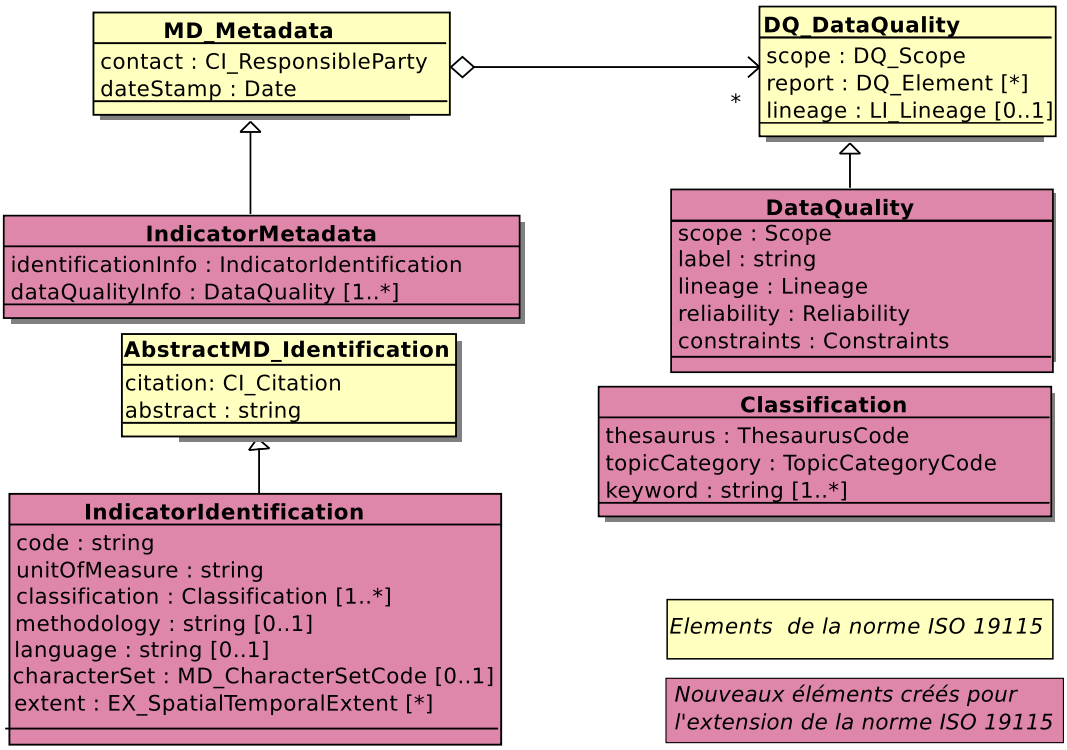


FIGURE 6.3 – Informations redéfinies au niveau indicateur dans l'extension.

Pour l'indicateur, il devient obligatoire de préciser :
- son nom avec *citation*, un champ restreint en longueur,
- un résumé avec *abstract* qui décrit l'indicateur de façon plus complète,
- son code dans le jeu de données avec *code*,
- l'unité de mesure associée avec *unitOfMeasure*,
- un ou plusieurs classifications avec *classification* pour l'indexer par thèmes et mots-clés.

La classification de type ***Classification*** propose à l'utilisateur de choisir un thésaurus parmi ceux référencés par le système d'information, et un des thèmes ou sous-thèmes que propose ce thésaurus, thème auquel il peut associer un à plusieurs mots clés (des chaînes de caractères). De façon optionnelle, la langue et l'encodage des chaînes de caractères peuvent être précisés, tout comme la couverture spatio-temporelle de l'indicateur (dans l'attribut *extent*) et la méthode de calcul de l'indicateur (dans l'attribut *methodology*). En vue de faciliter l'acquisition et le retraitement des métadonnées avec ce profil, il nous semble important d'harmoniser la langue des métadonnées : elles sont toutes saisies en anglais, avec UTF8 comme encodage. C'est pourquoi la langue et l'encodage sont optionnels dans notre profil.

12. Sur cette figure, les éléments originaux de la norme apparaissent en jaune, ceux étendus en rose foncé.

Dans cette extension, le souci de créer le moins de nouveaux éléments est présent. Cependant, les mécanismes d'extension par héritage ne permettent pas de modifier la multiplicité de certains attributs. Or, nous avons besoin de rendre obligatoires les informations essentielles (qui ne le sont pas forcément dans la norme ISO 19115), et, au contraire, de rendre facultatives d'autres informations qui sont imposées.

Par exemple, l'élément MD_DataIdentification de la norme spécialise *AbstractMD_Identification* et propose également des champs pour la langue (mais la langue est obligatoire), l'encodage, la couverture et un attribut *supplementalInformation* qui peut être utilisé pour décrire la méthode de calcul d'un indicateur. Or, nous souhaitons modifier la multiplicité du champ *language* afin qu'il devienne optionnel, ce que le mécanisme d'extension dans les schémas XSD ne permet pas. De plus, au niveau de la classification, les mots-clés se spécifient de façon indépendante des thèmes, et ne sont pas obligatoires. Ainsi, dans *AbstractMD_Identification*, des mots-clés (*descriptiveKeywords*) peuvent être spécifiés, de façon optionnelle, et dans l'élément MD_DataIdentification, des thèmes (*topicCategory*) peuvent être spécifiés de façon optionnelle. Le nouvel élément de classification permet de toujours associer un thème avec un ou plusieurs mots-clés, et de l'associer à l'indicateur de façon obligatoire.

6.1.4.3 Simplification des éléments renseignant sur la qualité

L'utilisateur doit ensuite mentionner des informations renseignant sur la qualité des valeurs associées à l'indicateur, de façon obligatoire, mais simplifiée : l'élément ***DataQuality*** qui étend DQ_DataQuality est prévu à cet effet. Le champ *label* référence des étiquettes placées dans le jeu de données en face de chaque valeur et permet de dispenser l'utilisateur de l'énumération fastidieuse des éléments géographiques concernés par cette qualité (via le champ *scope*), ou bien du calcul de la couverture spatio-temporelle (*extent*) de cette rubrique. L'utilisateur ne doit fournir qu'un ensemble minimal d'informations. La figure 6.4 illustre l'ensemble des informations que porte l'élément ***DataQuality***, et les nouveaux éléments du profil apparaissent en couleur rose foncé.

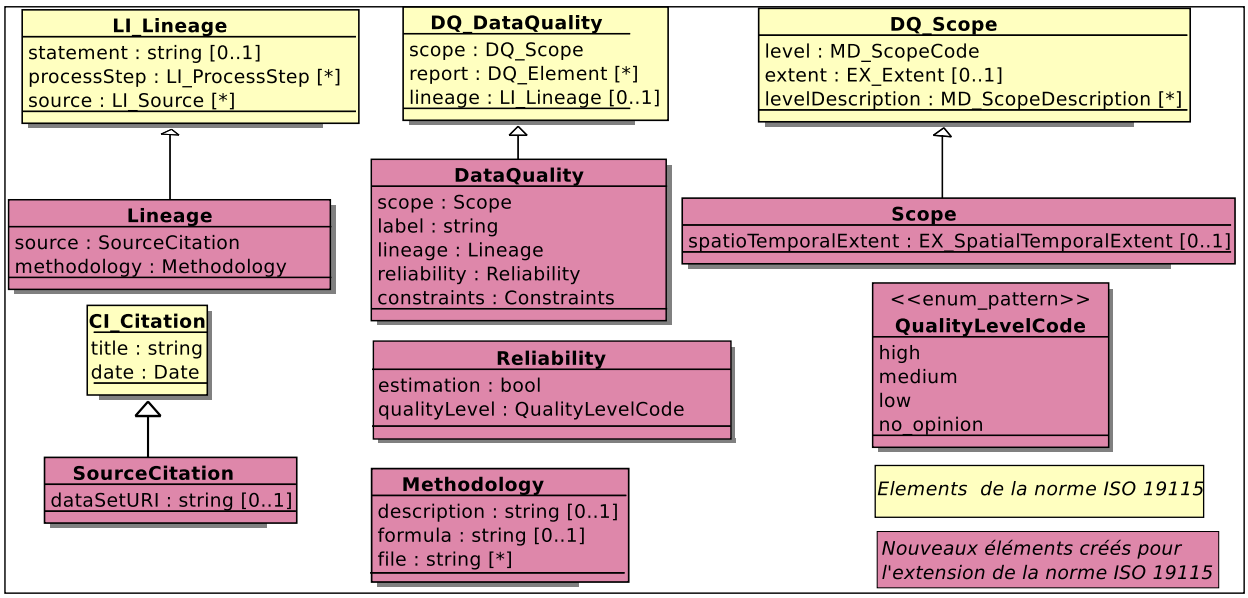


FIGURE 6.4 – Informations redéfinies au niveau indicateur dans l'extension.

Les informations sur la qualité sont largement simplifiées par rapport à l'élément DQ_DataQuality.

Le **rapport de qualité** (*report*), qui donne des éléments complets et factuels concernant l'évaluation de la qualité des données est facultatif dans la norme comme dans l'extension. Cependant, dans l'extension, l'élément Reliability, qui exprime la confiance que l'auteur des métadonnées accorde à ces données, de façon subjective, est obligatoire. Il indique si les données sont issues d'une estimation, (*estimation* vaut vrai), et donne son opinion (*qualityLevel*) sur une échelle de valeur codifiée (QualityLevelCode).

Le **lignage** (*lineage*) des données doit mentionner une source (*source*) et la méthode de mesure/collecte des données (*methodology*). La source précise obligatoirement le nom du fournisseur (CI_Citation.*title*) et la date de récupération des données (CI_Citation.*date*). Le champ *methodology* est un équivalent simplifié de LI_ProcessSteps, qui est présent dans la rubrique DQ_DataQuality de la norme, et permet de décrire les transformations ou les méthodes de calcul des valeurs. Ceci est fait par divers moyens : soit, avec un champ textuel qui en donne une description (*description*), qui peut aussi être une URL, soit avec une formule (*formula*) qui peut s'exprimer dans un langage semi-structuré (comme MathML [13]), soit avec un ensemble de fichiers et documents multimédias que l'utilisateur pourra adjoindre à la fiche de métadonnées simplifiées. Cet expédient, en attendant des moyens automatiques plus efficaces pour traiter l'information, permet au moins de collecter la connaissance, pour les utilisateurs humains qui accèderont à cette documentation, lors de la restitution des métadonnées.

6.1.4.4 Modification des contraintes légales portant sur l'usage et la publication des données

Les données statistiques sont soumises à des contraintes de diffusion de type légales (et non pas de sécurité) qui sont imposées par le producteur des données. Celui-ci rédige un *copyright* (droit de reproduction) concernant l'usage qu'il autorise de ses données, et ce droit doit s'appliquer dans le jeu de données à toutes les données qui sont issues de ce producteur. Mais la section 3.3.1 page 78 montrait que les contraintes légales définies par la norme ISO19115 peuvent se résumer à des codes dont l'interprétation est ambiguë. Or, dans notre cas de figure, il s'agit de déterminer si oui ou non les données sont diffusables au grand public, et si la découverte de l'existence de ces données est même autorisée.

Les contraintes associées à l'usage des données et des métadonnées ont donc été redéfinies pour les rendre plus opérationnelles. Ainsi, au lieu de les définir au niveau du jeu de données ou de chaque indicateur, via l'élément MD_Constraints, elles sont attachées au niveau des valeurs du jeu de données, dans la rubrique ***DataQuality***, et sont obligatoires. Ceci permet de répondre au besoin de définir des droits de diffusion des données, et des *copyrights* spécifiques à des sous-ensembles de valeurs ayant la même provenance. Il était envisageable également de créer plusieurs éléments MD_Constraints définissant plusieurs groupes de droit d'accès et de diffusion des données, rattachés au niveau du jeu de données. Le problème étant d'associer chaque valeur à son groupe, par une étiquette par exemple. En plaçant les contraintes dans l'élément **DataQuality** qui est associé à chaque valeur par une étiquette, nous bénéficions de ce mécanisme d'association, sans avoir à le dupliquer. Le nouvel élément ***Constraints*** est une composition de contraintes sur les métadonnées (***MetadataConstraints***) et sur les données (***DataConstraints***), voir figure 6.5.

Le nouvel élément propose un mécanisme permettant de diffuser des métadonnées, sans les données, ou bien de cacher complètement les métadonnées (et les données ne sont alors pas accessibles non plus). Ce mécanisme existe dans la norme, cependant, c'est à travers un attribut de type booléen que ce droit est précisé, et non plus avec les codes ambigus proposés par la norme. Si l'existence même des données doit être cachée au public, *readRight* vaut faux. Les données peuvent être diffusées librement à tout

13. http://www.w3.org/Math/

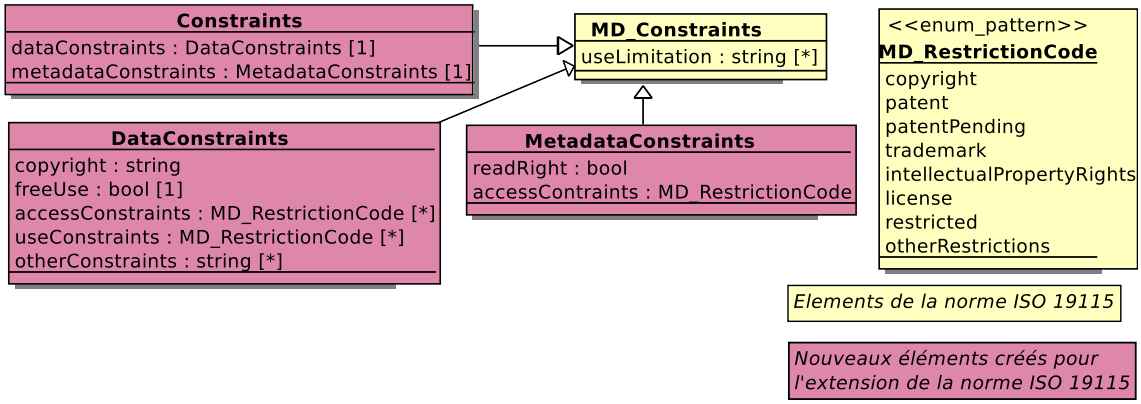


FIGURE 6.5 – Définition de contraintes d'usage pour des valeurs du jeu de données.

public (*freeUse* vaut vrai) ou bien être réservées à une communauté d'utilisateurs enregistrés auprès du distributeur des données (*freeUse* vaut alors faux). Les données sont toujours accompagnées d'un *copyright*, champ textuel équivalent de *useLimitation*, mais obligatoire. L'auteur des métadonnées est encore libre de réutiliser les codes spécifiques pour les mentions légales que propose la norme avec les champs *accessConstraints*, et *useConstraints*.

L'opérationnalisation de la norme ISO 19115 signifie donc la création d'un profil plus simple, mais aussi plus contraignant : l'utilisateur est forcé de renseigner certains champs qui, auparavant, étaient optionnels. En contrepartie, l'information qu'il doit livrer est simplifiée au maximum, et une partie sera calculée après acquisition des données. Techniquement, l'extension se présente sous la forme d'un schéma XSD dans le fichier *esponMDExtension.xsd* (fichier en Annexe, page 259). C'est une grammaire qui peut être utilisée pour structurer et valider un fichier XML conforme à notre profil.

6.2 Proposition d'un flux d'acquisition et de diffusion des données et métadonnées

Le profil proposé doit être utilisé pour l'acquisition des données statistiques dans un système d'information, connecté sur le Web, qui peut ensuite re-diffuser les métadonnées. L'usage des données et métadonnées que nous proposons consiste à les utiliser pour construire un système de *métainformation actif*, au sens où l'entend l'ONU, [UN/ECE 00] [14] :

> « Un système de métainformation actif intègre les données comme les métadonnées dans un même stockage physique, tandis qu'un système de métainformation passif ne contient que les références aux données, non pas les données elles-mêmes. »

Le principal atout d'un tel système d'information (ou métainformation) est de pouvoir répondre à des requêtes qui mélangent des données issues de plusieurs jeux de données différents, et rendent ainsi indépendant l'usage des données du mode d'acquisition des données. En effet, dans un nombre conséquent de systèmes d'information existants, parmi ceux permettant de découvrir les données par les métadonnées, la récupération des données se fait par lots, et ces lots correspondent à ceux intégrés

14. Traduction libre de « *An active metainformation system is physically integrated with the information system containing the data that the metadata in the metainformation system informs about. A passive metainformation system contains only references to data, not the data themselves* »

lors de l'acquisition. Or, pour l'information statistique, cela peut avoir un sens de chercher à récupérer un sous-ensemble de variables pour un territoire donné (la région Rhône-Alpes, par exemple) sur une époque couvrant les dix dernières années d'étude, et de sélectionner ces variables parmi plusieurs jeux de données. Il est ainsi plus aisé de se questionner sur la nature et la cohérence des données, de détecter des phénomènes d'harmonisation propres à chaque jeu de données qui amènent des variables identiques à prendre des valeurs différentes sur la même unité spatiale à la même période.

La figure 6.6 illustre le flot de données idéal pour un système d'information actif connecté au Web. Notre proposition pour ce flux de données est de :

1. faciliter l'acquisition des données et des métadonnées ;
2. structurer les données et les métadonnées dans une même base de données ;
3. diffuser les données dans un ou plusieurs formats de diffusion standards.

Il s'agit de proposer un éditeur de métadonnées en vue de faciliter l'édition d'un fichier de métadonnées conforme à l'extension de la norme ISO 19115 définie. Ensuite, l'acquisition des données dans le système doit se faire par l'analyse et le traitement de la paire de fichiers données/métadonnées dans le système que l'utilisateur dépose par le biais d'un portail Web. Cette opération a pour conséquence le stockage dans une base de données spatio-temporelles de toutes les informations, données et métadonnées. Ces données (dites thématiques) se rattachent à une information spatiale dont le modèle est présenté dans le chapitre 5 page 141. L'interface de requête spatio-temporelle devra respecter les directives INSPIRE et proposer des critères d'interrogation sur le lieu, la date, les acteurs, et la nature des données. Enfin, un format de diffusion automatique basé sur SDMX pour la diffusion de données est proposé en vue de contribuer au développement du partage de données entre systèmes statistiques actifs et distribués.

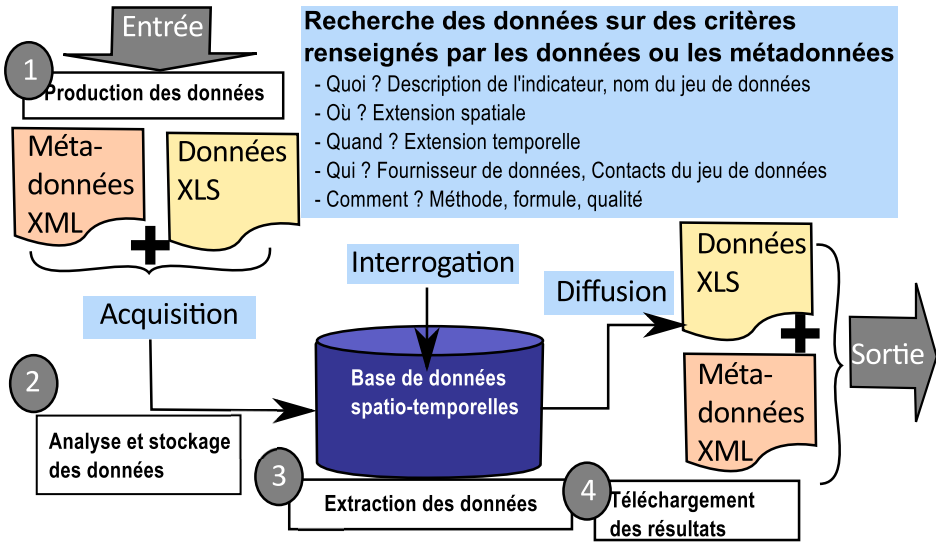


FIGURE 6.6 – Schéma du flot de données.

6.2.1 Structuration des données

Nous souhaitons simplifier au maximum la phase de saisie des métadonnées. Dans le flot de données proposé, le premier document contenant les données peut être analysé afin d'aider à générer des métadonnées dans un document complémentaire, le « fichier de métadonnées », dont une partie devra encore

être complétée par l'auteur des métadonnées (*a priori* le producteur de données). En effet, de précédents travaux ont montré l'intérêt que pouvait représenter l'extraction semi-automatique d'un maximum d'éléments descriptifs des fichiers de données [Manso 04], [Taussi 07], [Laura Díaz 07]. Bien que ces travaux aient été proposés pour des données d'imagerie satellitaires, il est envisageable de découvrir certaines informations par l'analyse des fichiers de données statistiques, si le fichier de données est formaté de façon adéquate.

Notre proposition s'appuie donc sur la définition d'un modèle de document, produit par un tableur tel qu'Excel, pour des données socio-économiques. Ce format s'apparente aux fichiers que les acteurs du domaine statistique s'échangent (voir section A.3.1 page 16) et a été élaboré après de nombreuses discussions avec des acteurs du système d'information (les producteurs de données, et les usagers) en vue de contrarier le moins possible leurs habitudes [Grasland 09]. Le format tabulaire que nous proposons contient une première colonne avec la liste des unités statistiques, codifiées suivant la version de nomenclature utilisée, puis un nombre non limité de colonnes d'indicateurs, représentés par leur code dans la base de données du fournisseur de données, et, au croisement de la ligne et de la colonne, la valeur de l'attribut associé. Ce modèle est présenté dans le tableau 6.2.

TABLE 6.2 – Modèle de document tabulaire proposé pour les données socio-économiques.

	code	level	version	**area**	label	**pop**	label	**gdp**	label
startDate				2003		2004		2005	
endDate				2003		2004		2005	
	LI000	NUTS3	NUTS2003	701,5	1	37,5	3	677	3
	AT112	NUTS3	NUTS2003	1792,6	1	141,7	1	3108,3	1
	FR411	NUTS3	NUTS2003	1471,4	1	97,4	1	1571,9	1
	AT121	NUTS3	NUTS2003	3356,7	1	237,9	1	4585,8	1
	AT12	NUTS2	NUTS2003	3367,1	1	247,8	1	5352,8	1
	DK011	NUTS3	NUTS2006	1230,1	2	143,5	4	4127,7	5
	...	...	...	...	...	...	...	...	...

Ce nouveau modèle formalise certaines informations implicites : la version de la nomenclature (*version*), le niveau d'information pour l'unité géographique codifiée (*level*) sont ajoutés après le code de l'unité spatiale sur chaque ligne. Concernant les indicateurs, les entêtes des colonnes contiennent le code de l'indicateur en gras - dans l'exemple, ce sont « area », « pop » et « gdp » -, puis les deux lignes suivantes indiquent le début (*startDate*) et la fin (*endDate*) de la plage de validité de l'indicateur. On trouve ensuite sur chaque ligne, de façon classique, les valeurs d'indicateurs associées à chaque indicateur, et unité spatiale. Une colonne adjacente à la colonne de chaque indicateur a été ajoutée (*label*) afin de spécifier la provenance de chaque valeur, via une étiquette unique qui référence des informations de lignage qui sont précisées dans le document de métadonnées.

Par exemple, la valeur 1471,4 soulignée dans la table 6.2 correspond à la surface légale [15] (l'indicateur *area*) de l'unité FR411 mesurée en 2003, dans le niveau 3 de la Nomenclature des Unités Territoriales Statistiques (NUTS), version du NUTS datant de 2003, et les informations de provenance, qualité et contraintes de cette valeur sont associées à l'étiquette « 1 ». Cette étiquette pointe sur les informations

15. La surface légale d'une unité est établie par les géomètres et prend en compte des contraintes légales (la surface d'un lac est incluse dans l'unité par exemple). Le calcul d'une surface à partir de sa représentation géométrique ne tient pas compte de ces règles, et sera de plus imparfait, car dépendant du niveau de généralisation des contours de l'unité.

qui sont développées dans le document de métadonnées.

Le « fichier de métadonnées » accompagne le fichier de données, et il comporte les informations descriptives sur les trois niveaux d'information définis, jeu de données, indicateur, et valeur. Dans un souci de simplification de l'acquisition des métadonnées, nous n'exigeons pas de l'utilisateur de remplir certaines rubriques qui sont pourtant très utiles pour la découverte des données par les métadonnées. Par exemple, la couverture spatio-temporelle du jeu de données ou de chaque indicateur (enregistrée dans l'élément EX_Extent), ainsi que le système de référence (enregistré dans l'élément MD_ReferenceSystem) sont à déduire du jeu de données, et non pas à saisir dans les métadonnées. Ainsi, la couverture spatiale correspond à la liste des unités spatiales mentionnée dans le fichier de données, qui sont supposées codifiées et répertoriées, leurs géométries étant connues dans la base de données spatio-temporelles.

Par ailleurs, étant donné qu'un volume conséquent de documentation sur les processus de transformation existe dans des documents textuels et ne peut encore être formalisé (voir section A.1.5), il est suggéré de conserver ces documents textuels en attendant de créer les outils pour le traitement de ces documents. Cet ensemble de documents qui explicitent les procédures de production et de transformation des données sont référencés dans la section lignage par leur nom, dans un champ prévu à cet effet. Il s'agit par la suite de se donner les moyens de conserver, voire de retraiter ces fichiers dans un répertoire ou une base multimédia, en plus des métadonnées structurées qui sont, elles aussi, à conserver pour le traitement ultérieur.

6.2.2 Contrôle de la saisie des métadonnées

L'édition des métadonnées, qui peut être une opération fastidieuse pour un opérateur humain, est également une opération où de nombreuses erreurs ou omissions peuvent intervenir. À partir du profil allégé (on parle aussi de gabarit) que nous avons défini, il s'agit ici de proposer un outil pour l'édition des métadonnées qui facilite leur édition, mais également contrôle la conformité des métadonnées saisies à ce profil. L'objectif est de produire des métadonnées conformes à ce profil, puis de pouvoir exporter ces métadonnées. Si l'éditeur est directement relié au système d'information, l'export pourrait se faire dans la base de données du système d'information. Cependant, dans le cas contraire, il est préférable, pour une meilleure interopérabilité, de proposer un export dans le format XML, basé sur le schéma de la grammaire définie dans notre extension (voir le fichier *esponMDExtension.xsd* en Annexe, page 259). Notre proposition établit le cahier des charges auquel devrait répondre un éditeur de métadonnées performant. Puis nous proposons une analyse des principaux éditeurs connus à ce jour, et nous montrons enfin comment un outil adapté peut faciliter l'édition de métadonnées conformes.

6.2.2.1 Cahier des charges de l'éditeur idéal

De précédents travaux de recherche sur la gestion des métadonnées ont établi un ensemble de fonctionnalités requises ou bien recommandées pour un éditeur de métadonnées, pour des données à caractère plutôt environnemental [Zarazaga-Soria 03], [Wilde 04]. Ces différents travaux plaident d'abord pour une architecture modulaire en vue de faciliter l'intégration des différentes fonctionnalités, offrant ainsi un outil évolutif, adaptable à de nouveaux besoins. Les fonctionnalités décrites sont les suivantes :

– la validation syntaxique automatique de la fiche de métadonnées vis à vis de la grammaire (correspondant au profil de métadonnées) choisie. En effet, en fonction de la norme ou du profil choisi, (Dublin Core, CSDGM, ISO 19115), certains éléments sont ou ne sont pas optionnels ;

- un outil pour l'import ou l'export de fiches de métadonnées au format XML, avec une grammaire se conformant au format choisi : Dublin Core, CSDGM, ISO 19115. Grâce à l'usage de transformations XSL, la présentation des métadonnées dans un format HTML pour être adaptée suivant les préférences de l'utilisateur ;
- une gestion des différents niveaux d'information ;
- un thésaurus, car il aide à classifier et décrire les indicateurs en utilisant des mots-clés définis et reconnus par une communauté scientifique, comme le GEMET pour les ressources environnementales. Il s'agit de proposer un outil pour l'édition et la visualisation du vocabulaire dans une structure hiérarchique (par ordre alphabétique), et l'import ou l'export de champs textuels ;
- un annuaire de contact avec les noms, adresses, mail, et téléphone des personnes susceptibles de servir de contact (comme responsable de jeux de données par exemple) ;
- une aide en ligne pour expliciter les champs à remplir (définition et multiplicité de chaque champ, avec des exemples) ;
- une génération automatique de certaines métadonnées par analyse des fichiers de données. Dans CatMDEdit par exemple, un analyseur a été développé pour les formats de données au format matriciel (ECW, FICC, GeoTiff, GIF/GFW, JPG/JGW, PNG/PGW) ou bien vectoriel (Shapefile, DGN), basé sur les travaux de Manso, [Manso 04], qui permet de récupérer la couverture géographique du jeu de données, l'auteur, la date de création du fichier de données, et parfois aussi des informations plus techniques (le système de référencement spatial, le nombre et le type des objets géométriques, etc.) ;
- un outil pour définir de la couverture spatio-temporelle attachée à un élément de métadonnées. Dans les interfaces actuelles, l'utilisateur peut sélectionner la région géographique via le dessin d'une boîte englobante sur une carte numérique ;
- un contrôle d'accès via la création de rôles, et l'authentification des utilisateurs dans les cas où l'éditeur permet de créer ou supprimer des métadonnées du système d'information ;
- l'internationalisation de l'application, pour traduire l'interface dans toutes les langues.

Relativement aux spécificités de l'information statistique territoriale, et dans la cadre de la création d'un système d'information actif, nous ajoutons les précisions suivantes :

- l'existence d'un lien entre l'éditeur et le système d'information permettrait de lister les indicateurs et les fournisseurs présents dans la base, et pourrait ainsi éviter la saisie d'informations descriptives (nom, code, URL) déjà connues, concentrant l'attention de l'utilisateur sur des informations importantes comme l'unité de mesure et la méthodologie de calcul de l'indicateur, et les transformations des valeurs.
- la définition de la couverture spatio-temporelle pourrait être faite en fournissant le nom ou/et le code des unités spatiales, ou en utilisant une carte territoriale numérique dans laquelle l'utilisateur sélectionnerait les unités concernées par l'élément de métadonnées (par clic de souris, par exemple).
- un contrôle d'accès est facultatif si les métadonnées ne sont pas stockées via l'éditeur. Il ne devient nécessaire que si l'éditeur active l'acquisition des données et métadonnées dans la base de données.
- la génération automatique de certaines métadonnées devrait exploiter le format tabulaire de données que nous avons défini.

6.2.2.2 Etude des éditeurs disponibles

La norme a été implémentée dans un éditeur en ligne [16] développé pour la commission européenne. Au regard des besoins définis pour l'information statistique, cet éditeur propose des champs qui ne sont pas pertinents, et il est impossible de rapporter toute l'information associée aux tables statistiques, car c'est un profil basé sur un seul niveau, celui d'un jeu de données. Par exemple, pour définir la couverture géographique des données (*i.e* l'extension spatiale), seule la sélection d'une boîte englobante est proposée, alors qu'il serait nécessaire de pouvoir sélectionner un ensemble d'unités territoriales sur une carte. De même, les champs associés à la résolution spatiale ne sont pas pertinents car, en statistique, on définit la résolution par le niveau de la nomenclature utilisée, et non pas avec un rapport d'échelle, comme proposé dans l'interface (voir figure 8.6). Cet éditeur ne permet pas non plus d'intégrer notre profil de métadonnées.

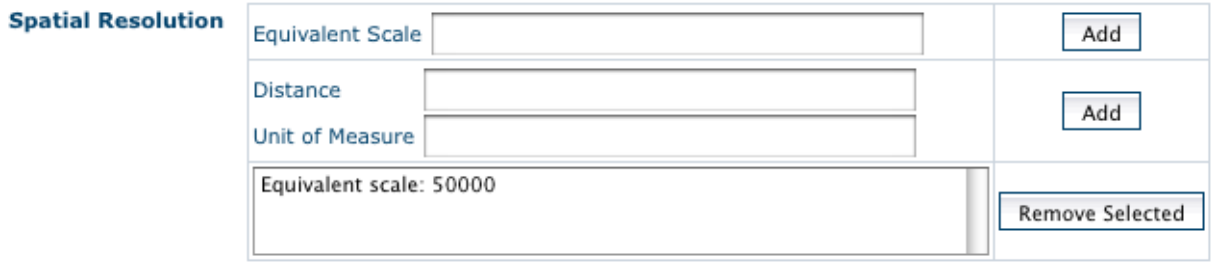


FIGURE 6.7 – Un aperçu de l'interface de l'éditeur INSPIRE, onglet « *Quality&Validity* ».

Il existe, en revanche, une liste conséquente d'outils, commerciaux ou libres pour l'édition, le stockage et la gestion des métadonnées. Les sites suivants en référencent la plus grande partie, et proposent également des comparaisons des outils.

- `http://marinemetadata.org/tools`
- `http://www.fgdc.gov/metadata/iso-metadata-editor-review`
- `http://sco.wisc.edu/wisclinc/metatool`

Nos exigences concernent l'adaptabilité du logiciel et l'ouverture du code, mais aussi l'utilisation de logiciels libres de droits. Ceci écarte d'emblée les logiciels commerciaux ou basés sur l'emploi de systèmes sous licence commerciale. Dans notre étude, les trois éditeurs suivants ont été retenus pour une analyse plus avancée, du fait de leur caractère libre mais également de leur audience large, et parce qu'ils semblaient répondre à une bonne partie de nos exigences (thésaurus et de répertoires de contact intégrés, validation automatique du formulaire de saisie, etc.). Ces éditeurs sont :

- *CatMDEdit*, disponible sur `http://catmdedit.sourceforge.net/`
- *MDWeb*, disponible sur `http://www.mdweb-project.org/`
- *GeoNetwork* disponible sur `http://geonetwork-opensource.org/`

Les avantages qu'offrent ces éditeurs spécialisés ont été présentés dans la littérature [Zarazaga-Soria 03], [Wilde 04], [Barde 05], [Desconnets 07], [Christophle 09]. En dehors de *CatMDEdit*, ces outils sont des plate-formes Web conçues pour le catalogage de données qui proposent également des fonctionnalités de stockage des données. Cependant, ils ne permettent pas d'extraire et de mixer des informations issues de plusieurs jeux de données, car, dans ces systèmes, les données sont conservées dans leur format initial dans un répertoire ou une base de données dédiée, tandis que les métadonnées sont structurées dans une base séparée, et servent essentiellement à indexer les fichiers de données de

16. En ligne sur `http://www.inspire-geoportal.eu/index.cfm/pageid/342`.

façon sémantique. Nous ne nous intéressons donc ici qu'aux possibilités d'édition de métadonnées que ces outils offrent.

Avec les logiciels mentionnés, il n'est pas envisageable de répondre à tous les points énoncés dans le cahier des charges. Par exemple, nous devons renoncer à l'analyse de notre format de données pour produire une fiche de métadonnées pré-remplie. Egalement, cette solution ne permet pas non plus forcément de disposer d'une connexion sur la base de données proposant la liste des indicateurs déjà présents dans le système d'information. Toutefois, il est tout de même envisageable de produire une fiche de métadonnées conforme qui pourrait être intégrée au système d'information.

CatMDEdit est un outil très complet qui, par ailleurs, a été développé à partir d'un cahier des charges qui correspond à nos besoins [Zarazaga-Soria 03]. Cependant, il fonctionne comme un client java qui doit être déployé chez les potentiels utilisateurs. Afin d'éviter le déploiement et la maintenance d'un logiciel sur un ensemble de clients potentiellement hétérogènes, un éditeur déployé sur un unique serveur accessible depuis une interface Web semblait préférable. *MDweb* est un outil de catalogage et de moissonage de l'information environnementale. La nouvelle architecture que propose *MDWeb*, dans sa version v2.2.2, semble beaucoup mieux convenir à l'intégration d'un profil que celle de la version v1.6.0. que nous avions testé. Cela semblait plus facile avec *GeoNetwork* d'après sa documentation. C'est sur ce dernier critère que *GeoNetwork* a été retenu pour expérimenter l'intégration du profil de la norme ISO 19115 que nous proposons.

Ainsi, dans le cadre du projet *ESPON 2013 database*, il a été développé une adaptation de *GeoNetwork* (v2.4.2) pour l'édition des métadonnées [Grasland 10b, Plumejeaud 10], entre Septembre 2009 et Février 2010. L'adaptation de *GeoNetwork* nécessite l'ajout du profil de métadonnées (c'est-à-dire du schéma XSD correspondant) à l'ensemble des profils déjà supportés. Il s'agit en théorie de suivre une procédure documentée et de manipuler des feuilles de styles XSL s'appliquant aux différents éléments de la grammaire, pour obtenir la présentation désirée de modèle pré-rempli (ou *template* en anglais) de métadonnées, et sa validation automatique. Un nombre illimité de fiches peuvent ensuite être instanciées à partir de ce modèle, exportés ou sauvegardées dans un SGBD (McKoi [17] dans la version v2.4.2). Il est possible d'intégrer le profil *esponMD* pour créer notre modèle [18], voir la figure 6.8, mais il apparaît que le logiciel supporte assez mal les adaptations introduites.

Malgré des modifications de son code, le comportement de l'édition des métadonnées reste un peu erratique. En particulier, la sauvegarde de la fiche de métadonnées entraîne la perte de certaines informations, et la recherche des fiches de métadonnées sauvegardées pose problème. Ce problème d'instabilité devrait être résolu *a priori* par l'analyse et la réparation du code défectueux. Cependant, ce code complexe est assez mal documenté : il n'existe aucun document d'architecture. Notre analyse soulève, de plus, un problème fondamental dans l'architecture de la plate-forme : il semblerait que le *template* et les fiches de métadonnées soient régulièrement confondus. Ainsi, la sauvegarde d'une fiche de métadonnées peut détruire le *template*. Or, la résolution de ce problème implique une refonte d'une partie du système. Il n'existe donc plus de valeur ajoutée à utiliser GeoNetwork.

Par conséquent, notre étude conclut à la nécessité de développer un éditeur Web léger, en lien avec la base de données, et capable de reconnaitre le format tabulaire des données pour les analyser et pré-remplir les fiches de métadonnées. Ces conclusions sont également celles d'autres projets de recherche, [Kazakos 03], [Wilde 04], qui, pour des profils et des formats particuliers, ont été conduits à développer

17. `http://www.mckoi.com/index.html`

18. Cette tâche n'est pas aussi simple que le laisse entendre la théorie, et elle implique une modification du code de l'application afin de pouvoir, par exemple, mettre en oeuvre des règles de contrôle et de mises en forme adaptées à notre profil.

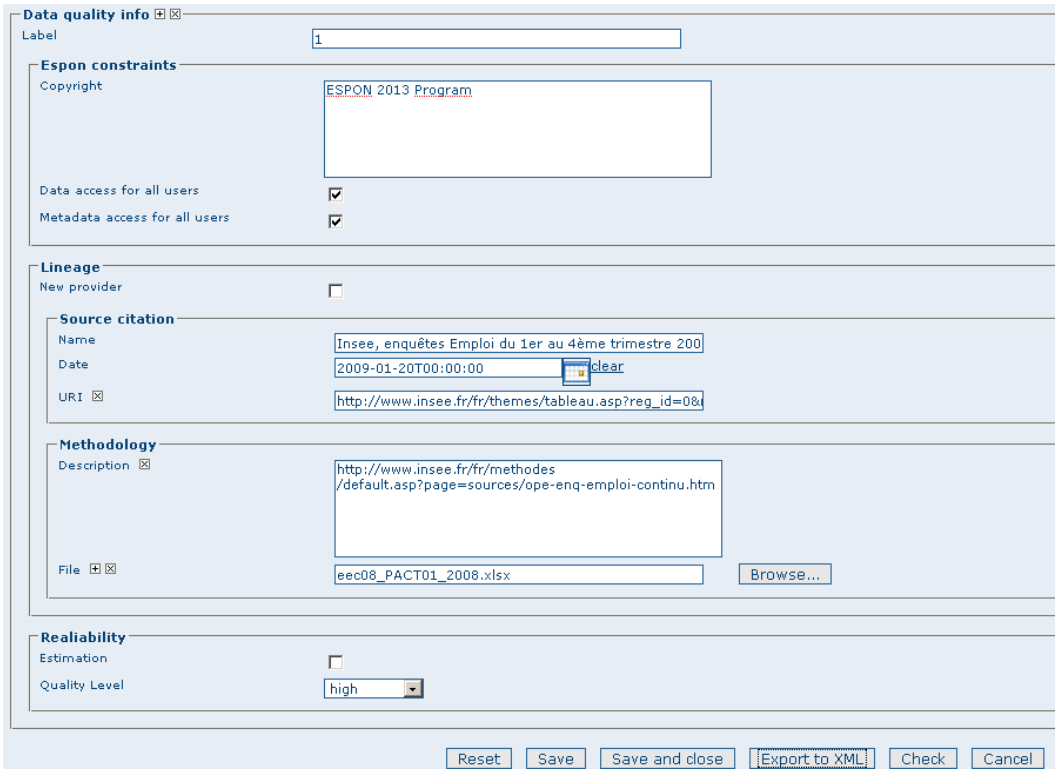


FIGURE 6.8 – Interface Web de *GeoNetwork* intégrant le profil *esponDB*.

un éditeur de métadonnées « sur mesure ».

6.2.2.3 Un éditeur dédié au profil *esponMD*

Ce paragraphe décrit l'intégration du profil *esponMD* dans une plate-forme Web entièrement réalisée par le projet *ESPON 2013 database* [Grasland 10c], avec Anton Telechev comme développeur principal de la partie dédiée à l'édition de métadonnées. Reposant sur la technologie Ajax avec l'usage de la bibliothèque javascript JQuery [19], l'éditeur de métadonnées utilise le profil comme modèle de structuration de l'information éditée. À chaque rubrique correspond un objet, dont la vue est intégrée dans le *Document Object Model* (DOM) de la page HTML. La présentation du formulaire est adaptée à la structure d'un jeu de données : elle comprend 4 onglets, le premier (« Dataset ») pour les informations sur le jeu de données, figure 6.9, le second (« Contact ») pour décrire les contacts associés à ce jeu de données dans les différents rôles qui peuvent leur être dévolus, figure 6.10, le troisième (« Indicator ») pour renseigner les indicateurs, figure 6.11, et le dernier onglet (« Value ») pour les groupes de valeurs (auxquels seront associées les informations de l'élément ***DataQuality***), figure 6.12.

Du fait de la multiplicité des instances d'indicateurs, de contacts et des groupes de valeurs, le formulaire d'édition des métadonnées peut devenir extrêmement long et pénible à lire. L'idéal est de visualiser une instance unique de chaque objet à la fois dans le formulaire (que ce soit un contact, un indicateur,

19. `http://jquery.com/`

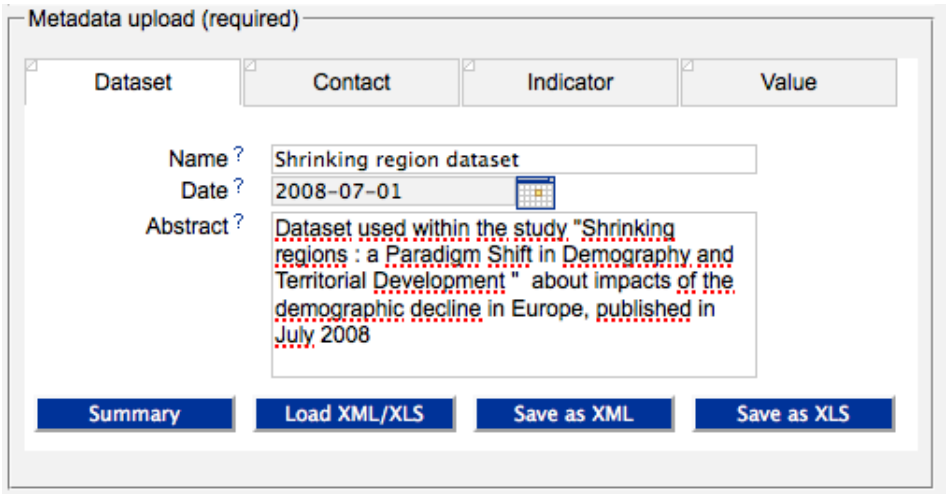


FIGURE 6.9 – Présentation de l'éditeur de métadonnées ESPON, vue du jeu de données.

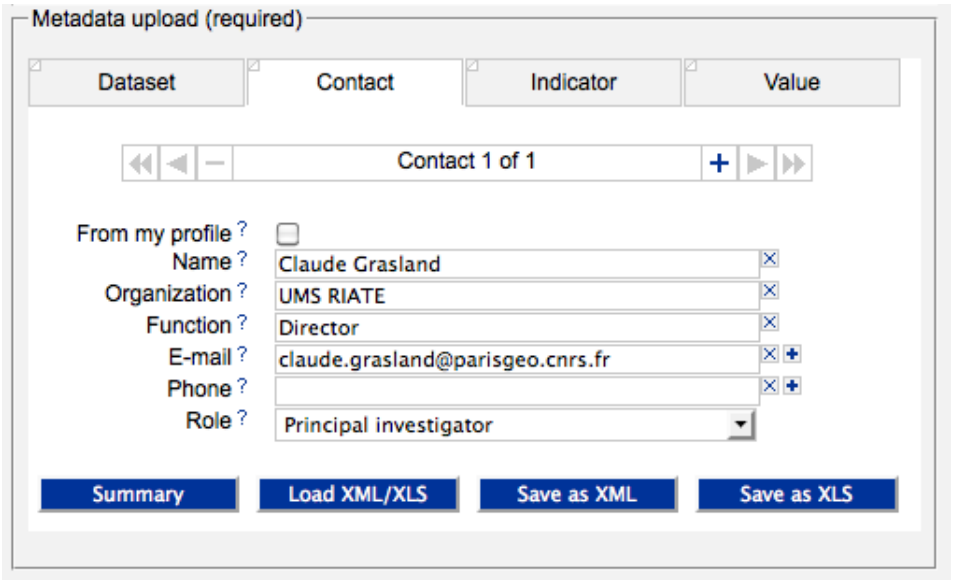


FIGURE 6.10 – Présentation de l'éditeur de métadonnées ESPON, vue sur un contact.

ou bien les informations de qualité liés à un groupe de valeurs). Ainsi, un système d'ajout (signe "+"), suppression (signe "-") d'instance et navigation par des flèches en arrière ("«", "<") ou en avant (">", "»") dans les listes d'instances d'un même type d'objet est intégré aux onglets « Contact », « Indicator », et « Value » afin de gérer la multiplicité de leurs instances, limitant ainsi la longueur du formulaire visualisé pour une plus grande visibilité. Etant donné que des champs peuvent ou doivent (dans le cas de thésaurus) être pré-remplis à partir d'informations contenues dans la base, comme la liste des contacts, des indicateurs ou des fournisseurs de données, ces informations sont chargées depuis la base de données spatio-temporelles à chaque édition de fiche de métadonnées. Les points d'interrogation " ?" sous forme d'hyperlien signalent l'accès à l'aide en ligne sur chaque champ de la fiche. En vue de faciliter l'acquisition et le retraitement des métadonnées avec ce profil, il a été imposé que toutes les métadonnées soient saisies en anglais, avec l'encodage UTF8.

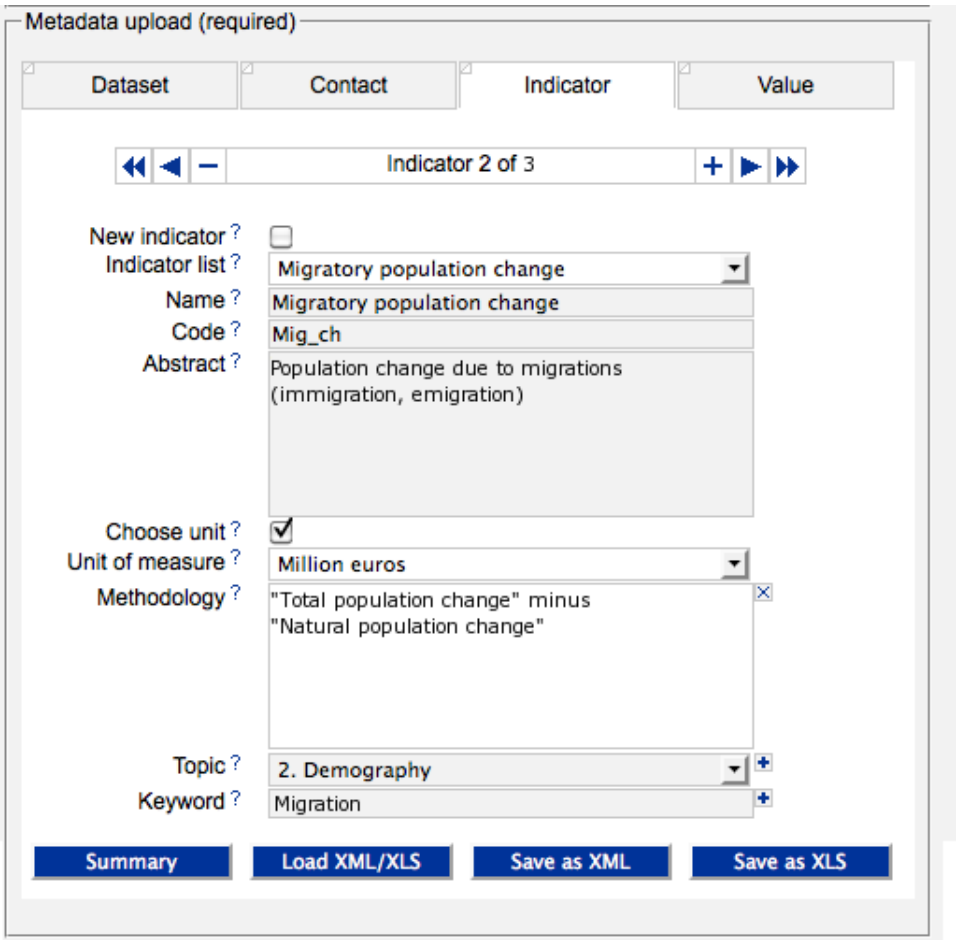


FIGURE 6.11 – Présentation de l'éditeur de métadonnées ESPON, vue sur un indicateur.

Plusieurs opérations sont possibles pour l'utilisateur :
- avec "*summary*", il visualise un résumé de son édition sur une seule page, dans une présentation textuelle lisible ;
- "*load XML/XSL*", il donne à analyser une fiche de métadonnées existante dans un format tabulaire ou XML, en vue de son édition ;
- "*save as XML*", il sauvegarde sur son poste une fiche de métadonnées dans un format XML ;
- "*save as XSL*", il sauvegarde sur son poste une fiche de métadonnées dans un format tabulaire XSL (compatible avec Excel) ;

Les opérations de sauvegarde ne sont possibles que si toutes les informations obligatoires sont présentes (il y a un contrôle de conformité au profil). Lorsque des champs obligatoires sont manquants, ils sont entourés de rouge. L'extrait de code 8.1 en annexe page 269 montre un extrait du fichier XML généré correspondant à l'élément édité dans l'onglet « Value » et conforme au profil *esponMD*.

Accessible en ligne depuis la plate-forme scientifique ESPON, un utilisateur souhaitant produire des métadonnées peut agir selon deux scénarios. Dans le premier scénario, l'utilisateur entre directement sur la page d'édition des métadonnées, et remplit tous les champs, comme nous venons de le décrire, avant d'exporter dans l'un des deux formats (XML ou XLS) la fiche de métadonnées. Dans le second scénario, qui devrait être bientôt implémenté, il télécharge ses données dans le format tabulaire préconisé pour qu'elles soient analysées par l'éditeur de métadonnées en vue de préparer le formulaire de métadonnées.

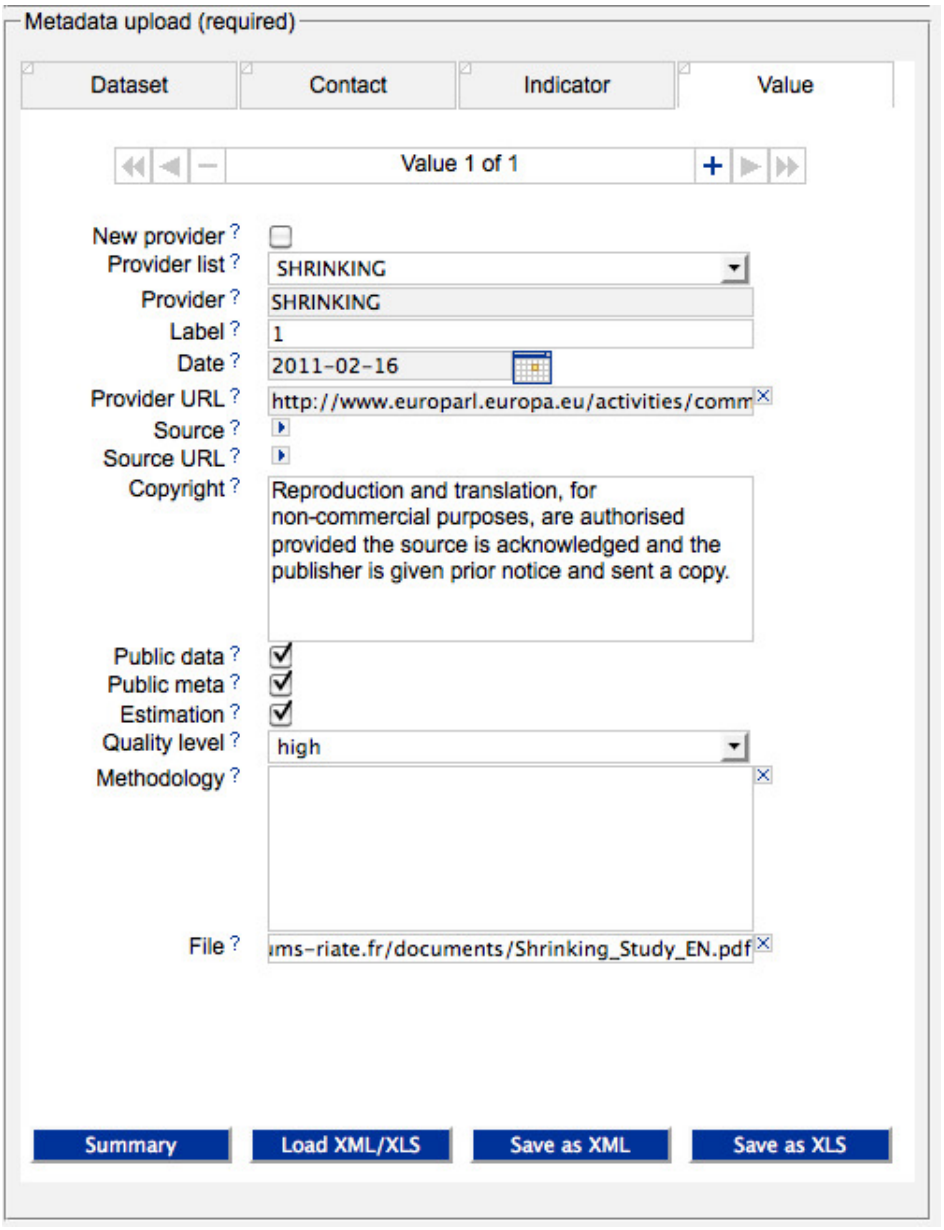


FIGURE 6.12 – Présentation de l'éditeur de métadonnées ESPON, vue sur un groupe de valeurs.

Ainsi, les étiquettes associées aux groupes de valeur sont pré-remplies, tout comme les informations liées aux indicateurs lorsque les codes utilisés dans les jeux de données ont été reconnus. Dans ces deux scénarios, les métadonnées visualisées correspondent toujours à un ensemble minimal d'informations afin de faciliter l'édition et la vérification par l'utilisateur. Également, il n'existe aucun accès en écriture dans le système d'information, et l'édition des métadonnées n'est associée à aucun enregistrement des données dans le système. Ceci permet un couplage faible de l'éditeur avec le reste du système, et augmente la ré-utilisabilité de ce module [20].

20. Le code est la propriété d'ESPON et du groupe de recherche HyperCarte.

6.2.3 Stockage conjoint des données et des métadonnées

Après avoir exposé comment produire des métadonnées conformes au profil *esponMD* de la norme ISO 19115, nous montrons comment ces métadonnées pourraient être structurées dans une base de données spatio-temporelles en vue de fabriquer un système d'information actif, où notamment, les métadonnées seraient exploitées pour répondre à des requêtes fusionnant des sous-ensemble de données. La figure 6.13 présente dans un formalisme UML le modèle associé aux métadonnées dans la base de données. Des informations spatiales sont associées aux unités géographiques (*GeographicUnit*) : elles n'apparaissent pas dans ce schéma, mais sont décrites dans le chapitre 5 page 150. Le schéma proposé sert à structurer les informations de la base de données du projet ESPON 2013 database [Plumejeaud 10, Grasland 10c].

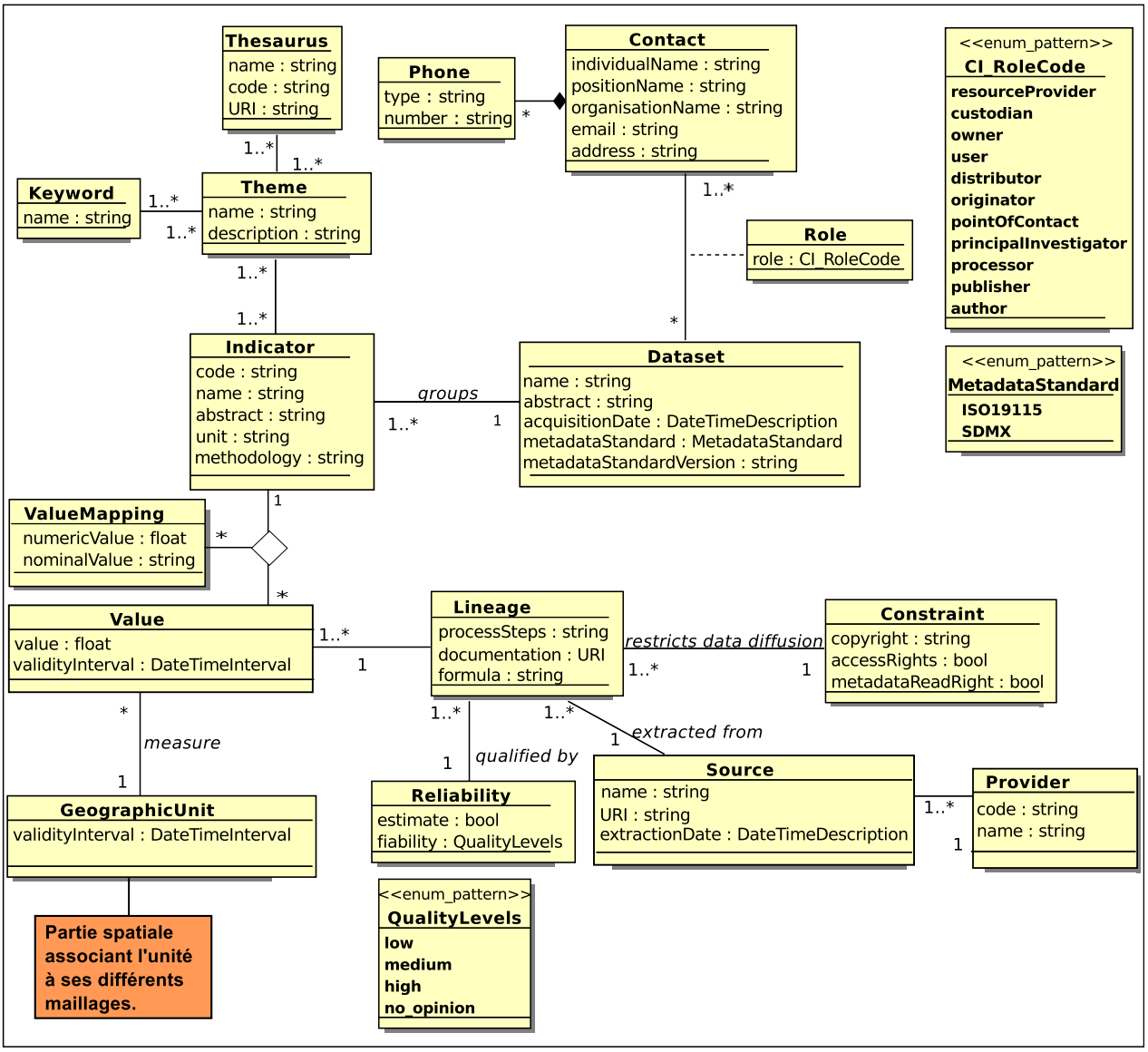


FIGURE 6.13 – Modèle de classes UML structurant les données et métadonnées.

6.2.3.1 Le niveau des valeurs

Le modèle décrit de façon la plus complète possible les valeurs (Value) qui sont associées aux unités géographiques (*GeographicUnit*). La valeur précisée avec *Value* est un numérique de type réel (*value*) auquel est associée une période ou date de validité « officielle » (attribut *validityInterval*). Chaque valeur est associée à un indicateur (*Indicator*) et une unité géographique (*GeographicUnit*). Pour les indicateurs de type qualitatif nominal, comme par exemple la typologie urbain/rural des unités spatiales que publie le projet de recherche européen TIPTAP [Camagni 10], des fonctions de conversion entre valeurs numériques et ces valeurs qualitatives nominales sont prévues. L'association entre valeurs nominales et leurs traductions numériques est conservée dans la classe *ValueMapping* associée à l'indicateur et l'ensemble des valeurs qu'elle décrit.

Chaque valeur possède une provenance (ou lignage), *Lineage*, et cette provenance peut s'appliquer à un ensemble de valeurs (au moins une). Par exemple, les valeurs associées à l'ensemble des unités statistiques de niveau départemental en France, pour le jeu de données « données Eurostat » et l'indicateur « chômeurs » peuvent avoir une même provenance. Cette provenance est qualifiée par un niveau de confiance (*Reliability*) renseignant sur le fait que la donnée soit une estimation, ou bien un chiffre donné pour exact, et par un niveau de fiabilité (*fiability*) donné par un type énuméré (*QualityLevels*) à quatre valeurs :

- *low* pour une confiance faible dans la fiabilité de cette donnée,
- *medium* pour une confiance moyenne dans la fiabilité de cette donnée,
- *high* pour une confiance élevée dans la fiabilité de cette donnée,
- *no_opinion* pour ceux qui ne souhaitent pas ou ne peuvent pas émettre d'opinion.

La provenance de la valeur permet aussi de documenter le processus de transformation spécifique à la valeur (*processSteps*), la formule de calcul de cette valeur (*formula*), et enfin fournit une référence vers un document décrivant le processus de constitution de la valeur (documentation). L'attribut *processSteps* peut préciser des informations simples, sous forme textuelle, comme par exemple indiquer que la valeur est issue de l'agrégation de NUTS de niveau(x) inférieur(s). Le champ *formula* est une chaîne de caractères qui pourra être formatée selon un schéma XML (comme MathML [21] ou une extension) dans de futurs développements.

Le lignage de la valeur indique aussi la source des données (*Source*), connue par son nom (*name*), mais aussi la date d'extraction (*extractionDate*). La source est publiée par un unique fournisseur de données (*Provider*), soit dans des fichiers, soit en ligne sur un site Web, décrits par des Unique Resource Identifier (*URI*). Le fournisseur est décrit par son nom (*name*) et son URI (*URI*).

Le lignage informe également sur les contraintes de diffusion de la valeur (*Constraint*). *Constraint* donne le texte du *copyright* associé aux données, qui pourra être montré dans une interface de diffusion des données aux utilisateurs, mais qui possède aussi deux champs booléens indiquant si la donnée est en accès libre au grand public (*accessRights* vaut alors vrai), et si l'existence de cette donnée peut être mentionnée au grand public (*metadataReadRight* vaut alors vrai). Par exemple, dans le cas de données confidentielles, *metadataReadRight* vaut faux, et obligatoirement, du fait de la directive INSPIRE, *accessRights* vaut alors faux. On peut, par contre, stocker des données non confidentielles (*metadataReadRight* vaut vrai) mais ayant un accès restreint (*accessRights* vaut faux). Cette contrainte, placée au niveau des valeurs de chaque unité géographique, permet de définir différents niveaux de diffusion pour un jeu de données et même pour un certain indicateur collecté sur une région géographique. Certains sous-ensembles géographiques peuvent avoir été renseignés d'après des sources qui ne souhaitent pas

21. `http://www.w3.org/TR/MathML3/`

divulguer leurs données au grand public, et, par le biais d'une contrainte en accès restreint, l'existence de ces données peut être connue, mais les utilisateurs non autorisés doivent s'adresser directement à la source pour obtenir les données.

6.2.3.2 Le niveau des indicateurs

L'indicateur (*Indicator*) est décrit par son nom (*name*) et son code (*code*), un résumé sur le sens de l'indicateur (*abstract*), une unité de mesure (*unit*), une méthode de mesure et/ou de calcul (*methodology*) des mots clés (*Keyword*) et des thèmes (*Themes*). Chaque thème est associé à au moins un thésaurus (*Thesaurus*), et ce thésaurus, décrit par son nom (*nom*) et code (*code*) est publié sur un support référencé par son URI (*URI*).

Deux jeux de données peuvent contenir le même indicateur : il aurait donc été légitime d'associer un unique indicateur avec l'ensemble des jeux de données où il peut être présent. Cependant, dans chacun de ces jeux de données, un « même » indicateur peut avoir été mesuré suivant une méthodologie différente, et dans une unité de mesure différente. Par ailleurs, ni la comparaison des codes (qui ne sont pas standardisés à un niveau international, ni exempts de fautes de frappe), ni la comparaison des noms (où des coquilles peuvent se glisser), ni celle des résumés, ne garantit d'identifier la même instance d'indicateur entre deux jeux de données différents. Il surgit là un problème d'appariement sémantique que nous n'avons pas résolu à ce niveau. Ainsi, par prudence, nous préférons proposer de stocker toutes les instances d'indicateurs, sans qu'elles ne soient harmonisées. En revanche, le traitement de ces informations, comme expliqué dans les perspectives, peut permettre de déterminer si deux instances d'indicateurs sont équivalentes et de construire ainsi une ontologie spatio-temporelle des indicateurs.

6.2.3.3 Le niveau du jeu de données

Le jeu de données est représenté par la classe *Dataset*, dont les attributs sont le nom (*name*), le résumé (*abstract*) et la date d'acquisition (*acquisitionDate*) dans le système. Ces métadonnées sont décrites suivant une norme (*metadataStandard*) et une version de la norme (*metadataStandardVersion*), et l'ensemble du jeu de données est importé dans la base de données à la date *acquisitionDate*.

Le jeu de données possède un ou plusieurs correspondants renseignés dans *Contact* dont le rôle est précisé dans *Role*. Ce contact correspond au *CI_ResponsibleParty* de la norme ISO 19115. Il peut être renseigné par le nom d'une personne (*individualName*) ou le nom de l'organisation (*organisationName*), ainsi que sa fonction dans l'organisation (*positionName*) dans laquelle il travaille pour produire ce jeu de données. Il possède une adresse de courrier électronique (*email*) et une adresse postale (*address*). Ce contact peut être associé à zéro ou plusieurs numéros de téléphone, (*Phone*), dont on précise le type (*mobile, phone, fax*).

6.2.3.4 Exploitation du modèle

Pour les organismes en charge de collecte *et de diffusion* de données, dont le projet *ESPON 2013 database* fait partie, il s'agit d'abord de se conformer à la directive INSPIRE [22], qui s'applique à l'in-

22. `http://inspire.jrc.ec.europa.eu/`

formation statistique territoriale et exige *la découverte des données spatiales par les métadonnées*. Ceci signifie que tout utilisateur doit pouvoir prendre connaissance des éléments descriptifs du jeu de données (des indicateurs et même des valeurs) avant de se voir proposer l'envoi des données qui l'intéresse. Lorsque cette récupération des données a lieu, les données doivent à ce moment-là être accompagnées de leurs métadonnées.

Plus spécifiquement, INSPIRE indique les différents types de requêtes auxquels un système d'information publiant des données à références spatiales doit désormais être en mesure de répondre - plus précisément le service en charge de la publication des données sur le Web (*discovery service*). Ce service doit permettre de rechercher les données géographiques selon les critères suivants :

- les mot clés ;
- la classification des données ;
- la qualité et la validité des jeux de données [23] ;
- le degré de conformité aux règles d'implémentation fournies par INSPIRE ;
- la localisation géographique ;
- les conditions légales d'accès et d'usage des jeux de données ;
- les correspondants connus du jeu de données (qu'ils soient responsables de la création, gestion, maintenance, ou de la distribution des jeux de données).

Avec le modèle que nous venons de décrire, nous sommes en mesure de répondre à de telles requêtes, mais en *fusionnant des sous-ensembles de jeux de données*, ce qui est moins classique, et en y associant les métadonnées correspondantes. En effet, ce schéma peut se traduire par le modèle relationnel suivant (en souligné apparaissent les clés primaires, en italique les clés étrangères) :

- Value(valueId, value, validityInterval, *lineageId*, *indicatorId*, *guId*)
- Lineage(lineageId, processSteps, documentation, formula, *reliabilityId*, *sourceId*, *constraintId*)
- Reliability(reliabilityId, estimate, fiability)
- Source (sourceId, name, URI, extractionDate, *providerId*)
- Provider (providerId, name, providerURI)
- Constraint (constraintId, copyright, accessRights, metadataReadRight)
- Indicator(indicateurId, code, name, abstract, unit, methodology, *datasetId*)
- ValueMapping(mappingId, numericValue, nominalValue)
- ValueMappingIndicator(*mappingId, valueId, indicatorId*)
- Theme(themeId, name, description)
- Keyword(keywordId, name)
- Thesaurus(thesaurusId, name, code, URI)
- ThemeThesaurus(*themeId, thesaurusId*)
- ThemeKeyword(*themeId, keywordId*)
- Dataset(datasetId, name, abstract, acquisitionDate, metadataStandard, metadataStandardVersion)
- Contact(contactId, individualName, positionName, organisationName, email, address)
- Role(*contactId, datasetId, role*)
- Phone(phoneId, type, number, *contactId*)

23. La notion de qualité est hautement subjective, car elle correspond à l'adéquation entre la vue que donne un système d'information d'une réalité, et la réalité perçue par les utilisateurs [Wand 96]. La satisfaction de ce critère n'est donc pas évidente : il s'agira avant tout de donner les moyens aux utilisateurs de comprendre si les données proposées sont en adéquation avec leurs besoins.

Par exemple, on peut souhaiter récupérer l'ensemble des noms, codes, et résumés d'indicateurs tels que la personne associée aux données dans le rôle « originator » soit par exemple le projet TIPTAP. Le résultat est donné par la requête 6.2 qui utilise le résultat de la requête 6.1

$$T1 \leftarrow \pi_{\text{datasetId}}(\sigma_{\text{organisationName='TIPTAP'}}(Contact \underset{\text{contactId=contactId}}{\bowtie} (\sigma_{\text{role='originator'}} Role))) \quad (6.1)$$

$$T2 \leftarrow \pi_{\text{name, code, abstract}}(T1 \underset{\text{datasetId=datasetId}}{\bowtie} Indicator) \quad (6.2)$$

La directive INSPIRE impose, entre autres, de pouvoir interroger tout système d'information statistique selon des critères *spatiaux* de localisation. En répondant à des requêtes spatio-*temporelles*, donc en incluant des critères temporels dans la requête, nous pensons anticiper les évolutions futures de la norme, en exploitant le modèle spatio-temporel proposé dans le chapitre 5 page 141. Par exemple, si on souhaite récupérer l'ensemble des valeurs pour les indicateurs de chômage couvrant une période couvrant 1995 à 2005 et un espace concernant la France, à partir du niveau départemental, il faut :

- trouver la liste des identifiants des indicateurs se rapportant au chômage [24] : $I = \{I_1, I_2, \ldots, I_p\}$
- sélectionner toutes les unités géographiques correspondant à la France entre 1995 et 2005, de niveau 3 et plus : en tout 127 unités, soit 100 départements, 26 régions, 1 état : $G = \{u_1, u_2, \ldots, u_n\}$
- pour les unités de G, récupérer les valeurs dont la plage de validité est incluse dans l'intervalle [1995, 2005], et qui sont associées à l'un des indicateurs de la liste I, comme présenté dans la requête 6.3.

$$\pi_{\text{value}}(Value \underset{\text{indicateurId=indicatorId}}{\bowtie} I) \cap \pi_{value}(\sigma_{\text{[1995, 2005]}\supseteq\text{validityInterval}}(Value \underset{\text{geographicId=guId}}{\bowtie} G)) \quad (6.3)$$

Le projet *ESPON 2013 database* a implémenté une interface de requêtes sur une telle base de données, [Grasland 10c], accessible sur `http://database.espon.eu/data`, qui respecte complètement la directive INSPIRE, et qui est basée sur le modèle de données que nous venons de décrire. C'est un système d'information dit « actif » qui permet de fusionner des sous-ensembles de données issus de jeux de données (les lots d'acquisition) différents. Il reste toutefois à définir des formats de diffusion pour ces informations, qui soient intéropérables et peut-être plus adaptés au traitement automatique de l'information que des fichiers dans un format tabulaire propriétaire tel qu'Excel.

6.2.4 Diffusion des données et des métadonnées via SDMX

Par ailleurs, nous souhaitons proposer la dissémination des données dans un mode interopérable, avec l'ouverture d'un service électronique de diffusion de données dans un format adapté au traitement automatique de l'information, et à la découverte des données par les métadonnées. Il apparaît que la fourniture de données statistiques basées sur le modèle que propose SDMX serait en complète adéquation avec INSPIRE. En effet, SDMX propose de publier les métadonnées dans des registres publics (dont un

24. Pour cette tâche, l'ontologie des indicateurs devrait permettre de reconnaitre les instance d'indicateurs répondant au critère de la requête, par l'analyse des informations portées dans les entités *Indicator*, *Theme*, *Keyword* et *Thesaurus*, comme nous l'expliquons en perspective.

registre qui est implémenté par EUROSTAT). À partir de la consultation de ces registres, l'utilisateur (ou un automate) peut alors récupérer l'adresse électronique d'un service Web diffusant des données statistiques sur requête des utilisateurs (dans un modèle « *pull* »), comme le montre la figure 3.4 page 86. SDMX est un format auto-décrit puisque la grammaire du document XML proposant les données est incluse dans les documents qui sont diffusés à l'utilisateur. De ce fait, le format est très interopérable. C'est un fonctionnement très similaire à celui que propose l'OGC avec les services Web de catalogage et de fourniture des données (WFS, [ISO 10]).

Pour pour pouvoir diffuser les données collectées dans un modèle SDMX, il est nécessaire de traduire le contenu du profil *esponMD* vers un modèle SDMX. La création d'un modèle SDMX correspondant à l'information territoriale statistique et ses métadonnées (celles qui ont été définies dans le profil *esponMD* de la norme ISO 19115) consiste essentiellement à créer le méta-modèle du document DSD qui sert à établir le schéma des documents SDMX-ML qui contiendront les données et leurs métadonnées. Il s'agit donc en premier lieu de recenser les concepts qui sont utilisés dans la norme ISO 19115, et leurs équivalents, s'il existent, dans la norme SDMX. Si les concepts sont différents, ou s'il faut définir des codes pour les décrire, nous le signalons. En second lieu, il faut proposer une structure de l'information. Pour cela, on distingue ce qui participe à l'identification d'une valeur, constituant ainsi une dimension de l'information, de ce qui qualifie une valeur et s'apparente donc à un attribut SDMX. Enfin, il faut rattacher les concepts à chaque niveau d'information.

6.2.4.0.1 Identification des concepts Il s'agit ici de recenser toutes les informations rendues obligatoires ou optionnelles dans la norme ISO 19115 et de préciser le concept existant équivalent en SDMX. Le tableau 8.1 en annexe page 271 associe à chaque élément du profil ISO 19115 de *esponMD* son concept [25] équivalent. Un tiret « - » signale l'absence du concept équivalent à l'information présente dans le profil. Par exemple, pour le champ « rôle » d'un contact, il n'existe pas d'équivalent dans les concepts SDMX. Parfois, plusieurs concepts sont susceptibles de correspondre à l'information issue du profil, ils sont alors tous énumérés. Réciproquement, il existe certains concepts présents dans SDMX mais absents de la norme ISO 19115, qui ne sont pas listés ici. Nous en mentionnons toutefois deux qui nous semblent pertinents pour notre profil. Par exemple, SDMX définit le concept de période de référence (BASE_PER) pour un indicateur qui peut-être très utile lorsque sont calculés des taux de croissance ou des indices en référence à une période. De même, les unités de mesure sont traitées de façon plus précises dans SDMX, et un concept UNIT_MULT sert par exemple à spécifier le multiplicateur des unités de mesures : sa valeur indique par quelle puissance de 10 les unités doivent être multipliées.

La table 8.1 page 271 montre qu'il existe des informations dans la norme ISO 19115 qui ne trouvent pas directement leur pendant dans les concepts SDMX. Pour ces informations, nous proposons de définir des concepts (dont le nom remplacera ISO19115_CONCEPTNAME), et de leur associer une description, en anglais et en français. Ces définitions pourront être utilisées dans la création des fichiers DSD, de la manière suivante :

```
<Concept agencyID="ESPON" id="ISO19115_CONCEPTNAME">
   <Name xml :lang="fr">Description en français</Name>
   <Name xml :lang="en">Description en anglais</Name>
</Concept>
```

25. La définition complète de chaque concept est disponible dans un document accessible en ligne : `http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf`

La table 6.3 présente notre proposition de nouveaux concepts en vue de traduire des métadonnées au format du profil *esponMD* de la norme ISO 19115 vers des métadonnées au format SDMX, avec leur traduction française.

TABLE 6.3 – Nouveaux concepts SDMX et leur équivalent du profil *esponMD*.

Attribut utilisé dans le profil *esponMD*	Nouveau concept SDMX	Traductions
individualName	CONTACT_PERSON	Nom d'une personne qui sert de contact pour ce jeu de données
role	CONTACT_ROLE	Rôle de ce contact par rapport au jeu de données
label	QUALITY_VALUE_ID	Etiquette attachée à un groupe de valeurs ayant la même provenance, pour lesquelles la qualité, ainsi que les contraintes légales de diffusion sont identiques
level	QUALITY_VALUE_SCOPE	Niveau du groupe de valeurs - de valeur constante « *tile* »
date	PUBLI_DATE	Date de publication des observations (millésime)
readRight	METADATA_ACCESS_RIGHT	Droit d'accès aux métadonnées (libre ou non)
code	INDICATOR_ID	Code de l'indicateur
thesaurus	THESAURUS	Nom d'un thésaurus
topicCategory	THEME_CODE	Code d'un thème sélectionné dans le thésaurus pour décrire l'indicateur
keyword	KEYWORD	Mot-clé décrivant l'indicateur

6.2.4.0.2 Structure de l'information Cette section définit quelles sont les informations qui doivent être considérées comme des dimensions (Dimension) au sens SDMX, de celles qui ne servent que pour la description des données et qui sont donc des attributs (Attribut) au sens SDMX. Une valeur (OBS_VALUE) est identifiée par :

- le nom du jeu de données (DSI) auquel elle appartient,
- la date d'acquisition du jeu de données dans le système d'information (DATA_UPDATE)
- le nom de l'indicateur qu'elle décrit (INDICATOR_ID),
- la date ou période de validité (TIME_PERIOD),
- l'unité territoriale qu'elle décrit (STAT_UNIT),
- la source (ORIG_DATA_ID) de cette donnée,
- le producteur (COMPILING_ORG) de cette donnée,
- la date d'extraction (ou de publication) de ces données (PUBLI_DATE).

Les autres informations sont donc toutes des attributs au sens SDMX. Le tableau 8.2 en annexe page 274 résume l'ensemble des concepts introduits précédemment en indiquant leur statut (D pour Dimension, ou A pour Attribut), leur niveau de description : Dataset pour jeu de données, Group pour groupe de valeurs, Series pour séries de valeurs et Obs pour les observations, et le mode de description : textuel ou bien le nom de la liste de codes à employer.

Dans la liste des concepts précédents qui ont été retenus pour faire partie du modèle SDMX correspondant au profil *esponMD*, de nouvelles listes de codes ont été introduites, qu'il s'agit de valuer. La liste CL_QUALITY_LEVEL correspond aux valeurs définies pour le profil *esponMD* dans l'élément QualityLevelCode du profil. La liste CL_ROLES correspond aux rôles définis dans la norme ISO 19115, dans l'élément CI_RoleCode, et la liste CL_SCOPES correspond à l'ensemble des portées qu'une information sur le jeu de données peut avoir et qui sont définies dans l'élément MD_ScopeCode de la norme ISO 19115 [26].

À partir de ces informations, la création du fichier DSD est directe, ainsi que la génération du fichier XSD établissant la grammaire des fichiers SDMX-ML qui serviront à échanger des données. Le modèle SDMX proposé reprend les informations minimales du profil *esponMD* de la norme ISO 19115, mais il est envisageable de l'enrichir avec d'autres informations, comme par exemple des listes codifiant les unités de mesure.

6.3 Conclusion

Dans cette étude, nous nous sommes intéressés à la possibilité de mettre en œuvre la norme ISO 19115 pour établir des métadonnées pour des données statistiques de type socio-économiques, à référence spatiale et temporelle. Nous répondons à la question de la compatibilité de ce type d'information avec la norme ISO 19115 de façon positive, moyennant une adaptation de la norme dans une extension, c'est-à-dire un profil. Le profil *esponMD* créé a pour le but de prendre en compte les trois niveaux d'information identifiés et de simplifier l'acquisition des métadonnées dans un profil de la norme.

Nous présentons aussi un cadre opérationnel dans lequel ce profil peut être utilisé, avec le flot de données associé. Sans entrer dans certains détails (interface d'extraction des données de la base), qui sont expliqués dans des rapports techniques [Grasland 10b, Plumejeaud 10, Grasland 10c], nous montrons que ce profil peut-être utilisé pour construire un système d'information dit « actif » respectant complètement la directive INSPIRE et facilitant l'acquisition de métadonnées. Cette proposition définit également un cahier des charges pour un éditeur de métadonnées, et elle propose un schéma XSD conforme au modèle SDMX pour l'export des données dans un format interopérable.

Cependant, l'interopérabilité n'est assurée que dans le sens de l'export en SDMX de métadonnées saisies avec le profil *esponMD*. À l'heure actuelle, traduire un fichier SDMX vers le profil *esponMD* de la norme ISO 19115 impliquerait une perte d'information. En effet, certains concepts présents dans un modèle SDMX tels que FREQ, le concept de fréquence de publication des données, ou bien le concept BASE_PER qui indique l'année de l'indice de référence dans le cas du taux d'évolution d'un indicateur sont absents de la première version du profil actuellement utilisé. Il aurait été préférable d'assurer tout de suite cette interopérabilité, mais cette recherche effectuée dans le cadre d'un projet Européen devait aussi respecter certains impératifs opérationnels, en terme de délais. Également, il fallait tenir compte des utilisateurs qui allaient renseigner les métadonnées. En effet, comme l'introduction de ce chapitre l'explique, les standards de métadonnées n'étaient pas vraiment compris ni acceptés par l'ensemble des utilisateurs, et la mise au point de ce profil représente à la fois un travail de pédagogie important, mais aussi quelques concessions faites à sa complétude. Cependant, avec la disponibilité actuelle de l'éditeur de métadonnées, et le recul, les utilisateurs seront plus facilement convaincus de l'utilité de nouveaux

26. Disponibles en ligne sur l'adresse suivante : `http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml`

champs d'information. Nous pouvons donc envisager d'étendre le profil *esponMD* dans une nouvelle version intégrant les concepts SDMX qui nous semblent utiles, voir indispensables.

Par ailleurs, si l'acquisition de métadonnées dans un format structuré permet de résoudre l'interopérabilité syntaxique, pour l'interopérabilité sémantique, notre proposition a montré qu'il reste encore à traiter ces métadonnées de façon à manipuler automatiquement leur contenu. En effet, suivant les producteurs de données, les descriptions qu'elles contiennent ne sont pas harmonisées (aussi bien en termes de classification qu'en termes de description des transformations par exemple), et ne permettent pas un traitement automatique de cette information. Nous esquissons en perspectives une solution basée sur l'exploitation de ces métadonnées pour la construction d'une ontologie des indicateurs qui pourrait résoudre le problème d'interopérabilité sémantique.

Chapitre 7

Méthodes pour l'exploration et l'analyse de l'information statistique territoriale

Ce chapitre est l'aboutissement des deux précédentes propositions qui offrent un modèle pour formaliser l'information statistique issue de sources hétérogènes, associée à différents maillages changeant au cours du temps, et décrite par un ensemble de métadonnées renseignant sur sa provenance et sa nature. Cette modélisation de l'information peut être très utile à l'analyse des données, et suscite de nouveaux modes d'exploration des données.

Nous proposons d'abord une plate-forme pour l'analyse interactive des données et de leur qualité qui permet de repérer les valeurs exceptionnelles. Ensuite, nous montrons l'intérêt que l'intégration de notre modèle spatio-temporel du support représenterait pour l'analyse et l'interprétation de l'évolution des écarts territoriaux, en lien avec les changements territoriaux qui ont été identifiés, et la recherche d'évolutions exceptionnelles.

7.1 Exploration interactive de la qualité des données

Dans cette section, nous proposons un système pour vérifier la consistance (ou cohérence) des données collectées dans le système d'information spatio-temporel via des méthodes d'exploration interactives, et la mise en oeuvre de méthode de détection de valeurs exceptionnelles. Nous rappelons d'abord les motivations de cette proposition, puis nous décrivons en détail l'architecture d'un tel système, et nous montrons les résultats que nous avons obtenus avec notre premier prototype, nommé QualESTIM.

7.1.1 Motivations

L'organisation de l'activité sociale et économique d'une société requiert la connaissance d'un grand nombre d'indicateurs de développement sur des périodes temporelles étendues. Ces indicateurs (démographiques, économiques, financiers, etc.) sont en effet utilisés pour établir des scénarios d'évolution à plus ou moins long terme sur des espaces géographiques d'intérêt : l'Europe ou le Maghreb, par exemple. Des données peu fiables, incomplètes, avec de nombreuses erreurs de saisie ou simplement issues d'une

méthodologie biaisée peuvent produire des scénarios d'anticipation approximatifs, voire faux. Dans ce contexte, la qualité des données statistiques joue un rôle important [Chrisman 84]. En vue de documenter cette qualité, les jeux de données peuvent être accompagnés de métadonnées qui donnent des informations sur la provenance des valeurs, sur la méthode de calcul utilisée pour les obtenir, ou encore sur une estimation de la qualité par le fournisseur lui-même. La norme ISO 19115 offre une structure pour ces rapports, voir la figure 3.3 page 83, mais celle-ci est trop complexe pour les utilisateurs. Ainsi, dans le cadre de l'opérationalisation de la norme nous avons été amené à collecter des rapports non structurés (des fichiers, des documents multimédias), voir la section 6.1.4.3 page 202. Mais ces rapports d'expertise dans un format très descriptif ne sont pas forcément exploitables de façon automatique [Dean 96]. Nous nous intéressons ici aux autres moyens permettant de vérifier la qualité interne des données, et plus particulièrement la précision sémantique des données.

Cette évaluation de la qualité peut reposer sur une étape de détection de valeurs exceptionnelles, qui vise à identifier les valeurs qui sont très différentes de leur voisinage (temporel, spatial et thématique). Plusieurs méthodes (géo)statistiques permettent l'évaluation de la qualité d'un jeu de données en qualifiant chacune des valeurs comme étant exceptionnelle ou non. Cependant, suivant le type de méthode employée, et/ou son paramétrage, les résultats d'évaluation d'une même valeur ne sont pas forcément concordants. Il faut alors envisager d'exécuter plusieurs méthodes afin de vérifier si ces méthodes convergent vers l'attribution du qualificatif « exceptionnelle » à la valeur. Néanmoins, cela n'est pas toujours suffisant pour qualifier la valeur d'erreur de mesure : par exemple, la valeur du Produit Intérieur Brut (PIB) par habitant du Liechtenstein est très haute par rapport à son voisinage spatial, sans être une erreur. En réalité, à ce point de l'analyse, seul un expert peut trancher et décider du niveau de fiabilité de chaque valeur remarquée. L'intégration de l'expertise humaine dans un système d'évaluation de la qualité semble donc indispensable. L'usage de méthodes de recherche de valeurs exceptionnelles permet en tout cas de délimiter des sous-ensembles de valeurs *a priori* suspectes, et qui peuvent se révéler très intéressantes sur le plan thématique. À ce point de l'analyse, seul un expert peut trancher et décider du niveau de fiabilité de chaque valeur remarquée. Pour l'aider dans son analyse, il nous paraît pertinent de présenter la valeur suspecte avec ses métadonnées, expliquant la provenance de la valeur par exemple, mais aussi le niveau de confiance que le fournisseur lui-même accordait à cette valeur.

Dans ce cadre, un système permettant de paramétrer et d'exécuter des méthodes sur des sous-ensembles de données choisis par un utilisateur, et lui permettant de visualiser aisément leurs résultats dans une interface cartographique dynamique, serait un apport certain pour l'analyse de la qualité des données. En effet, via ce type d'interface, nous pourrions répondre aux questions suivantes : quels sont les lieux géographiques pour lesquels toutes les méthodes révèlent la présence d'une valeur exceptionnelle pour l'indicateur choisi ? Ces valeurs suspectes proviennent-elles du même fournisseur ? Via cette interface, l'expert serait amené à se questionner sur l'exactitude et la véracité d'un sous-ensemble réduit de valeurs. Par ailleurs, si ces valeurs exceptionnelles ne sont pas erronées, elles sont alors particulièrement intéressantes pour le thématicien, dans une optique d'aménagement du territoire, parce qu'elles signifient qu'un phénomène spécifique est à analyser.

7.1.2 Un système interactif dédié à l'évaluation de la qualité

Avec cette proposition dédiée à l'évaluation de la qualité l'information statistique, il s'agit de montrer l'intérêt que la connaissance des métadonnées peut représenter pour l'ESDA (*Exploratory Spatial Data Analysis*) pour l'identification de valeurs exceptionnelles, et l'analyse de la qualité des données. Nous souhaitons donc articuler la fonction de recherche de valeurs exceptionnelles, classique en ESDA, avec :

– des fonctions d'extraction de données d'une base de données spatio-temporelles, accompagnées de leurs métadonnées,
– la production de rapports d'analyse,
– l'affichage des résultats basé sur une visualisation cartographique interactive.

La figure 7.1 présente une vue globale de l'architecture de notre système dédié à l'évaluation de la qualité, nommé *QualESTIM*.

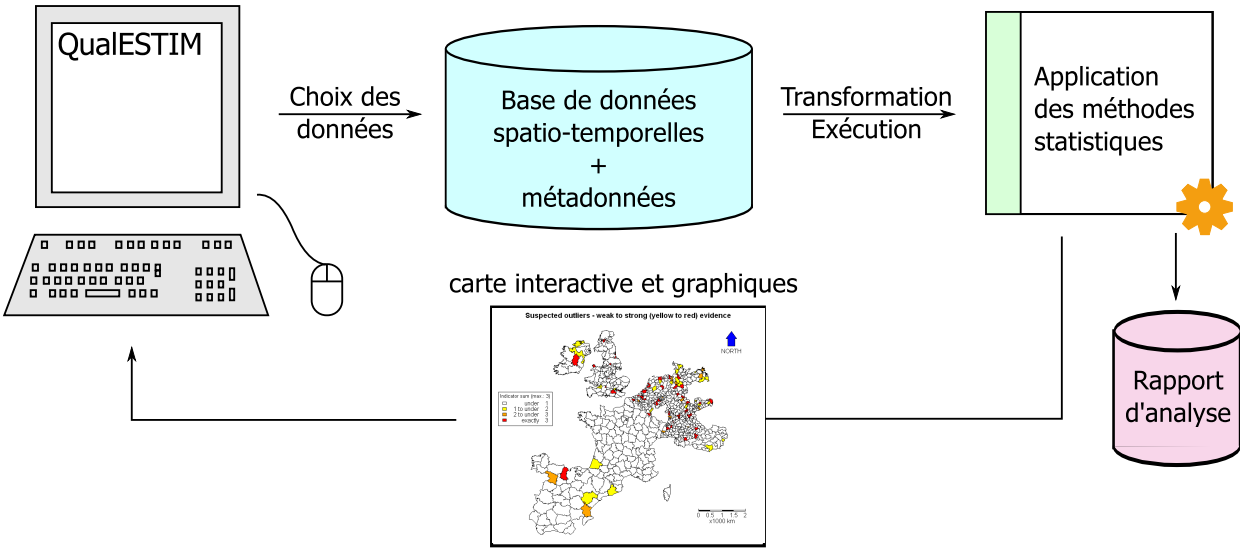


FIGURE 7.1 – Architecture de QualESTIM - vue générale.

Le cycle itératif d'analyse proposé se base sur l'approche du célèbre mantra « *Overview, Zoom and Filter, Details on demand* », [Shneiderman 96], et intègre la capacité à archiver des analyses. Il s'agit d'abord de se donner une vue générale de l'ensemble des données, de pouvoir concentrer son attention sur des sous-ensembles, et de filtrer l'information selon certains critères, et enfin de demander des informations supplémentaires sur certaines données ainsi repérées. Shneiderman ajoute que ce processus peut et doit être réitéré à l'infini et qu'il est important que l'utilisateur puisse avancer puis revenir en arrière dans ses explorations. Ces allers-retours impliquent des sauvegardes des visualisations de différentes sous-parties (un historique des photos produites), mais également une sauvegarde du chemin (en quelque sorte la carte mentale des manipulations) ayant conduit à cette photographie des données. Natalia et Gennady Andrienko insistent également très fortement sur *l'aspect itératif* de l'analyse exploratoire [Andrienko 06]. Ils ajoutent que l'utilisateur se donne un but général d'investigation qui motive l'analyse. Par celle-ci, d'autres questions peuvent être suscitées, à des niveaux de détail (et/ou d'abstraction) différents, et ces questions ne sont pas formulées à l'avance par l'utilisateur. Ce dernier découvre ainsi ses données, et s'offre les moyens de les interpréter dans un processus itératif, interactif et dynamique.

Nous proposons ici que l'objectif de cette analyse exploratoire soit de repérer des valeurs exceptionnelles en combinant plusieurs types d'analyses, puis d'exporter les résultats de ces analyses dans un rapport conforme à la norme ISO 19115. Ce rapport peut, en retour, servir à enrichir automatiquement les données. Par ailleurs, les conclusions que l'utilisateur tire des relations semblant exister entre les données peuvent être réutilisées dans le cadre d'un processus d'harmonisation ou de transfert des données. Etant donné que l'utilisateur peut intégrer sa propre expertise dans l'analyse des données, l'outil que nous proposons ne peut être utilisé qu'après authentification de l'utilisateur.

Les étapes du processus d'analyse sont les suivantes :
- L'utilisateur choisit le jeu de données qu'il souhaite analyser, via une interface qui lui permet d'interroger la base de données spatio-temporelle. Un premier affichage cartographique avec curseur temporel lui permet d'avoir un aperçu de la distribution des données, de la quantité de valeurs manquantes. Il est encore dans une vue générale (*Overview*) .
- Par la suite, il peut s'intéresser à un sous-ensemble de valeurs qui sont mises en évidence par l'usage de méthodes de recherche de valeurs exceptionnelles. Dans cette phase de filtrage (*Filter*), il choisit une méthode qu'il paramètre, et demande son exécution. En retour, les unités dont les valeurs sont considérés comme exceptionnelles sont surlignées en rouge dans la carte choroplèthe. Le rapport d'analyse est affiché sous la forme de cartes et de diagrammes.
- L'utilisateur peut demander plus de détails (*Details on demand*) sur la provenance de valeurs qui semblent exceptionnelles pour une ou plusieurs méthodes : en cliquant sur une unité, les métadonnées correspondant à cette unité, cet indicateur et le jeu de données sont affichées.
- Ce processus de filtrage peut être réitéré. L'exécution de chaque méthode sélectionnée et paramètrée par l'utilisateur produit un rapport d'analyse, qui est conservé et peut être combiné avec d'autres.
- Enfin, l'utilisateur peut souhaiter archiver ces rapports, en intégrant les conclusions de son analyse propre, dans un espace qui lui est dédié (l'onglet « expertise »).

L'interface graphique de QualESTIM (voir figure 7.2) est structurée selon la logique de ce processus. Une première zone (zone 1 de la figure 7.2) de cette interface est dédiée à la sélection des données à analyser : l'utilisateur choisit un espace d'étude, un niveau de zonage, un jeu de données, l'indicateur dont il souhaite évaluer la qualité, et la période temporelle. Dans une seconde zone, il choisit la méthode d'évaluation qu'il souhaite appliquer à l'échantillon sélectionné et spécifie les valeurs des paramètres des scripts (si nécessaire). Dans la troisième zone, les résultats de ses requêtes et analyses sont affichés. La quatrième zone est dédiée aux métadonnées associées au contexte de cette analyse.

7.1.2.1 Sélection des données

L'interface de requête proposée ne permet que la sélection de données issues d'un même jeu de données. Bien qu'il soit possible de proposer d'autres modalités pour l'interrogation de la base de données en vue de combiner des indicateurs issus de jeux de données différents, ce choix vise à centrer l'intérêt de l'utilisateur sur la cohérence des données issues d'un même jeu de données.

Les données statistiques sont extraites d'une base de données spatio-temporelles dont le schéma est établi à partir du modèle identitaire évolutif décrit dans le chapitre 5. Le modèle est structuré de façon à rendre compte de la dimension hiérarchique de l'organisation spatiale, qui sera réutilisée lors de l'interrogation du modèle. En effet, il décrit les zonages (et leurs versions) comme des sous-ensembles d'unités formant un découpage d'un certain niveau du territoire, découpage valide durant une certaine période. Par exemple, la NUTS (Nomenclature des Unités Territoriales Statistiques) comprend six versions de nomenclature (1980, 1988, 1995, 1999, 2003, 2006), qui définissent cinq niveaux administratifs chacune (des communes aux états). Ce modèle est centré sur les unités géographiques, qui possèdent une période de validité, une identité (nom, code, centre, etc.), une extension spatiale (la géométrie des unités) et une partie thématique qui décrit les indicateurs statistiques disponibles sur ces unités, les valeurs associées, avec leurs différentes périodes de validité. La géométrie des unités est versionnée car les méthodes d'analyse spatiale tiennent compte des géométries des unités sur lesquelles sont collectés les indicateurs et ces géométries changent suivant les versions de nomenclature.

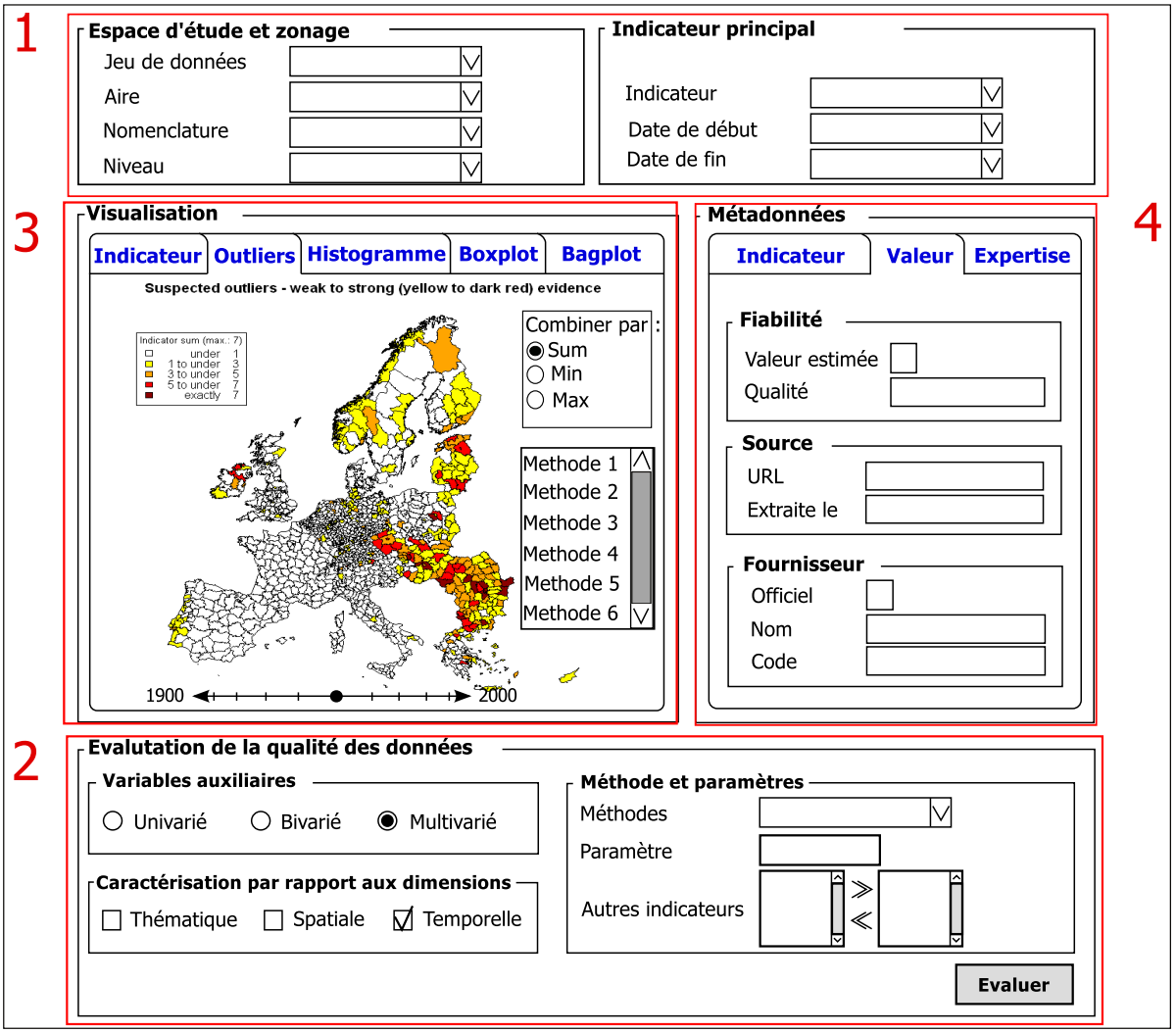


FIGURE 7.2 – Schéma de l'interface graphique de QualESTIM.

La version de nomenclature peut être configurée par l'utilisateur, et ne sont retournées que des données existant dans la version configurée. L'interface contraint l'utilisateur à d'abord choisir un jeu de données, puis l'aire d'étude, puis la nomenclature, puis le niveau auquel il souhaite interroger la nomenclature, puis enfin l'indicateur existant pour les conditions prédéfinies (jeu de données, aire d'étude, nomenclature, et échelle). Cette contrainte évite de poser des requêtes qui n'ont pas de réponse. Pour l'indicateur sélectionné, l'intervalle temporel pour lequel des données existent, délimité par la date de début et la date de fin, est renvoyé. L'utilisateur a la possibilité de mener son étude sur cette plage temporelle, ou bien de restreindre cette plage. Lorsque pour l'espace, la date et l'indicateur sélectionné, des valeurs sont manquantes, l'interface cartographique signale par un motif hachuré l'absence de valeurs. Idéalement, l'interface devrait permettre de choisir d'évaluer des dérivées temporelles d'indicateurs de type quantitatifs absolus : il faudrait pour cela autoriser l'utilisateur à sélectionner un même indicateur ayant deux plages de validité différente, et à en calculer le ratio. Cependant, l'interface que nous proposons est pour l'instant limitée au choix d'une seule plage de validité par indicateur et ne permet pas d'exprimer des opérations entre indicateurs (telles que le rapport entre deux variables) comme le fait HyperAtlas.

Par ailleurs, cette base de données conserve également un ensemble de métadonnées qui ont été collectées suivant un profil de la norme ISO19115 adapté pour les données statistiques, que nous avons décrit dans le chapitre 6, et qui apportent des informations sur la provenance (ou lignage) des valeurs, sur les trois niveaux d'information (jeu de données, indicateur et valeur). Nous ne proposons pas de rechercher les données par les métadonnées, mais en revanche, l'affichage des données se fait de façon conjointe avec les métadonnées associées dans la zone 4 de l'interface. L'analyse des métadonnées permet, en principe, de distinguer les indicateurs de type quantitatif relatif (les taux) des indicateurs de type quantitatif absolu (les stocks). En effet, l'analyse des stocks et des taux relève de procédures de test de nature profondément différentes.

7.1.2.2 Choix, paramétrage et exécution de méthodes statistiques

Dans la zone 2 de l'interface graphique, l'utilisateur est invité à choisir le type de méthode d'évaluation de la qualité qu'il souhaite exécuter sur le jeu de données sélectionné. L'ensemble des méthodes statistiques a été structuré en fonction de deux critères. Le premier porte sur la dimension que la méthode permet d'explorer (thématique, spatiale ou temporelle). Pour la dimension spatiale, les méthodes vérifient la variabilité des valeurs en fonction de voisinages spatiaux, mais pour des taux uniquement, puisque les méthodes pour la mesure de l'autocorrélation spatiale ne s'appliquent qu'à des taux. Il est aussi possible de pratiquer une analyse du taux de variation d'une variable de stock, qui sera alors à la fois spatiale et temporelle. Le second critère porte sur le nombre de variables auxiliaires que la méthode requiert (aucune pour une méthode univariée, une pour une méthode bivariée, et plus d'une pour une méthode multi-variée). Par exemple, dans le cadre d'une étude bivariée, l'utilisateur pourrait choisir comme indicateur principal, le Produit Intérieur Brut (PIB) par habitant, et comme variable auxiliaire, le taux d'urbanisation ou l'espérance de vie ou encore la part des services dans la population active (mais seulement si ces variables existent sur une même période temporelle).

Ainsi, en fonction du nombre de variables auxiliaires sélectionnées et des dimensions choisies, la liste des méthodes disponibles est mise à jour. En effet, une correspondance a été définie dans le modèle des méthodes entre le type d'analyse (dimensions et nombre de variables) et les méthodes implantées dans QualESTIM. Par exemple, la méthode du *Boxplot* prend en paramètre un jeu de données univarié et analyse suivant la dimension thématique, alors qu'une technique comme la *Régression Géographiquement Pondérée* considère la variabilité d'un jeu de données multivarié selon la dimension spatiale. De plus, chaque méthode requiert des paramètres qui lui sont spécifiques et l'interface est mise à jour en fonction de la méthode sélectionnée.

7.1.2.3 Exploitation et interprétation des résultats

Les résultats de l'analyse statistique apparaissent dans la partie « Visualisation » de l'interface (zone 3). L'onglet "Indicateur" représente la carte avec les données brutes choisies par l'utilisateur. L'onglet "Outliers" permet la visualisation de la dernière évaluation effectuée par la méthode sélectionnée. Les autres onglets comportent toutes les autres formes de graphiques calculés par cette méthode (histogramme, *boxplot*, etc.).

L'interface est prévue pour autoriser la comparaison de plusieurs résultats de méthodes. Pour cela, nous associons un drapeau [1] à chaque rapport d'analyse produit par l'exécution d'une méthode. L'utilisateur peut revoir un ancien rapport en sélectionnant le drapeau correspondant. En effet, au survol du drapeau par la souris (voir figure 7.3), une bulle apparaît et contient des informations identifiant le rapport, comme sa date d'exécution, et d'autres informations comme la méthode et les paramètres utilisés.

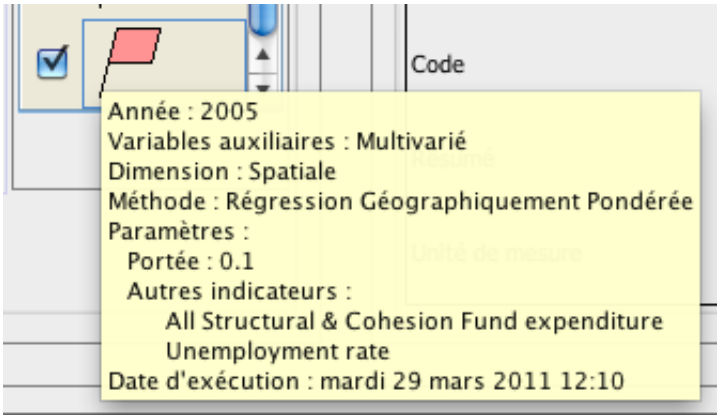


FIGURE 7.3 – Bulle d'information associée à une exécution de la Régression Géographiquement Pondérée.

Les rapports d'analyse de chaque méthode n'étant pas forcément concordants pour une même valeur, nous proposons plusieurs méthodes de combinaison des résultats, afin de restituer une vue d'ensemble de ces analyses. La combinaison des résultats consiste à superposer les cartes dans l'onglet « Outliers » suivant trois méthodes d'agrégation différentes, qui sont proposées à l'utilisateur. La « somme » consiste à assombrir l'aplat de couleur d'une unité territoriale lorsqu'augmente le nombre d'exécutions de méthodes considérant sa valeur comme exceptionnelle. Le « minimum » représente les unités territoriales dont la valeur d'indicateur n'est jamais évaluée comme exceptionnelle par l'ensemble des analyses exécutées, et le « maximum », celles dont la valeur d'indicateur est toujours considérée comme exceptionnelle.

Le *slider* temporel (ou barre de progression temporelle) permet à l'utilisateur de définir la date à laquelle il souhaite visualiser le résultat de l'analyse pour l'indicateur choisi. En effet, pour un indicateur annuel couvrant une période temporelle de dix ans par exemple, *dans la version de nomenclature choisie pour l'analyse des données*, la méthode fournit une analyse des valeurs pour chaque année et les résultats de l'analyse peuvent varier en fonction des années. L'utilisateur peut déplacer le curseur du *slider* et visualiser les résultats de l'analyse de l'indicateur pour chaque année de l'intervalle de temps choisi.

7.1.2.4 Exploitation des métadonnées

Pour rappel, l'objectif est d'enrichir les métadonnées qui accompagnent les valeurs statistiques à partir des résultats de l'évaluation de la qualité des données. Dans un premier temps, l'utilisateur accède aux métadonnées fournies avec le jeu de données dans la zone 4 de l'interface. Lorsque l'utilisateur parcourt la carte avec la souris, les onglets de métadonnées sont mis à jour en fonction de l'unité géographique survolée. Ils présentent des informations sur les indicateurs ou sur les valeurs (identité du fournisseur, description, etc.), voir figure 7.2. Concernant les indicateurs, nous affichons le nom de l'indicateur, le

1. La couleur rouge du drapeau ne correspond à aucune sémantique particulière.

code de l'indicateur, un résumé, une unité de mesure, le nom du jeu de données et la date d'acquisition. Ces informations sont importantes car elles indiquent à l'utilisateur la provenance des valeurs analysées, selon plusieurs niveaux d'information (jeu de données, indicateur et valeur elle-même).

Cette exploitation des métadonnées issues de la base de données permet de mettre en relation ces connaissances avec l'estimation faite par les méthodes de la qualité des valeurs observées. Par exemple, l'utilisateur peut observer que, très fréquemment, une source de données particulière apporte de nombreuses valeurs exceptionnelles, et ceci peut aussi l'amener à réviser le niveau de qualité estimé d'un groupe de valeurs, à la baisse ou à la hausse. Ainsi, suite à l'exécution d'une ou plusieurs méthodes sur un jeu de données, l'utilisateur a la possibilité de réévaluer le niveau de qualité des données dans l'onglet « Expertise » de la zone 4, en y associant le rapport d'analyse généré.

Les rapports d'analyse des différentes exécutions de méthodes de détection de valeurs exceptionnelles sont conservés, durant la session de l'utilisateur, dans des tables temporaires (voir figure 7.4). Celles-ci sont associées à un contexte d'analyse défini par les choix de l'utilisateur dans la première zone (voir figure 7.2, zone 1). À chaque exécution (*Run*) est associée un utilisateur (*userId*), et une date d'exécution (*date*), mais aussi un contexte d'analyse qui est défini par le niveau de zonage (*zoningLevel*), le nom de l'aire d'étude, (*studyAreaName*), la période de l'analyse représentée par l'intervalle [*startTimePeriod, endTimePeriod*], les indicateurs (*Indicator*) impliqués et les jeux de données (*Dataset*) dont ils sont issus chacun. On qualifie le rôle de chaque indicateur dans l'exécution, afin de signaler si les valeurs de l'indicateur étaient l'objet de l'étude (*inspected*), ou bien des variables auxilliaires (*ancillary*). La méthode de calcul employée (*Method*) et les valeurs de paramètres utilisés (*ParameterInstance*) sont associées à cette exécution. Les résultats de l'exécution (*Analysis*) donnent lieu à une notation de la valeur inspectée (*Value*) de type numérique entier : plus la note est élevée, plus la méthode considère comme suspecte la valeur. À zéro, la valeur est considérée comme normale durant cette exécution.

Un modèle des méthodes est introduit ici, pour décrire les instances des méthodes utilisées. Chaque méthode est modélisée par son nom (*name*) et sa description (*description*), le nombre d'indicateurs auxilliaires qu'elle peut prendre en compte (*nbAncillaryVar*), et les dimensions explorées : *timeDim*, *spaceDim* ou *thematicDim* qui valent vrai si la dimension temporelle, spatiale ou thématique respectivement sont explorées. De même, les paramètres (*Parameter*) que la méthode doit utiliser sont décrits par leur nom (*name*), et leur type (*EnumType*). Le type peut être un type de base parmi l'énumération suivante : booléen, entier, réel ou chaîne de caractères. Lorsque le paramètre est de type numérique, sa valeur minimum (min), maximum (max) sont précisées, et une formule stockée dans une chaîne de caractères, *optimumFormula*, décrit comment spécifier une valeur adaptée au jeu de données sélectionné.

Après l'exécution d'un ensemble de méthodes de détection, une synthèse des méthodes d'estimation peut être établie. Elle est représentée sous la forme d'une classe *SynthesisReport*, associée à la valeur inspectée (*Value*), et à l'ensemble des rapports d'analyse (*Analysis*) qui servent de base à cette synthèse. La synthèse produit automatiquement un indicateur, *systemQualityLevel*, qui représente le rapport entre la somme des notations appliquées à la valeur, et le nombre d'exécutions considérées. Ce rapport augmente avec le nombre de méthodes qui estiment la valeur exceptionnelle. Néanmoins, l'utilisateur peut ensuite rectifier ce rapport avec sa propre estimation de la qualité, établie d'après ses connaissances (*userQualitylevel*). L'objectif des valeurs *systemQualityLevel*, qui sont affichées dans l'interface sur une carte choroplèthe, est d'attirer l'attention de l'expert, qui peut trancher après analyse avec sa propre notation sur la valeur. Nous enrichissons ainsi progressivement les métadonnées avec des rapports d'expertise documentés. Il est ainsi possible de savoir quelles méthodes ont produit les rapports d'analyse, et dans quelles conditions (liste des paramètres utilisés).

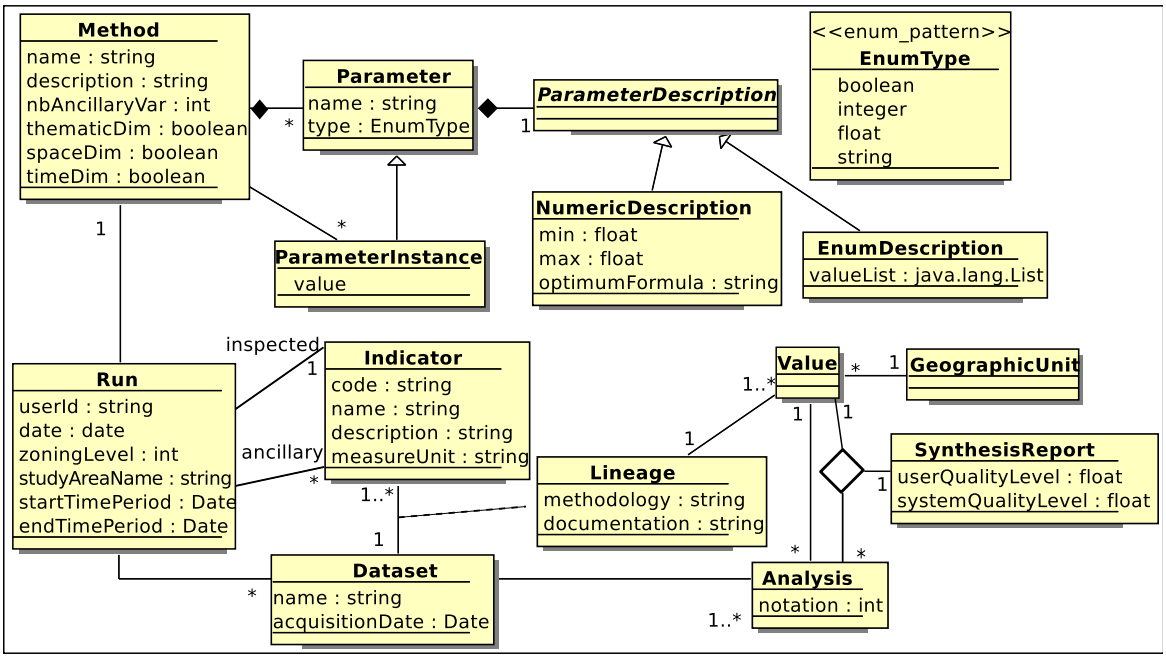


FIGURE 7.4 – Modèle UML du rapport généré, décrivant les méthodes, le contexte d'exécution, et le résultat des analyses.

Ce rapport peut être intégré dans l'élément DQ_Report de la norme ISO 19115, attaché à chaque élément de *DataQuality* du profil *esponMD*. Il peut également être exporté dans un fichier au format XML, contenant l'ensemble des métadonnées associées aux valeurs observées pour ce jeu de données.

7.1.3 Mise en œuvre dans QualESTIM

Le prototype qui implémente notre proposition a permis de valider l'approche consistant à coupler des scripts écrits dans un langage et un environnement adapté au calcul scientifique, R. A travers ce prototype, qui n'intègre que partiellement la dimension temporelle, nous pouvons souligner ce que pourrait apporter une intégration spatio-temporelle et thématique complète des données, mais également les difficultés qui doivent encore être surmontées. Nous discutons également le rôle que peuvent ou doivent jouer les utilisateurs de tels logiciels.

7.1.3.1 Implementation de QualESTIM

QualESTIM a été développée en Java, et utilise un connecteur JDBC pour interroger la base de données relationnelle PostgresSQL avec extension spatiale PostGIS implémentant le schéma spatio-temporel qui a été décrit dans le chapitre 5. L'interface graphique a été réalisée avec JavaSwing, tandis que les méthodes de recherche de valeurs exceptionnelles sont développées avec R. Utiliser R comme langage de programmation des méthodes de recherche de valeurs exceptionnelles avait l'avantage de réutiliser du code qui existe et qui a été validé dans la communauté statistique qui travaille avec R. Bien que R puisse s'interfacer avec des bases de données, ce n'est pas un langage prévu pour le développement d'interfaces interactives. Nous souhaitions donc coupler R avec Java, langage de programmation objet, mieux adapté

au développement d'interfaces graphiques, et ceci d'une façon faible, rendant le code modulaire. Nous détaillons ici deux points techniques qui n'ont pas été simples à résoudre, qui sont relatifs à l'articulation de deux environnements d'exécutions (Java et R) assez différents, mais que nous pensons complémentaires.

7.1.3.1.1 Intégration des scripts R dans la JVM

Les scripts R que nous avons utilisés ont été développés par Martin Charlton et Paul Harris, du laboratoire du National Centre for Geocomputation [2], dans le cadre du projet *ESPON 2013 database* [Harris 10], pour la détection des valeurs exceptionnelles. Le tableau 7.1 liste l'ensemble des méthodes qui ont été utilisées, et spécifie leur classification (dimension et nombre de variables auxilliaires).

TABLE 7.1 – Liste de méthodes de recherche de valeurs exceptionnelles implémentées dans QualESTIM.

Méthode	Dimension	Nombre de variables auxiliaires
Boxplot standard	thématique	0
Boxplot ajusté	thématique	0
Bagplot	thématique	1
Distance de Mahalanobis	thématique	1 ou plus
Analyse en composantes principales	thématique	1 ou plus
Régression linéaire multiple	thématique	1 ou plus
Test de Hawkins	spatiale	0
Moyenne Locale	spatiale	0
Regression locale	spatiale	0 ou plus
Régression géographiquement pondérée	spatiale	0 ou plus

Aucun de ces scripts n'implémente d'exploration suivant la dimension temporelle. Il est possible d'intégrer des méthodes explorant la dimension temporelle, cependant, il existe de fortes restrictions sur l'intervalle temporel qui peut être utilisé pour ce type d'analyse, que nous discutons dans la section 7.1.3.3.

Nous détaillons ici comment ces scripts ont été intégrés dans le système QualESTIM développé en Java. Initialement, nous souhaitions créer un *wrapper* Java de ces méthodes implémentées dans des scripts R permettant simplement de leur passer les paramètres d'entrée, sans modifier le code de ces scripts. À cette fin, l'adaptateur JRI [3] (Java R Interface) qui permet d'embarquer l'environnement de R dans une JVM, peut être utilisé pour invoquer l'exécution de scripts R depuis la JVM. Cependant, des difficultés connues des utilisateurs de R nous ont obligés à changer d'approche en recopiant les scripts dans du code Java. Ce code est exécuté via la JRI, qui permet d'embarquer un objet *rengine* dans le code Java, qui appelle l'exécution de code R. Cette approche n'est pas particulièrement satisfaisante du point de vue de la modularité du code, comme le montre l'extrait de code Java 7.1 appelant la méthode de calcul de la régression géographique pondérée.

2. `http://ncg.nuim.ie/index.php`
3. `http://rosuda.org/JRI/`

```
        /**
         * Geographically weighted regression
         * @param bwd3 : bandwidth, must be inside [0, 1] range.
         */
        public void gwr(double bwd3) {

            // With coordinate data as explanatory variables (i.e. first-
                order polynomial).
            // Using spgwr
            String rfunction = "bisquare";
            double cutoff = 0.05; // 0.05 for 5% tails

            rengine.eval("data.1<-data1copy@data");

            // Defining the coordinates...
            rengine.eval("coords.1<-cbind(data.1[\"X\"],data.1[\"Y\"])");
            rengine.eval("coords.1<-as.matrix(coords.1)");

            rengine.eval("bwd.3<-"+bwd3);
            rengine.eval("gwr.p<-gwr(stock1~X+Y,data=data.1,coords=coords
                .1,adapt=bwd.3,gweight=gwr."+rfunction+",predictions=T)");

            // GWR raw residuals...
            rengine.eval("raw.resids.gwr<-stock1-gwr.p$SDF$pred");
            rengine.eval("summary(raw.resids.gwr)");

            // Identifying and updating outlier information in one file
            rengine.eval("cut.off.4<-quantile(raw.resids.gwr,probs=seq
                (0,1,"+cutoff+"),na.rm=T)");

            rengine.eval("indicator.25GWR<-ifelse(raw.resids.gwr>=cut.off
                .4[2]&raw.resids.gwr<=cut.off.4[20],0,1)");
            rengine.eval("data1copy@data<-cbind(data1copy@data,raw.resids.
                gwr,indicator.25GWR)");
            rengine.eval("data1copy@data<-as.data.frame(data1copy@data)");
            rengine.eval("attach(data1copy@data)");
            rengine.eval("data.1<-data1copy@data");

            rengine.eval("detach(data1copy@data");
        }
```

Listing 7.1 – Extrait de code Java appelant le script R pour le calcul de la régression géographiquement pondérée.

L'extrait de code Java 7.1 montre également que nous avons simplifié le paramétrage de la régression géographiquement pondérée. En effet, le seuil de test de significativité des résidus (*cutoff* est à 5%) ainsi que le modèle de régression (*bisquare* représente un modèle gaussien) ont été prédéfinis, et ne peuvent pas être paramétrés par l'utilisateur, contrairement à la portée (*bwd3*). Nous discutons dans la section 7.1.3.3 de la question du paramétrage des méthodes.

7.1.3.1.2 Transformation des données pour R Nous avons développé un module de connexion à la base de données permettant d'interroger et d'extraire les données sur la base des critères définis dans la section 7.1.2.1 concernant le contexte d'analyse. Le modèle de la base de données est conforme au schéma spatio-temporel qui a été décrit dans le chapitre 5. Ce module reconstitue un objet *SpatialData-PolygoneFrame* pour R, dont la figure 7.5 décrit la structure.

En effet, le paquetage sp de R sur lequel s'appuient les bibliothèques géostatistiques de R propose un modèle non topologique, où sont décrites d'une part les géométries des unités d'étude (le SpatialPolygons) sous la forme d'une liste de polygones, et d'autre part les valeurs des variables statistiques associées (dans un data.frame). Le lien entre le data.frame et le SpatialPolygons est réalisé à travers l'identifiant (ID) de chaque unité dans les deux structures. Cette structure est complètement expliquée dans [Bivand 08].

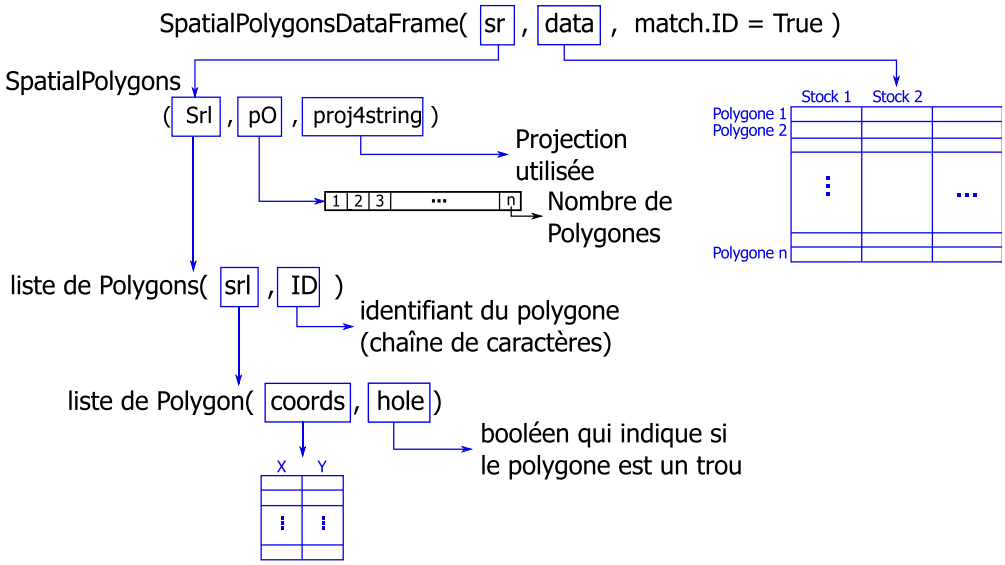


FIGURE 7.5 – Structure de données utilisée par R pour le calcul sur des données à références spatiales.

Nous détaillons ici le processus de reconstruction de cet objet à partir de la sélection des données extraites de la base de données. L'espace d'étude est composé d'un ensemble d'unités territoriales. Chacune de ces unités est considérée comme un multi-polygone. En effet, un premier polygone caractérise le contour de cette unité, et un ensemble d'autres polygones désignent les trous qu'elle contient, s'il y en a (par exemple, les lacs). Ces unités territoriales sont ensuite ajoutées à la liste des polygones (*Srl*) qui représente l'espace géographique. Cette liste est un des composants de l'objet *SpatialPolygons*, dans lequel est précisée aussi le type de projection utilisé, *proj4string*, ainsi que l'ordre de dessin des polygones, *pO*. En parallèle, une matrice (*data*) contenant un ensemble de valeurs d'indicateurs choisis est créée : chaque colonne de la matrice contient les données d'un indicateur pour une année (*stock1*, *stock2*, etc.) pour un jeu de données et chacune de ses lignes caractérise une unité territoriale (*Polygone1*, *Polygone2*, etc.). Cet objet représente un unique espace d'étude pour un jeu de données choisi par l'utilisateur. Lors de la constitution de cet objet, les unités dont la valeur est manquante pour au moins un des indicateurs sélectionnés sont retirées de l'analyse.

7.1.3.2 Validation de l'approche exploratoire QualESTIM

Un cas d'étude réalisé avec QualESTIM montre l'apport que l'usage de méthodes de recherche de valeurs exceptionnelles combinées avec la consultation des métadonnées peut représenter pour l'analyse exploratoire de données spatio-temporelles. Notre étude de cas porte sur l'évolution du Produit Intérieur Brut (PIB) par habitant sur les pays de l'Union Européenne de 2000 à 2005, dans la version de nomenclature des NUTS 2006 au niveau 3. L'indicateur observé est un pourcentage et a été calculé en effectuant un ratio du PIB par habitant de 2005 par le PIB par habitant de 2000. Comme sa distribution est non gaussienne, le logiciel en prend automatiquement le logarithme. Il est valide en 2005.

Dans la zone 1 de l'interface, nous choisissons de visualiser l'évolution du PIB par habitant, en euros, sur l'union européenne élargie (l'espace ESPON 31) issu d'un jeu de données de test (QualESTIM demo). Ce jeu de données comporte une vingtaine de variables socio-économiques harmonisées dans la version 2006 de la nomenclature NUTS. L'interface est mise à jour automatiquement : une carte choroplèthe apparaît désormais dans la zone 2 de visualisation, et les unités sont colorées en fonction de la valeur d'évolution du PIB par habitant. L'Islande apparaît hachurée car sa valeur est manquante. Cette vue générale des données permet à l'utilisateur de se rendre compte de la distribution de cette évolution sur l'espace géographique, voir figure 7.6. Les informations concernant le jeu de données et l'indicateur étudié sont exposées à côté de la carte, dans la zone 4.

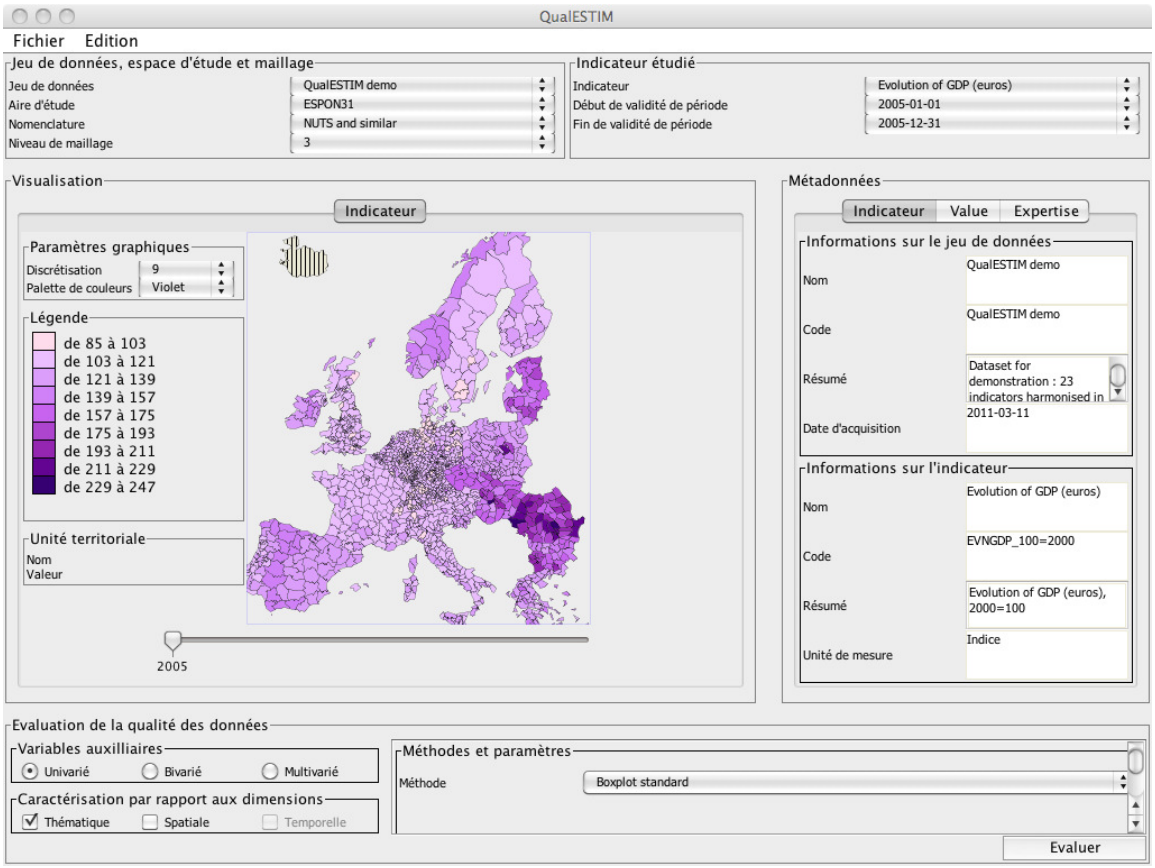


FIGURE 7.6 – Vue générale de la distribution de l'évolution du PIB par habitant entre 2000 et 2005.

Pour filtrer, et détecter des évolutions exceptionnelles (anormalement hautes ou basses), il est ensuite possible d'utiliser une des méthodes de détection de valeurs exceptionnelles qui ont été implémentées. Leur paramétrage se fait dans la zone 2, en bas de l'interface. L'utilisation de la méthode de la boîte à moustaches produit une carte de la distribution des valeurs loin des frontières, et dans les onglets supplémentaires sont affichés le diagramme de fréquence et la boîte à moustaches.

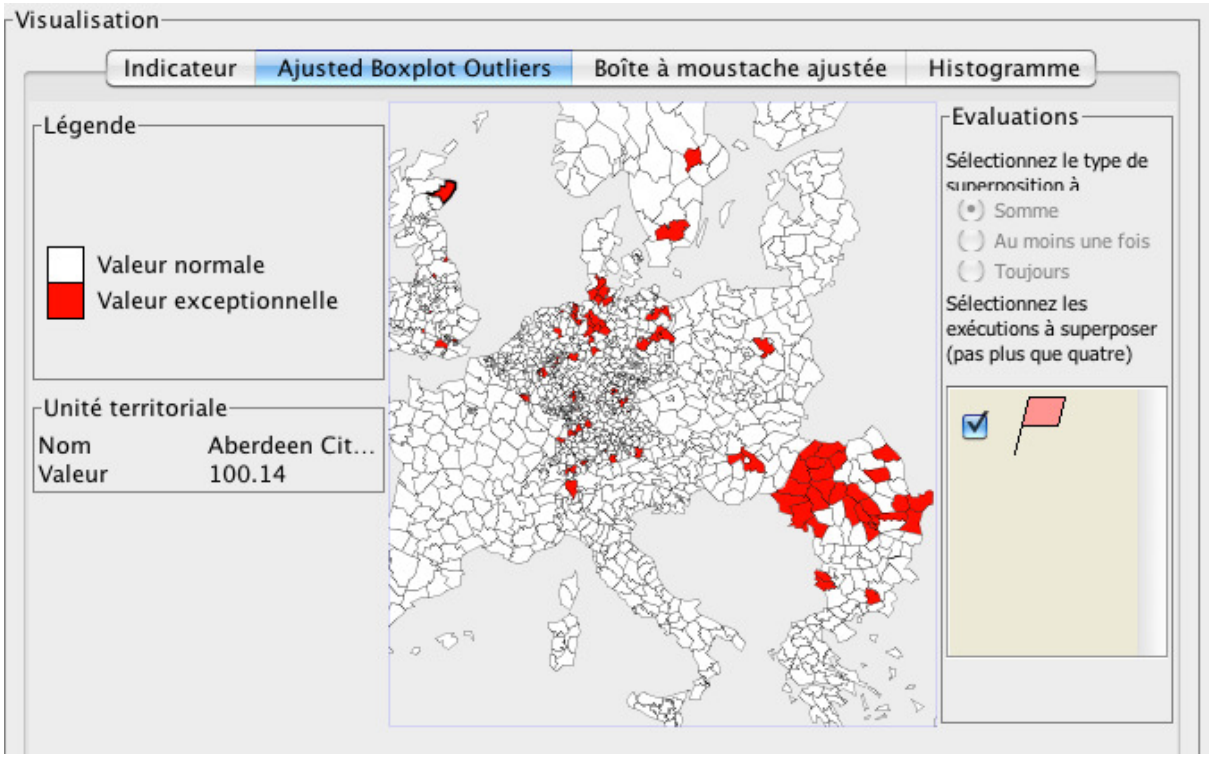


FIGURE 7.7 – Résultat d'analyse par la méthode du boxplot (thématique, univariée).

Par cette méthode, l'unité « Aberdeen City » apparaît comme exceptionnellement basse pour l'évolution de son PIB par rapport à l'ensemble de toutes les unités. Cependant, la méthode de la boîte à moustache (*boxplot*) indique des valeurs hautes ou basses, sans expliquer si ces valeurs sont exceptionnelles aussi par rapport à leurs voisines. Or, la carte des unités exceptionnelles que renvoie la méthode de Hawkins, paramétrée avec un test du chi2 paramétré à 3.84146 et une portée de 300 kilomètres montre qu' « Aberdeen City » n'est pas une valeur exceptionnelle par rapport à l'ensemble de ses voisines, voir figure 7.8. En revanche, elle indique que d'autres valeurs d'indicateur sont exceptionnelles, dont l'unité territoriale « Kyustendil » en Bulgarie.

Pour tirer des conclusions plus fermes, il est donc nécessaire de combiner les résultats de différentes méthodes. Nous employons ensuite la moyenne locale, la régression linéaire multiple, la régression locale et la régression géographiquement pondérée. Cette dernière méthode est utilisée avec deux variables auxiliaires, la typologie urbaine/rurale établie pour les unités géographiques de l'espace européen par le projet TIPTAP [Camagni 10], ainsi que la quantité de fonds structurels européens perçue par habitant par chaque unité à ce niveau, connue en 1999 (mais que nous avons daté à 2005 pour les besoins de l'étude).

Le choix des fonds structurels comme variable auxiliaire est thématiquement justifié par le fait que le PIB par habitant est le seul instrument de mesure de l'éligibilité des régions européennes à ces fonds. Ces fonds ont comme objectif d'améliorer la cohésion territoriale, c'est-à-dire, dans l'esprit de la com-

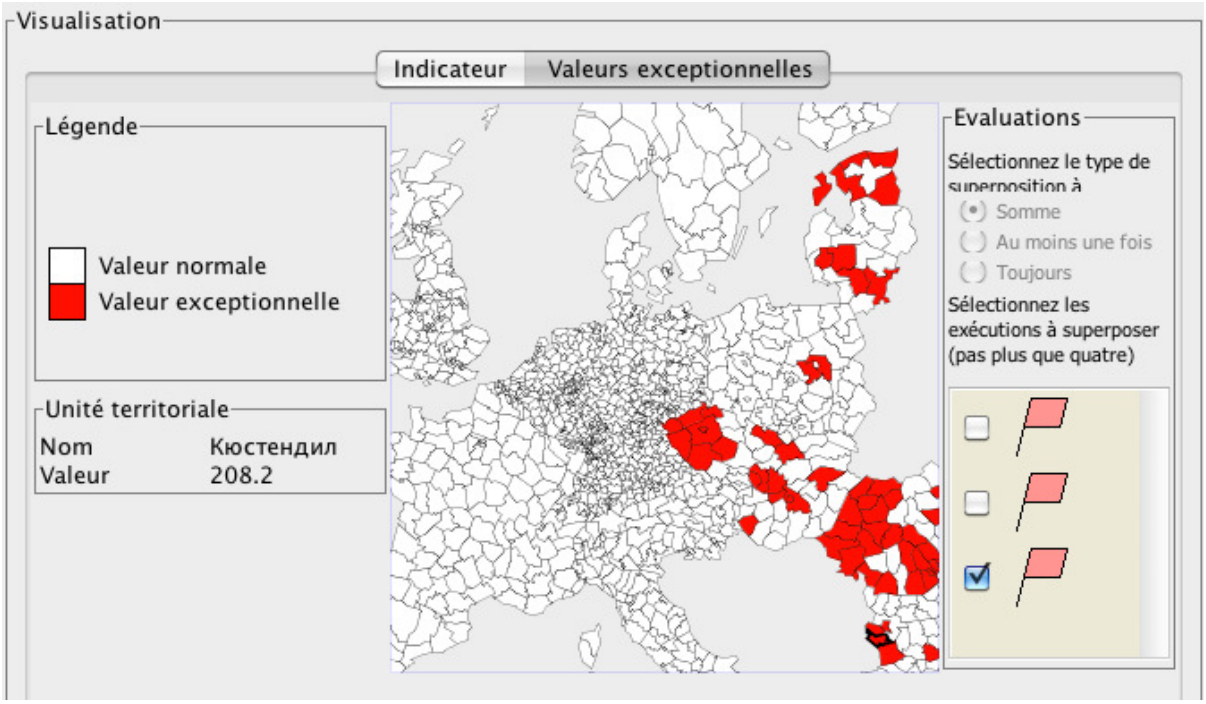


FIGURE 7.8 – Résultat d'analyse par la méthode de Hawkins (spatiale, univariée).

mission européenne, de réduire les écarts de richesse entre les régions européennes : ils sont versés en majorité aux régions dont le PIB par habitant est faible (bien que les règles d'attribution ne soient pas très claires) [Proisy 11]. Par ailleurs, le concept de « cohésion territoriale », bien qu'inscrit dans l'agenda de Lisbonne, n'est pas non plus très clair. Cependant, l'article 174 du traité sur le fonctionnement de l'Union Européenne définit en outre les régions auxquelles il convient d'apporter « une attention particulière » :

> Il s'agit des « zones rurales », de celles où s'opèrent « une transition industrielle » et des régions « qui souffrent de handicaps naturels ou démographiques graves et permanents telles que les régions les plus septentrionales à très faible densité de population et les régions insulaires, transfrontalières et de montagne. »

(voir l'article en ligne de Euractiv [4]). Or, la typologie urbaine/rurale que nous utilisons comme variable auxiliaire distingue les régions densément peuplées, des régions moins peuplées, dites rurales. Cette variable qualitative a été transformée en variable qualitative ordinale dont les rangs s'échelonnent de 1 à 6, 1 s'appliquant aux régions très densément peuplées, 6 pour au contraire les régions les moins densément peuplées.

Les résultats successifs de ces méthodes sont agrégés par la somme, comme montre la figure 7.9. Il apparait que les unités ayant vu leur PIB évoluer de façon exceptionnelle sont majoritairement dans le sud-est de l'Europe. En particulier, l'unité territoriale Kyustendil qui apparaît en rouge. Cette étape de filtrage permet à l'utilisateur de se poser des questions sur ces unités, du point de vue de la véracité de la donnée, comme du point de vue thématique. Cette région de l'Europe de l'Est a notamment bénéficié d'une forte croissance économique depuis le début des années 2000. Donc, il n'est pas si étrange que ces unités apparaissent comme exceptionnelles. Mais par ailleurs, si l'utilisateur s'intéresse à la source de ces données, en cliquant sur l'unité Kyustendil par exemple, il apprend que la valeur qui était initialement manquante est issue des estimations menées par le projet *ESPON 2013 Database*. De même, il peut

4. http://www.euractiv.fr/ue-cherche-definir-cohesion-territoriale-000062

vérifier qu'une majorité de valeurs exceptionnelles ne sont pas produites par Eurostat. En effet, la Bulgarie et la Roumanie ne sont devenues officiellement membres de l'Union Européenne qu'en 2006, date à partir de laquelle Eurostat a pu collecter des données économiques sur ces unités. La méthode d'estimation employée est peut-être trop optimiste. Celle-ci indique, si l'on reprend l'onglet « Valeur » des métadonnées, présenté dans la figure 7.10, que c'est une estimation basée sur la dernière valeur connue avant 2006 de la Bulgarie et le facteur supposé d'évolution temporelle.

En sa qualité d'expert, l'utilisateur peut alors réviser la fiabilité accordée à cette valeur, et par la suite réévaluer son niveau de qualité dans l'onglet « expertise ». Cette correction est automatiquement associée au rapport produit par les méthodes.

Cette étude met en évidence l'intérêt de notre proposition pour l'exploration et visualisation d'une base de données spatio-temporelles. L'expérimentation menée avec le prototype écrit en Java montre qu'il est possible d'interfacer un logiciel écrit dans un langage orienté-objet comme Java avec un environnement dédié au calcul scientifique, R. Il est ainsi envisageable d'étendre cette approche pour l'estimation de valeurs manquantes. L'objectif technique poursuivi était également d'améliorer la modularité des développements et de tirer parti de la richesse des développements statistiques menés dans l'environnement libre R. Cependant, cet objectif n'est pas entièrement atteint, car comme nous l'avons montré, le code exécutant les scripts R est pour l'instant peu générique. Par ailleurs, l'étude suivant la dimension temporelle n'est pas encore intégrée. Ces deux points font l'objet d'une discussion plus approfondie.

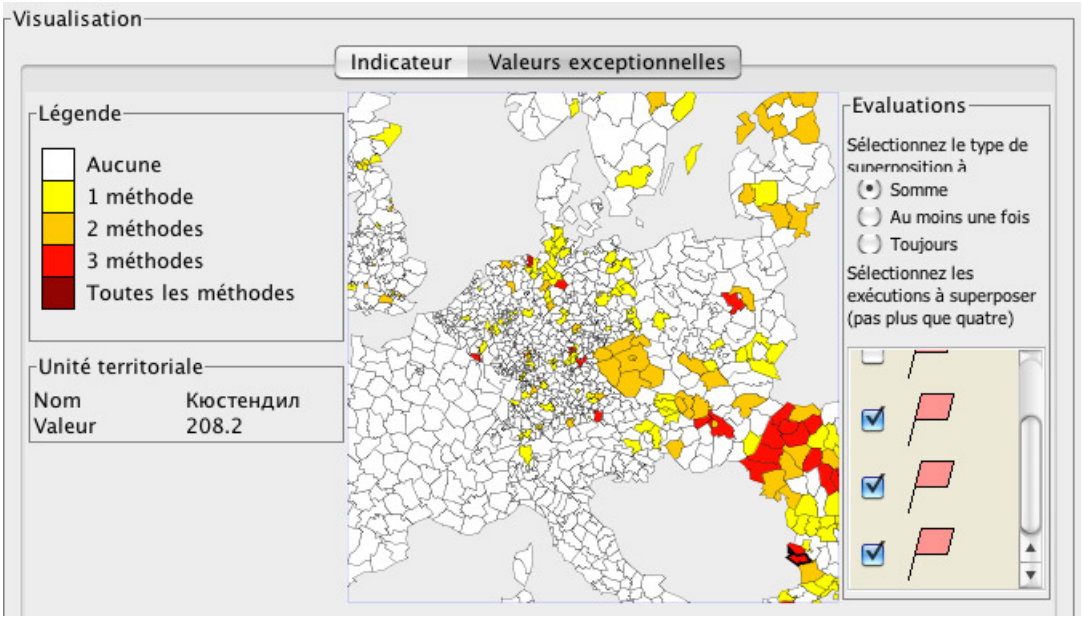


FIGURE 7.9 – Analyse combinée des valeurs exceptionnelles pour plusieurs méthodes dans Qualestim.

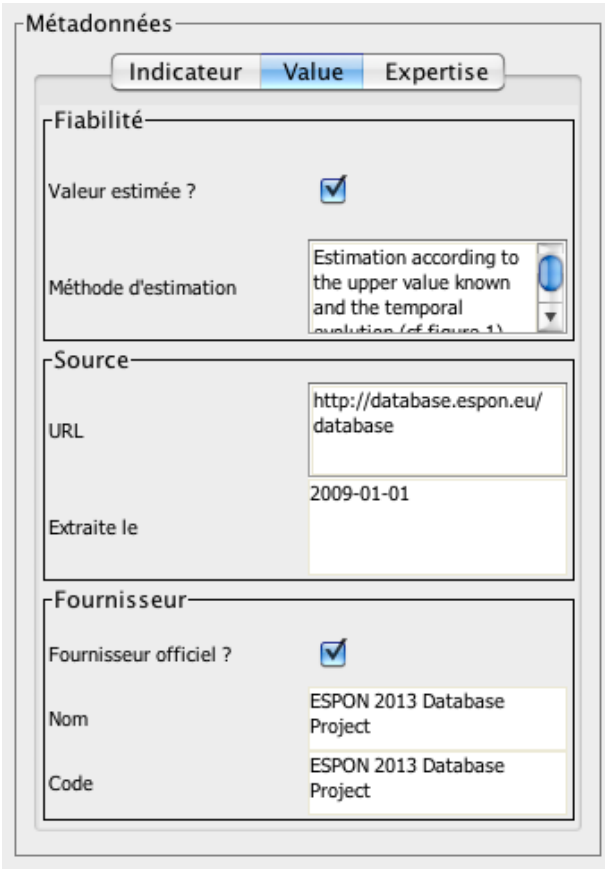


FIGURE 7.10 – Provenance de la valeur de l'unité Kyustendil.

7.1.3.3 Discussion

Nous proposons ici une discussion sur les questions et les problèmes que soulève une véritable exploration spatio-temporelle des données, ainsi que l'intégration de méthodes statistiques dont l'usage requiert un certain niveau de compétences du point de vue de l'utilisateur.

7.1.3.3.1 L'intégration de la dimension temporelle Dans le prototype que nous avons montré, l'étude est pour l'instant restreinte à une unique version de nomenclature, et bien que cette version fasse partie des options du logiciel modifiables par l'utilisateur, il lui est impossible de combiner des indicateurs connus dans d'autres versions. Par conséquent, les requêtes ne portent que sur les indicateurs disponibles dans la version de nomenclature configurée, ce qui réduit d'emblée l'étendue temporelle des requêtes. Ainsi, si 2006 est la version configurée pour l'étude, les indicateurs mesurés entre 1980 et 1990 en Europe, généralement associés à la version de maillage de 1980 ou 1988, ne sont pas disponibles. Par ailleurs, il est possible que des indicateurs soit disponibles pour la période d'étude spécifiée par l'utilisateur, par exemple [2000, 2005], mais dans une autre version de maillage. L'utilisateur ne bénéficie pas alors de la connaissance de ces indicateurs, qui existe au moins partiellement sur l'ensemble des unités appariées.

Il serait donc souhaitable au minimum de laisser voir à l'utilisateur s'il existe des indicateurs qui ont été créés pour le même espace d'étude et la même période temporelle, dans une autre version de maillage. Nous pensons aussi qu'il serait très intéressant de proposer l'activation « à la demande » de méthodes de transferts de certains indicateurs vers la version du maillage d'étude. Il faudrait alors distinguer les indicateurs « bruts » des indicateurs « transférés ». Ce point relève de la qualité et de l'objectivité de l'étude qui peut-être menée à partir de ces indicateurs harmonisés « à la volée » car il s'agit d'indiquer à l'utilisateur qu'avec des indicateurs « transférés », il travaille avec des données qui sont estimées.

Un autre point de discussion concerne les conditions d'interrogation du modèle : doit-on prendre des indicateurs valables uniquement à la même période, ou doit-on introduire plus de liberté dans le choix des variables. Par exemple, l'étude de l'évolution du PIB entre 2000 et 2005, à valeur de référence en 2000, est intéressante en conjonction avec des données auxiliaires comme le niveau d'éducation de la population, qui peut n'être connu qu'en 1999. La notion de *voisinage temporel « tolérable »* pour croiser des données socio-économiques est importante. En effet, les variables démographiques peuvent être utilisées avec une large plage de tolérance (une dizaine d'années par exemple), car les évolutions de ce type de variable sont lentes, de l'ordre d'une génération (20 ans). A l'opposé, le prix moyen du baril du pétrole, qui varie d'une semaine à l'autre dans des proportions importantes, devrait faire l'objet d'une restriction à un mois, par exemple. Ici, la connaissance de *l'inertie temporelle* d'une variable, c'est-à-dire la durée moyenne séparant des évolutions significatives, est essentielle. En supposant que la fréquence de mesure des indicateurs soit étroitement liée à cette inertie temporelle, il serait envisageable d'utiliser la fréquence de mesure pour déterminer une plage de tolérance temporelle pour chaque indicateur.

L'intégration de la dimension temporelle n'est donc pas simplement une question de requête spatio-temporelle, qui se résout assez facilement sur le plan technique. Il s'agit aussi de réfléchir en termes de voisinages temporels, et de combinaison de variables ayant une fréquence de mesure et une inertie différentes. Il est nécessaire de mener une réflexion sur les échelles temporelles, afin de spécifier ce qui est comparable et à quel rythme dans le temps.

7.1.3.3.2 L'intégration de méthodes statistiques L'intégration de méthodes statistiques dans un outil d'exploration spatio-temporelle pose des problèmes, tant sur le plan logiciel (du point de vue de la modularité, flexibilité et du coût de développement d'un tel logiciel) que du point de vue de l'utilisabilité de ces méthodes.

Sur le plan de l'architecture, l'interface proposée n'est pas aujourd'hui capable d'intégrer facilement de nouvelles méthodes : l'implantation d'une nouvelle méthode devrait faire l'objet de programmation, et l'espace réservé actuellement au paramétrage des méthodes est trop restreint dans l'interface. Ce manque de flexibilité existe pour d'autres environnements qui cherchent à intégrer des scripts R à leur plate-forme (SpatialAnalyst [5] de ESRI, ou *Marine Geospatial Ecology Tools* (MGET) [Roberts 10]). Ainsi, les coûts de développement sont proportionnels au nombre de méthodes intégrées. Si une description standardisée des scripts existait, il serait possible et moins couteux de générer à la volée autant de panneaux de configuration que nécessaire pour chaque méthode, sans surcoût de développement. Fondamentalement, il manque en général un méta-modèle des scripts R qui puisse permettre d'intégrer facilement de nouveaux scripts : si les scripts étaient munis d'une entête décrivant les paramètres d'entrée comme de sortie, spécifiant leur type et les valeurs possibles, ainsi qu'une description de l'objectif de la méthode, on pourrait alors envisager de brancher un nombre infini de scripts dans un mode interactif.

Sur le plan de l'utilisabilité de ces méthodes, la conception de systèmes destinés à aider des utilisateurs à activer des méthodes qui sont basées sur des hypothèses compliquées, peut-être hors du champ de leurs connaissances, pose problème. La question qui se pose est celle de l'adéquation entre le niveau d'expertise de l'utilisateur et celui que requiert l'emploi de ces méthodes, et l'écart que le logiciel est censé combler. Soit le logiciel est muni d'une aide en ligne très évoluée, permettant de comprendre tous les paramètres dans le détail, soit l'approche est simplifiée au maximum pour ne laisser à paramétrer qu'un nombre minimal d'éléments par l'utilisateur. La conception d'une interface intuitive permettant d'intégrer les paramètres des méthodes est un problème difficile, qui fait l'objet d'études approfondies dans le domaine de la didactique des sciences et de la mise à disposition d'outils pour le grand public [Lanter 91], [Bastien 01]. Ainsi, dans le prototype que nous avons présenté, la méthode de régression par pondération géographique (GWR) n'est paramétrée que par sa portée (le *bandwidth*), et l'utilisateur peut spécifier des variables explicatives supplémentaires. Cependant, la spécification du modèle complet de la GWR requiert en plus la définition du modèle de diffusion (bi-carré, gaussien, linéaire, ou autre). Nous avons souhaité présenter une interface simplifiée dans laquelle la GWR s'exécute à partir d'un modèle gaussien, qui ne peut pas être changé par l'utilisateur. Pour simplifier au maximum, l'utilisateur aurait pu ne même pas avoir à configurer la portée puisqu'il existe des tests statistiques en vue de déterminer la « meilleure » portée. Cependant ces tests sont coûteux en temps de calcul, et l'interactivité du logiciel est alors mise en danger. Il s'agit donc de réaliser deux compromis, le premier entre la finesse du paramétrage et celui de l'expertise attendue des utilisateurs, et le second entre l'automatisation de ce paramétrage et les performances attendues pour l'interactivité.

La construction d'une ontologie des méthodes statistiques par classification constituerait un progrès pour ce problème. En décrivant les méthodes, cette ontologie permettrait d'une part, de guider l'utilisateur dans le choix des modèles de distribution comme dans le choix des méthodes, et, d'autre part, faciliterait l'intégration de ces méthodes dans le logiciel, rendant adaptable à divers degrés d'expertise la méthode. Par exemple, la GWR pourrait être complètement décrite (par le modèle de diffusion, et sa portée), et l'usage d'un système de filtres ne proposerait que les paramètres en adéquation avec le rôle des utilisateurs. En instanciant cette ontologie sous la forme d'un méta-modèle décrivant les scripts R, il serait également plus facile de rendre flexible les logiciels souhaitant intégrer ces méthodes.

5. `http://www.esrifrance.fr/Spatial_Analyst.asp`

7.2 Une analyse multi-scalaire contextualisée

L'analyse que nous proposons ici repose sur l'exploitation de la connaissance des évolutions de la hiérarchie territoriale afin de mettre en lumière des processus d'évolution statistique contextualisés. En particulier, cette analyse permettrait de distinguer les évolutions exceptionnelles de celles qui seraient liées à des changements territoriaux. Dans cette proposition, le concept des cartes d'écarts territoriaux sur lequel repose HyperAtlas, présenté dans la section 4.3.2 page 132, est étendu en prenant en compte les modifications des relations d'appartenance des unités géographiques au cours du temps. En effet, l'organisation hiérarchique des territoires qui change a des conséquences sur les évolutions des variables statistiques associées, et cet aspect a jusqu'ici été négligé dans les outils existants.

Nous allons prendre un exemple pour illustrer en quoi la connaissance des modifications de relations d'appartenance peut être utile à l'analyse du changement, mis en rapport à la connaissance que nous avons de l'information statistique. Nous prenons ici l'exemple de l'Allemagne, et étudions le profil de l'unité régionale de Hambourg, codé 'DE6', en la comparant à son contexte territorial (dans la carte d'écart territorial). Hambourg est aujourd'hui une des régions les plus riches de l'Allemagne, et son PIB est exceptionnellement élevé par rapport à la moyenne nationale allemande, comme l'illustre la carte d'écart territorial d'HyperAtlas, figure 7.11.

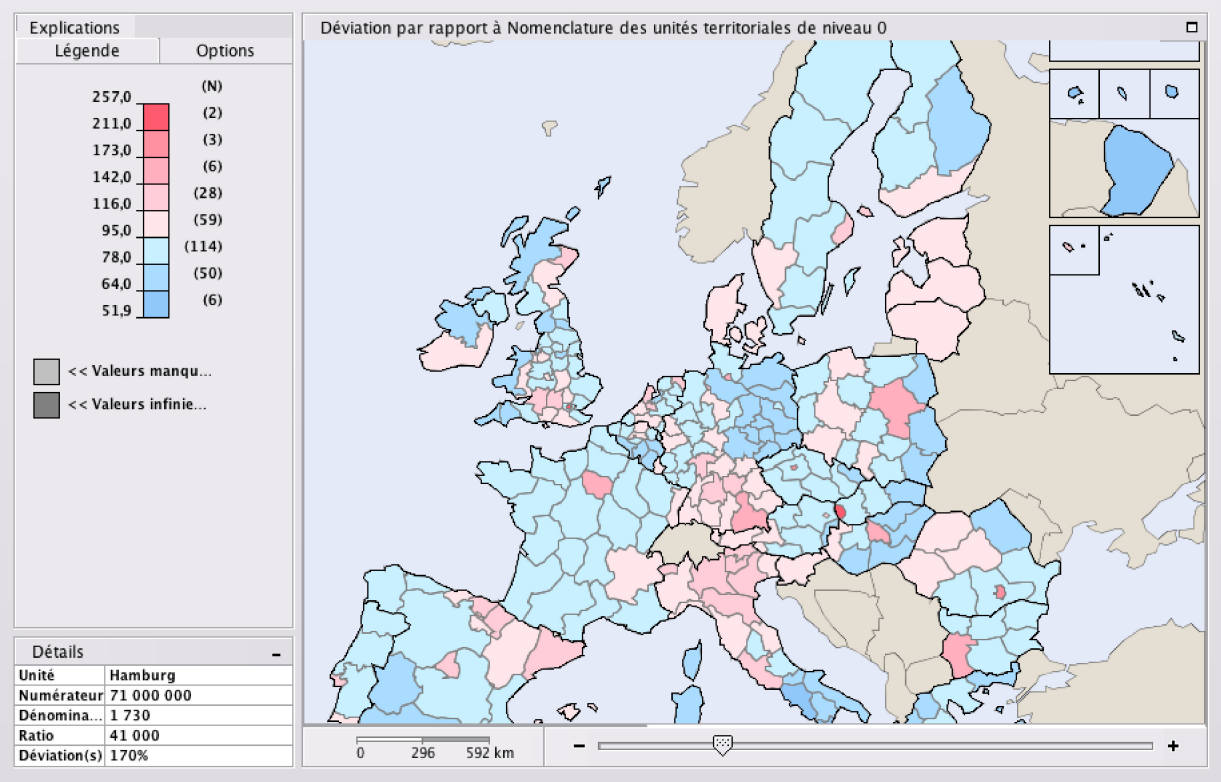


FIGURE 7.11 – Position de Hambourg pour son PIB par habitant en 2005, par rapport à son pays, dans HyperAtlas.

Nous souhaiterions découvrir comment cette richesse a évolué entre 1987 et 1996 relativement à un contexte territorial qui a changé. En effet, entre 1987 et 1990, Hambourg faisait partie de la République Fédérale d'Allemagne (RFA), et depuis la réunification allemande en 1990, Hambourg fait partie de

l'Allemagne réunifiée. Dans une analyse classique, l'utilisateur se verrait obligé de choisir un des deux contextes, et de s'y tenir pour analyser l'évolution de la variable d'intérêt (le PIB par habitant) sur toute la période. Or, aucun de ces deux contextes n'est vraiment satisfaisant, car ils ne reflètent que partiellement l'évolution de la position relative de l'unité.

Via notre modèle, il devient possible d'analyser le développement de l'unité 'DE6' suivant trois différents contextes, comme l'illustre la figure 7.12. La courbe étiquetée 'RFA' représente l'évolution de l'écart du PIB par habitant de Hambourg par rapport à l'ensemble des unités qui appartenaient à la RFA, entre 1983 et 1996. La courbe étiquetée 'Allemagne réunifiée' représente cette même évolution mais rapportée à l'ensemble de toutes les unités qui font aujourd'hui partie de l'Allemagne réunifiée (entre 1983 et 1996). La dernière courbe en pointillés représente l'évolution de l'unité par rapport à un contexte qui a changé, qui était avant 1990 la RFA, puis est devenu l'Allemagne réunifiée.

Depuis 1983, Hambourg a toujours été une des régions les plus riches d'Allemagne, que ce soit l'Allemagne réunifiée ou la RFA simplement. Cependant, sa richesse relative, si l'on s'en tient au contexte de la RFA, n'a pas évolué significativement, en dehors d'un léger déclin entre 1985 et 1990 avant une reprise de l'augmentation de l'enrichissement comparativement au reste de l'Allemagne. Par contre, les deux courbes qui tiennent compte de la réunification montrent que l'écart entre le niveau de richesse de Hambourg est celui de son contexte (l'Allemagne réunifiée) est devenu exceptionnellement élevé en 1991. En effet, la réunification allemande visait à absorber un pays (la République Démocratique Allemande, RDA), alors que son économie n'était pas adaptée au système capitaliste, et d'importantes restructuration des outils de production, comme du système financier devaient être menées pour ramener la RDA au niveau de richesse de la RFA. En 1990, le PIB de la RDA valait moins de la moitié de celui de la RFA. Donc en réalité, l'augmentation soudaine de richesse relative de Hambourg observée en 1991 ne correspond pas à un accroissement significatif de la richesse de Hambourg, mais plutôt à l'événement de Fusion entre la RFA et la RDA, qui par un pur effet mécanique modifie la position relative de Hambourg.

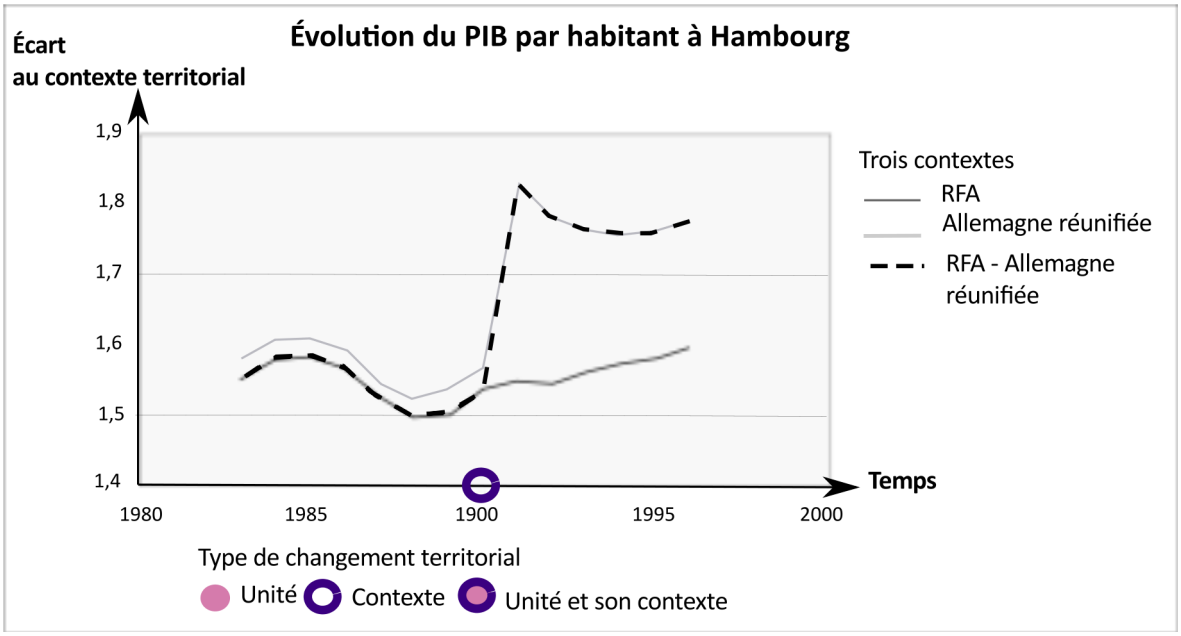


FIGURE 7.12 – Évolution différenciée de Hambourg suivant trois contextes spatio-temporels.

Cet exemple montre l'importance de mettre en évidence ces phénomènes "mécaniques" induits par les modifications territoriales. Pour mettre en oeuvre de façon systématique ce type d'analyse, il faut encore remplacer le modèle de données d'HyperAtlas par notre modèle multi-scalaire évolutif.

7.3 Conclusion

Dans ce chapitre, nous avons proposé quelques idées pour une analyse spatio-temporelle contextualisée de l'information statistique territoriale et de sa qualité. En considérant d'abord les métadonnées comme un apport important à la compréhension des données, nous avons montré la perspective nouvelle que ces informations pouvaient apporter à l'analyse de l'information, en prenant comme exemple l'évaluation de la qualité des données par recherche de valeurs exceptionnelles. Par ailleurs, avec le prototype que nous avons réalisé, *QualESTIM*, nous avons voulu expérimenter le couplage faible de codes scientifiques statistiques comme R avec java, un langage dédié à la programmation orientée objet, dédié à la réalisation de toute l'interface, interactive et dynamique. Cette expérience, qui montre la faisabilité de cette approche, fournit aussi des arguments pour l'établissement d'une ontologie des méthodes statistiques, qui permettrait de décrire et de piloter de façon plus générique les méthodes statistiques.

Par ailleurs, l'analyse proposée n'intègre pas encore véritablement la dimension temporelle, et nous posons certaines questions d'ordre théorique qui doivent être résolues avant d'entreprendre le développement d'un outil d'analyse spatio-temporelle. Il s'agit par exemple de définir en priorité la notion de voisinage temporel pour une variable statistique, et de déterminer l'inertie temporelle d'une variable, afin de pouvoir croiser des variables ayant des dates de validité différentes. Il manque également un outil pour le transfert de variables issues de versions de nomenclatures différentes afin de réaliser de façon systématique des études de variables harmonisées dans le maillage choisi par l'utilisateur. Sur le plan de la réalisation, nous n'avons pas intégré la possibilité d'étudier des taux de variation pour des variables de stocks. Les seuls taux de variations qui peuvent être étudiés sont ceux déjà présents dans la base de données. Ce type d'analyse nécessite en effet de filtrer les variables de stocks des variables de taux, chose qui pour l'instant nécessite une analyse fine des métadonnées, car il n'a pas été prévu de champ de type booléen dans les métadonnées distinguant les stocks des taux.

Par rapport à ces questions ou ces besoins, nous n'avons pas de réponse immédiate, mais nous avons eu également envie de montrer comment un modèle tenant compte des changements d'organisation hiérarchiques et des évènements territoriaux pouvait contribuer à la compréhension du changement. Pour cela, un exemple très simple illustre l'avantage qu'HyperAtlas aurait à évoluer pour réaliser une analyse du changement via des courbes d'écart relatif où le contexte territorial pourrait être choisi par l'utilisateur. Cette propositions souligne l'importance du contexte pour l'analyse de l'évolution de l'information statistique territoriale.

Troisième partie

Bilan et perspectives

Chapitre 8

Conclusion et perspectives

Pour conclure, nous allons résumer les contributions majeures de cette thèse et dégager les perspectives principales

8.1 Conclusion

Trois apports essentiels nous semblent découler de cette thèse. Le premier concerne la modélisation du support spatial de l'information statistique territoriale, prenant en compte son aspect hiérarchique et évolutif. Le second porte sur la définition d'un modèle de métadonnées pour l'information statistique territoriale, et des propositions pour la gestion (acquisition, stockage, diffusion) de ces métadonnées. Le dernier relève de l'intégration de méthodes statistiques pour l'analyse des données à une plate-forme qui repose sur le modèle de données proposé.

Ces résultats ont, pour une partie, été implémentés dans le projet *ESPON 2013 database*, qui propose une plate-forme complète avec, d'une part, une base de données PostgreSQL avec cartouche spatiale PostGIS, et, d'autre part, un éditeur de données Web basé sur le profil *esponMD* de la norme ISO 19115 que nous avons défini. Un prototype intégrant des méthodes statistiques écrites avec R interrogeant cette base de données a également été réalisé.

8.1.1 Gestion de hiérarchies multiples et évolutives

Partant des résultats d'anciens travaux de recherche qui promeuvent, d'une part la gestion du changement via des objets possédant une identité et une vie, indépendamment de leurs évolutions [Cheylan 93, Lardon 99, Worboys 98, Wachowicz 99], et, d'autre part, l'intégration des évènements et des processus de transformation [Claramunt 95, Wachowicz 99, Spery 01, Worboys 05] pour conférer un caractère explicatif aux modèles, nous avons étendu sur deux points ces approches.

Le premier point concerne la gestion de relations hiérarchiques entre unités d'un zonage, avec la modélisation de leurs multiples appartenances, et du changement de ces appartenances aux cours du temps. Ce travail est à mettre en relation avec les travaux menés dans le domaine des entrepôts de données

et de l'analyse multi-dimensionnelle [Pedersen 01, Tchounikine 05].

Le second point porte sur la création d'une méthode de mise à jour de ce modèle permettant d'assurer à la fois l'identification semi-automatique des unités géographiques à travers les différentes versions de nomenclature, et l'acquisition des évènements territoriaux dans le système. Cette contribution s'est appuyée sur les travaux réalisés dans le cadre d'appariement de bases de données hétérogènes, [Devogele 97, Raimond 07, Olteanu-Raimond 09].

Enfin, nous avons proposé un outil d'exploration interactive du changement territorial qui exploite les possibilités offertes par ce modèle et facilite l'analyse des modifications territoriales. Ce type d'outil offre des perspectives nouvelles à la compréhension du changement car l'expert est alors en mesure de bénéficier d'un niveau de lecture plus élevé du changement, d'établir des statistiques sur sa localisation dans le temps et l'espace, et de mettre ces statistiques en relation avec la connaissance du contexte historique et géographique qu'il détient.

8.1.2 Gestion de métadonnées pour l'information statistique territoriale

La nécessité de disposer de métadonnées afin de rendre compte du lignage complexe des informations collectées de sources hétérogènes, [McCarthy 82], [UN/ECE 95], [Kent 97] a motivé notre étude des standards disponibles pour la description de l'information statistique territoriale. Constatant soit leur inadéquation partielle à la structure composite de cette information, soit leur difficulté à être mis en oeuvre, nous avons proposé l'adaptation d'un de ces standards, la norme ISO 19115, dans un profil, le profil *esponMD*. Celui-ci permet, par l'entremise d'un éditeur dédié associé à un système de stockage de ces informations, d'assurer l'acquisition et le traitement d'un minimum d'information sur la qualité des données. Un premier pas vers l'interopérabilité vers le standard émergeant (SDMX) a également été franchi avec la traduction de notre profil dans le modèle SDMX.

8.1.3 Exploration et analyse interactive et contextualisée de l'information

Le modèle que nous avons proposé (gestion des supports et des métadonnées) a ensuite été intégré dans une plate-forme conçue pour l'analyse exploratoire des données, à références spatiales ou temporelles. Cette plate-forme est dédiée à l'analyse de la précision sémantique de l'information statistique territoriale par la recherche de valeurs exceptionnelles. Des travaux dans ce domaine, que ce soit par la fouille de données spatiale [Zeitouni 00], ou l'analyse exploratoire spatiale [Anselin 93, Rousseeuw 96], ont déjà proposé une batterie d'outils statistiques, qui peuvent être utilisés pour la recherche de valeurs exceptionnelles et l'évaluation de la qualité des données. Cependant, ces outils ne mettent pas en relation ces analyses ni avec métadonnées, ni avec le contexte territorial évolutif. Nous avons donc proposé une interface graphique intégrée à la plate-forme de données, programmée en Java, et facilitant l'exploration interactive *conjointe* des données et des métadonnées. De plus, pour cette analyse, qui suppose l'intégration de méthodes statistiques parfois avancées, nous avons couplé R avec Java, et mis en lumière les difficultés inhérentes à ce type de couplage. Enfin, nous avons suggéré de prendre en compte les changements d'appartenance des unités au cours du temps pour mieux juger du caractère exceptionnel de leur évolution.

8.2 Perspectives

Nos travaux ouvrent plusieurs voies pour la recherche sur l'information statistique territoriale, mais pourraient également trouver des applications pour le traitement d'autres données, à références spatiales et temporelles. Nous ne décrivons pas ici absolument toutes les perspectives que nous avons pu dégager au long des propositions, comme par exemple la nécessité d'approfondir la réflexion sur la notion d'échelle et de voisinage temporel, ou bien celle d'établir une ontologie des méthodes statistiques. Nous en développons seulement quatre, soit parce qu'elles sont la poursuite de cette thèse pour son amélioration, soit parce qu'elles sont transversales à ce travail et font écho à l'objectif initial de notre thèse.

8.2.1 Gestion de l'incertitude sur les évènements

Les évènements que nous avons modélisés ne sont que des évènements territoriaux, reconnus par un invariant géométrique. Il est donc nécessaire de disposer des géométries exactes pour procéder à l'inférence de ces événements. Or, dans le cas le plus général, que ce soit pour l'information statistique territoriale ou l'information géographique en général, ces géométries ne sont pas forcément disponibles. Deux cas se présentent :

– elles ne sont connues que de manière approximative,
– elles n'existent pas, mais d'autres documents apportent des informations de localisation relative.

Dans le premier cas, les travaux basés sur la logique des ensembles flous appliquées aux relations spatiales, [Zhan 98], [Alfred 04] pourraient servir de base au calcul des évènements territoriaux, intégrant l'incertitude sur les frontières des objets spatiaux. Il serait ainsi possible d'évaluer des marges de probabilité de réalisation de ces évènements.

Dans le second cas, on peut imaginer d'exploiter ces documents pour extraire et modéliser les informations sur la position relative des objets, et utiliser les travaux sur le calcul et l'utilisation de relations projectives [Clementini 08] pour raisonner sur cette information. Prenons, par exemple, la création d'un territoire qui serait identifié par un lieu ou un bâtiment emblématique. Sans connaitre les limites de ce territoire, ni la position exacte de cet emblème, l'analyse de documents historiques nous apprendrait que cette position est à l'Ouest d'une rivière, sur le trajet d'une route de commerce, au pied d'une montagne connue par son nom. On aurait alors la possibilité de positionner ce territoire entre ceux déjà existants, et d'imaginer un évènement entre ces territoires. Dans ce cas, l'invariant pourrait plutôt porter sur la position relative de chacune des unités territoriales. Par exemple, dans le cas de la création de Chamrousse, qu'illustre la figure 5.12 page 156, si Saint Martin d'Uriage et Vaulnaveys-le-Haut sont connues pour être voisines (par contiguïté d'ordre 1) à l'Ouest de Livet-et-Gavet, le fait que Chamrousse s'insère par la suite entre ces trois unités peut amener à émettre la conjecture que Saint Martin d'Uriage, Vaulnaveys-le-Haut et Livet-et-Gavet sont impliquées dans l'évènement territorial conduisant à la création de Chamrousse. Ensuite, d'autres informations pourraient apporter des preuves que Livet-et-Gavet n'est pas impliquée dans cet évènement, bien que l'étude soit délicate. En effet, la population de Livet-et-Gavet passe de 1853 à 1447 habitants entre les recensements de 1982 et 1990. Ces chiffres, qui pourraient aussi laisser croire que la population a diminué en raison d'échanges de surfaces avec les communes avoisinantes, correspondent en réalité à une baisse régulière de sa population depuis la fin de la seconde guerre mondiale. Dans ce type d'étude, les travaux sur la modélisation de l'incertitude de l'information spatiale [Goodchild 00, Dupin de Saint-Cyr 08, Jeansoulin 11] seraient d'un apport certain.

Enfin, du point de vue des métadonnées sur les évènements, nous avons été très peu diserts : un modèle, un standard peut également être recherché en vue de structurer au mieux ces informations. Le modèle pourra être adapté aux différents profils d'utilisation.

8.2.2 Aller plus loin que les métadonnées

Concernant le problème de l'interopérabilité sémantique, il nous semble crucial d'établir une ontologie statistique, permettant d'apparier les indicateurs issus de sources hétérogènes, afin de pouvoir raisonner sur la cohérence globale des données et non plus seulement sur la cohérence locale d'un jeu de données. Egalement, nous imaginons que l'intégration dans l'éditeur de métadonnées d'un système permettant de décrire de façon plus structurée les processus de transformation serait possible.

8.2.2.1 Calcul d'une ontologie de domaine

Les approches basées sur la construction d'ontologies depuis un corpus d'information pourraient être appliquées à l'ensemble des métadonnées collectées sur les indicateurs statistiques, pour l'analyse de similarité entre indicateurs, mais également pour l'alignement des catégories associées aux valeurs des indicateurs, à la condition que les textes descriptifs produits soient suffisamment longs. Il est ainsi possible d'établir une ontologie de concepts associés aux indicateurs, chaque indicateur formant une instance d'une classe d'indicateur (symbolisée par « is_a »). Chaque classe serait liée à d'autres par des relations de subsomption, « sub_class_of », ou des relations d'appartenance mais également une relation sémantique de domaine indiquant dans quelle mesure une catégorie recouvre l'autre, telle que « intersects », valuée par l'indice de recouvrement. Par exemple, les concepts de population au chômage, population active et population peuvent être identifiés via l'analyse des définitions des indicateurs statistiques, qui sont des instances de cette ontologie. Il apparaît que le nombre de chômeurs défini et mesuré par l'INSEE ou l'EUROSTAT se rapporte à un concept de « population_chômage » identique, et qu'un chômeur est un membre de la population active, qui elle-même fait partie de la population totale :

- *population_chômage_INSEE* is-a *population_chômage*
- *population_chômage_Eurostat* is-a *population_chômage*
- *population_chômage* sub_class_of *population-active*
- *population-active* sub_class_of *population*

Concernant les catégories, l'ontologie pourrait servir à modéliser la relation d'inclusion totale entre les personnes agées de 55 à 64 ans et les personnes agées de 50 à 64 ans, ou bien la relation d'intersection entre la catégorie "industrie de pêche, agriculture et exploitation forestière", notée *ABS.A* et la catégorie "activités d'agriculture, chasse et exploitation forestière", notée *SDMX.AYA* :

- *Age_55-64* sub_class_of *Age_50-64*
- *ABS.A* (intersects, 2/3) *SDMX.AYA*

L'opérateur (intersects, α) s'appliquant à deux classes A et B est une suggestion qui repose sur une extension au domaine des ontologies de la distance surfacique : si A inclut $a_1, a_2, ..., a_n$ concepts, et B inclut $b_1, b_2, ..., b_p$ concepts, alors

$$\alpha = \frac{\text{nombre de concepts en communs}}{n+p}.$$

L'analyse des métadonnées peut donc servir à établir, puis peupler une ontologie. Par rapport à l'emploi d'une taxonomie établie à l'avance pour regrouper les indicateurs par thème, l'établissement d'une telle

ontologie présente les avantages suivants :

- elle précise de façon nettement plus fine les concepts et les relations entre eux,
- elle peut aussi intégrer les différentes classifications utilisées pour indexer les données, que ce soit celles de thésaurus partagés comme celui que propose SDMX [1], mais également les classifications nationales d'indicateurs statistiques.
- elle est plus souple qu'une taxonomie établie à l'avance car elle est révisée à chaque nouvelle insertion de données dans le système d'information.

[Pattuelli 03] montre que l'existence d'une ontologie statistique facilite la compréhension des termes statistiques et la création d'un glossaire interactif.

8.2.2.2 Retranscription d'un lignage fin

Il s'agit ensuite de pouvoir raisonner sur les valeurs au niveau de leur transformation (modalités de calcul, réajustement, estimation). L'usage d'un graphe de flots de données semble adapté à la problématique du lignage, mais si la description des transformations nécessite la contribution d'un utilisateur, il faut absolument proposer des outils pour faciliter cette saisie. L'interface graphique pourrait, par exemple, être enrichie avec un dictionnaire des fonctions de transformation, usuelles dans le domaine des statistiques, et un dictionnaire des données disponibles qui sont utilisées comme entrées de la transformation. Ainsi, on simplifierait la saisie du type et du nom de l'indicateur par un simple « *drag and drop* » depuis une liste d'indicateurs recensés dans l'ontologie. L'ontologie des indicateurs précédemment évoquée est réutilisable pour retrouver rapidement l'instance d'un indicateur utilisé dans la formule comme ingrédient, comme ceci a déjà été souligné dans le travail de [Brilhante 06]. Il s'agit aussi d'établir le dictionnaire des transformations, le plus souvent des opérations mathématiques, qui peuvent être simples comme la division, ou plus complexes comme une formule de normalisation. La saisie de ces formules peut être envisagée à l'aide de langages comme MathML [2], (pour Mathématical Markup Language), ou bien OpenMath [3]. Par exemple, la formule de l'indicateur synthétique de vieillissement pourrait s'écrire (en utilisant Amaya [4] comme éditeur de formule WYSIWYG (*What You See Is What You Get*) de MathML) comme dans la figure 8.1 :

FIGURE 8.1 – Formule d'un indicateur composite, exprimée en MathML, en utilisant l'éditeur Amaya.

1. http://sdmx.org/wp-content/uploads/2009/01/03_sdmx_cog_annex_3_smd_2009.pdf
2. http://www.w3.org/TR/MathML3/
3. http://www.openmath.org/documents/bibliography.html
4. http://www.w3.org/Amaya/Overview.html

Ces formules peuvent être rattachées aux indicateurs dans l'ontologie, ou à leurs instances. Par exemple, les formules d'ajustement des variations saisonnières pourraient être accolées aux instances de l'indicateur *chômage* : *chômeur_INSEE* et *chômeur_Eurostat*.

8.2.3 La simulation de remembrements territoriaux

Il nous semble qu'un des cas d'application les plus intéressants du modèle des hiérarchies multiples et évolutives que nous avons proposé serait la simulation de remembrements territoriaux à des fins prospectives (que ce soit économiques, électorales ou environnementales).

En effet, l'objectif du travail sur les intercommunalités [Plumejeaud 09a], mené avec Guillaume Vergnaud, était de visualiser les valeurs de variables socio-économiques et démographiques dans les nouveaux zonages que forment les communautés de communes (ou EPCI) et les territoires de projet (ou Pays), afin d'analyser leur potentiel économique en comparaison avec le regroupement historique des communes en canton. Or, ces variables n'étaient connues que dans le maillage des communes. Grâce à notre modèle d'agrégation qui interdit les doubles comptes, il est aujourd'hui possible de reconstituer les valeurs dans les EPCI et les Pays. Si ce modèle était intégré dans HyperAtlas (ou un outil reprenant les principes de l'analyse des écarts), différents types d'écarts pourraient être analysés : l'écart d'une commune rapportée à sa communauté de communes ou bien à son canton, l'écart entre les cantons et les pays, ou les cantons et les communautés de communes.

Toutefois, une réflexion doit être menée sur les modalités de construction des cartes d'écarts territoriaux, car, comme le montre la figure 5.5 page 152, ce nouveau modèle d'agrégation n'indique pas à quelle unité supérieure l'EPCI e_4 doit être comparée pour les cartes d'écarts territoriaux. De même, si on imagine que plus d'une hiérarchie agrégative pourrait être analysée simultanément dans HyperAtlas, le concept des cartes d'écart territoriaux devrait être re-défini car il n'existe pas forcément de relations d'appartenance entre les unités d'une hiérarchie (les cantons par exemple) et une autre (les Pays par exemple). Par exemple, si une EPCI couvre plusieurs Pays, comme dans la figure 8.2, il existe plusieurs options pour choisir l'unité de comparaison :

- le pays qui comprend le plus de communes de l'EPCI : p_4 dans notre exemple,
- le pays couvrant la plus grande surface de l'EPCI : p_3 dans notre exemple,
- le pays dont la part de population (ou autre variable) est la plus importante dans l'EPCI,
- proposer un choix à l'utilisateur, complètement libre, ou bien limiter ce choix aux les unités qui couvrent complètement ou partiellement l'EPCI.

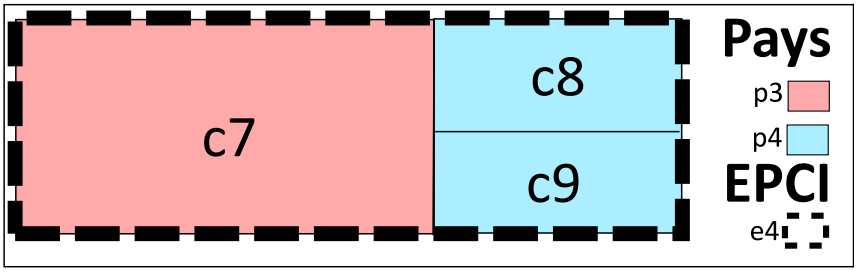


FIGURE 8.2 – La question du choix de unité de comparaison pour le calcul des cartes d'écarts territoriaux.

Une fois ces questions éclaircies, il est envisageable d'aller plus loin, c'est-à-dire de créer ou de supprimer des limites dans chaque zonage, comme de modifier dans une interface graphique les unités d'appartenance d'une unité, en respectant les contraintes posées par notre modèle : chaque unité d'un

niveau n'est définie que par des unités qui lui appartiennent totalement. La modification interactive des relations d'appartenance implique une mise à jour des valeurs statistiques par un processus d'agrégation dynamique. La modification de frontières, quant à elle, ne peut pas être réalisée sans activer des méthodes de transfert entre maillages non alignés. Il s'agit alors d'utiliser les capacités statistiques qu'offre notre plate-forme pour activer ou proposer à l'utilisateur une ou plusieurs méthodes de transfert adaptées à la nature de l'information traitée (des stocks dans HyperAtlas principalement). Par la suite, l'intégration de la dimension temporelle et de méthodes statistiques adaptées faciliterait la projection des données dans le futur et la construction de scénarios.

Le fruit de notre recherche pourrait servir le pouvoir politique à des fins peu honorables, comme par exemple l'optimisation de zonages électoraux. Cependant, comme nous l'avons exposé dès le préambule de cette thèse, si les frontières sont si souvent remodelées, c'est justement parce que déjà le pouvoir politique a compris l'intérêt qu'il existe à agir sur les zonages pour manipuler les chiffres qui sont censés être représentatifs de notre réalité. De précédents travaux se sont déjà intéressés à l'optimisation de zonages sur des critères spécifiés par l'utilisateur, avec des techniques d'agrégation par recuit simulé [Openshaw 88], [Martin 03], par régionalisation [Pumain 97], par algorithme génétique [Josselin 00], ou tenant compte des flux d'échange entre les unités [Terrier 80]. L'avantage de ce projet résiderait essentiellement dans l'interactivité offerte, incluant une capacité d'analyse des conséquences des modifications en terme de différenciation territoriale. L'accès libre pour tout citoyen à ce type d'outil de simulation nous semble en revanche essentiel, afin qu'un débat démocratique puisse être instauré, fondé sur l'ensemble de tous les scénarios produits par des simulations.

8.2.4 L'harmonisation de l'information statistique territoriale

Le sujet initial de cette thèse portait sur l'harmonisation de l'information statistique territoriale, qui permet par exemple de reconstituer des séries temporelles et qui exige de savoir estimer les valeurs manquantes. Constatant le caractère hétérogène des données (hétérogénéité des sources, des supports, des définitions, des classifications), notre travail a d'abord consisté à établir des bases solides d'un modèle destiné à l'estimation, en décrivant au mieux cette hétérogénéité, et les conséquences qu'elle peut avoir pour la qualité des données. Ainsi, la tâche d'estimation a été remise au lendemain.

Ce lendemain arrive, et pourtant un grand nombre de difficultés reste à surmonter pour achever cet objectif. Nous avons montré, par exemple, à travers l'estimation de la qualité des données, quelles seraient les méthodes à mettre en oeuvre et les difficultés que leur emploi pose. Nous avons dit qu'il manquait encore une ontologie des méthodes statistiques, et une ontologie des indicateurs statistiques. De même, nous nous apercevons que notre modèle de métadonnées n'est pas complet, qu'il faudra l'étendre, non seulement pour assurer l'interopérabilité complète avec SDMX, mais également pour intégrer des informations cruciales pour l'estimation (en différenciant les stocks des ratios par exemple).

Enfin, tout au long de cette thèse, nos travaux sont restés basés sur le paradigme orienté-objet. Or, pour l'estimation, il s'agira de modéliser les relations de distance (de toutes les sortes) entre les unités géographiques et de les conserver en vue d'optimiser les calculs [Ester 00], [Zeitouni 01]. De même, les relations de similarité entre les indicateurs statistiques, les relations d'équivalence entre les méthodes statistiques, en sus des relations historiques et hiérarchiques entre les unités géographiques que nous avons proposé, seront à calculer et à conserver également. De plus les relations à l'intérieur de l'ontologie statistique des indicateurs ou des méthodes par exemple ne seront certainement pas acycliques : l'héritage multiple doit avoir sa place dans un tel modèle.

Face à cette recrudescence de relations, nous avons l'impression qu'il faudra faire évoluer notre modèle orienté-objet vers un modèle d'hyper-graphe. En effet, la revue de quelques travaux dans ce domaine nous a convaincu que ce type de modélisation saurait gérer l'important volume de données qui sera à traiter désormais et la nature complexe de cette information, grâce à l'emploi des bases de données spécialisées pour les structures de graphes [Angles 08]. Ainsi, les outils de fouille de données, d'analyse spatiale et temporelle, et d'estimation pourraient alors profiter de la connaissance de tous les types de relations existant entre les noeuds du graphe de données [Ester 00], [Zeitouni 00], les unités géographiques étant un type de noeud de cet hyper-graphe, les indicateurs ou les méthodes un autre type.

L'harmonisation de l'information statistique territoriale est un travail de longue haleine, et nous avons tenté à travers cette thèse d'éclairer un peu la voie.

Quatrième partie

Annexes

Schéma XSD du profil ISO 19115 pour l’information statistique territoriale

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:gco="http://www.isotc211.org/2005/gco"
xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:esponMD="http://www.espon.eu/esponMD" targetNamespace="http://www.espon.eu/esponMD"
elementFormDefault="qualified" version="0.1">
    <!-- ================================== Annotation ================================== -->
    <xs:annotation>
        <xs:documentation>This file was generated by Christine Plumejeaud, 29 October 2009, using eclipse Europa 3.1.1
        This make an extent of the ISO 19115 standard for the ESPON 2013 DB project,
        allowing for the definition of new elements (all are prefixed with espon ),
        The template ESPON is provided to show how to use this extension. A word document explains this template.
        The principle is as follows: we use "series" element for our DATASET, and dataset element for our INDICATOR
            - at DATASET level, use the gmd:MD_Metadata element to describe. But replace inside the element gmd:MD_DataIdentification by
esponMD:datasetIdentification
            - at INDICATOR level, use the esponMD:indicatorIdentificationType instead of gmd:MD_DataIdentification
        The template and the XSD schema have been tested using the Xerces validator.
        It was then edited with XMLSpy v2009 sp1 (http://www.altova.com) by maria ramos (EMBRACE)
        </xs:documentation>
    </xs:annotation>
    <!-- ================================== Imports ================================== -->

    <xs:import namespace="http://www.isotc211.org/2005/gco" schemaLocation="gco/gco.xsd"/>
    <xs:import namespace="http://www.isotc211.org/2005/gmd" schemaLocation="gmd/gmd.xsd"/>
    <!-- ============================================================================ -->
    <xs:simpleType name="thesaurusCode_Type">
        <xs:annotation>
            <xs:documentation>List the available thesauri grouping High-level geospatial data thematic classification to assist in the grouping and
search of available geospatial datasets</xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:string">
            <xs:enumeration value="ESPON"/>
        </xs:restriction>
    </xs:simpleType>
    <!-- ..................................................................... -->
    <xs:element name="thesaurusCode" type="esponMD:thesaurusCode_Type" substitutionGroup="gco:CharacterString"/>
    <!-- ..................................................................... -->
    <xs:complexType name="thesaurusCode_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:thesaurusCode"/>
        </xs:sequence>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ============================================================================ -->
    <xs:simpleType name="topicCategoryCode_Type">
        <xs:annotation>
            <xs:documentation>High-level geospatial data thematic classification to assist in the grouping and search of available geospatial
datasets</xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:string">
            <xs:enumeration value="01. Agriculture and Fisheries"/>
            <xs:enumeration value="02. Demography (including Household, Population, ...)"/>
            <xs:enumeration value="03. Transport (including Accessiblity, Communication, Infrastructure, ...)"/>
            <xs:enumeration value="04. Energy and Environnement (including Climate, Consumption, Hazards, Pollution, Resources, ..."/>
            <xs:enumeration value="05. Land Use (including Land Cover, ...)"/>
            <xs:enumeration value="06. Social Affairs (including Culture, Education, Health, Literacy, ...)"/>
            <xs:enumeration value="07. Economy (including Employment, Finance, Industry, Labour, Technology, Trade, Tourism, R&amp;amp;D ...)"/>
            <xs:enumeration value="99. Non-/Cross-Thematic Data"/>
            <xs:enumeration value="01.01 Land Use"/>
            <xs:enumeration value="01.02 Farmer Structure"/>
            <xs:enumeration value="01.03 Employment"/>
            <xs:enumeration value="01.04 Livestock"/>
            <xs:enumeration value="01.05 Production"/>
            <xs:enumeration value="02.01 Population Structure"/>
            <xs:enumeration value="02.02 Population Movement"/>
            <xs:enumeration value="03.01 Transport Infrastructure"/>
            <xs:enumeration value="03.02 Passengers and Goods Transport"/>
            <xs:enumeration value="03.03 Accessibility"/>
            <xs:enumeration value="03.04 Impact of Transport Policies"/>
            <xs:enumeration value="04.01 Natural Hazards"/>
            <xs:enumeration value="04.02 Environmental quality"/>
            <xs:enumeration value="05.01 Land Use"/>
            <xs:enumeration value="06.01 Education"/>
            <xs:enumeration value="06.02 Poverty"/>
            <xs:enumeration value="07.01 Employment"/>
            <xs:enumeration value="07.02 Unemployment"/>
            <xs:enumeration value="07.03 Income and Consumption"/>
            <xs:enumeration value="07.04 Finances and Expenditures"/>
            <xs:enumeration value="07.05 Tourism"/>
            <xs:enumeration value="99.01 Integrative indices, typologies and scenarios"/>
            <xs:enumeration value="99.99 Geographical objects"/>
        </xs:restriction>
    </xs:simpleType>
    <!-- ..................................................................... -->
    <xs:element name="topicCategoryCode" type="esponMD:topicCategoryCode_Type" substitutionGroup="gco:CharacterString"/>
    <!-- ..................................................................... -->
    <xs:complexType name="topicCategoryCode_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:topicCategoryCode"/>
        </xs:sequence>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ============================================================================ -->
    <xs:complexType name="classificationType_Type">
        <xs:annotation>
            <xs:documentation>This group the thesaurus, thems and keywords inside an elements,
            so that we know from which thesaurus the themes and keyWords are extracted. </xs:documentation>
        </xs:annotation>
```

```
    <xs:complexContent>
        <xs:extension base="gco:AbstractObject_Type">
            <xs:sequence>
                <xs:element name="thesaurus" type="esponMD:thesaurusCode_PropertyType"/>
                <xs:element name="topicCategory" type="esponMD:topicCategoryCode_PropertyType" minOccurs="1" maxOccurs="1"/>
                <xs:element name="keyword" type="gco:CharacterString_PropertyType" minOccurs="1" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="classificationType" type="esponMD:classificationType_Type"/>
<!-- ..................................................................... -->
<xs:complexType name="classificationType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:classificationType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================= -->
<xs:complexType name="dataConstraintsType_Type">
    <xs:annotation>
        <xs:documentation>Restrictions and legal prerequisites for accessing and using the dataset.</xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:MD_Constraints_Type">
            <xs:sequence>
                <xs:element name="copyright" type="gco:CharacterString_PropertyType"/>
                <xs:element name="freeUse" type="gco:Boolean_PropertyType" minOccurs="1"/>
                <xs:element name="accessConstraints" type="gmd:MD_RestrictionCode_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                <xs:element name="useConstraints" type="gmd:MD_RestrictionCode_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                <xs:element name="otherConstraints" type="gco:CharacterString_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="dataConstraintsType" type="esponMD:dataConstraintsType_Type" substitutionGroup="gmd:MD_Constraints"/>
<!-- ..................................................................... -->
<xs:complexType name="dataConstraintsType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:dataConstraintsType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================= -->
<xs:complexType name="MetaDataConstraints_Type">
    <xs:annotation>
        <xs:documentation>Restrictions and legal prerequisites for accessing and using the metadata.</xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:MD_Constraints_Type">
            <xs:sequence>
                <xs:element name="readRights" type="gco:Boolean_PropertyType"/>
                <xs:element name="accessConstraints" type="gmd:MD_RestrictionCode_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="metadataConstraintsType" type="esponMD:MetaDataConstraints_Type" substitutionGroup="gmd:MD_Constraints"/>
<!-- ..................................................................... -->
<xs:complexType name="metadataConstraintsType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:metadataConstraintsType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================= -->
<xs:complexType name="constraintsType_Type">
    <xs:annotation>
        <xs:documentation>Restrictions and legal prerequisites for accessing and using the data and the metadata.</xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:MD_Constraints_Type">
            <xs:sequence>
                <xs:element name="dataConstraints" type="esponMD:dataConstraintsType_PropertyType"/>
                <xs:element name="metadataConstraints" type="esponMD:metadataConstraintsType_PropertyType"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="constraintsType" type="esponMD:constraintsType_Type" substitutionGroup="gmd:MD_Constraints"/>
<!-- ..................................................................... -->
<xs:complexType name="constraintsType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:constraintsType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================= -->
```

```
<xs:simpleType name="scopeCode_Type">
    <xs:annotation>
        <xs:documentation>Description of the class of information covered by the information</xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
        <xs:enumeration value="any"/>
        <xs:enumeration value="all"/>
        <xs:enumeration value="specifiedExtent"/>
    </xs:restriction>
</xs:simpleType>
<!-- ..................................................................... -->
<xs:element name="scopeCode" type="esponMD:scopeCode_Type" substitutionGroup="gco:CharacterString"/>
<!-- ..................................................................... -->
<xs:complexType name="scopeCode_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:scopeCode"/>
    </xs:sequence>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ================================================================================ -->
<xs:simpleType name="nomenclatureCode_Type">
    <xs:annotation>
        <xs:documentation>Description of the class of information covered by the information </xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
        <xs:enumeration value="any"/>
        <xs:enumeration value="WUTS0"/>
        <xs:enumeration value="WUTS1"/>
        <xs:enumeration value="WUTS2"/>
        <xs:enumeration value="WUTS3"/>
        <xs:enumeration value="NUTS0"/>
        <xs:enumeration value="NUTS1"/>
        <xs:enumeration value="NUTS2"/>
        <xs:enumeration value="NUTS2-3"/>
        <xs:enumeration value="NUTS3"/>
        <xs:enumeration value="NUTS4"/>
        <xs:enumeration value="NUTS5"/>
        <xs:enumeration value="LAU1"/>
        <xs:enumeration value="LAU2"/>
    </xs:restriction>
</xs:simpleType>
<!-- ..................................................................... -->
<xs:element name="nomenclatureCode" type="esponMD:nomenclatureCode_Type" substitutionGroup="gco:CharacterString"/>
<!-- ..................................................................... -->
<xs:complexType name="nomenclatureCode_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:nomenclatureCode"/>
    </xs:sequence>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ================================== Classes ================================== -->
<xs:complexType name="maintenanceInformation_Type">
    <xs:annotation>
        <xs:documentation>Information about the scope and frequency of updating</xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:MD_MaintenanceInformation_Type">
            <xs:sequence>
                <xs:element name="regularUpdates" type="gco:Boolean_PropertyType"/>
                <xs:element name="scopeCode" type="esponMD:scopeCode_PropertyType"/>
                <xs:element name="scopeExtent" type="gmd:EX_SpatialTemporalExtent_Type" minOccurs="0" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="maintenanceInformation" type="esponMD:maintenanceInformation_Type"/>
<!-- ..................................................................... -->
<xs:complexType name="maintenanceInformation_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:maintenanceInformation"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ================================================================================ -->
<xs:complexType name="datasetIdentification_Type">
    <xs:annotation>
        <xs:documentation>This is used to fill the Identification topic of the series (DATASET level)
    - all information are common to a set of the indicators </xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:AbstractMD_Identification_Type">
            <xs:sequence>
                <!--  INHERITED AND MANDATORY ELEMENTS-->
                <!--
                <xs:element name="citation" type="gmd:CI_Citation_PropertyType"/>
                <xs:element name="abstract" type="gco:CharacterString_PropertyType"/>
                -->
                <xs:element name="maintenance" type="esponMD:maintenanceInformation_PropertyType" minOccurs="0"/>
                <!-- at IMPORT, maintenance should not be filled -->
                <xs:element name="dataQualityInfo" type="esponMD:dataQualityType_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
```

```
    <!-- ..................................................................... -->
    <xs:element name="datasetIdentification" type="esponMD:datasetIdentification_Type" substitutionGroup="gmd:AbstractMD_Identification"/>
    <!-- ..................................................................... -->
    <xs:complexType name="datasetIdentification_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:datasetIdentification"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:complexType name="spatialRepresentation_Type">
        <xs:annotation>
            <xs:documentation>Information about the spatial objects in the dataset : can choose a nomenclature, and precise the levels used
            It can also be vector or grid (see the URL http://www.isotc211.org/2005/resources/codeList/
gmxCodelists.xml#MD_SpatialRepresentationTypeCode)  </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:AbstractMD_SpatialRepresentation_Type">
                <xs:sequence>
                    <xs:element name="spatialRepresentationType" type="gmd:MD_SpatialRepresentationTypeCode_PropertyType"/>
                    <xs:element name="spatialResolution" type="gmd:MD_Resolution_PropertyType" minOccurs="0"/>
                    <xs:element name="spatialNomenclatureName" type="gco:CharacterString_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                    <xs:element name="nomenclatureLevel" type="esponMD:nomenclatureCode_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                    <xs:choice>
                        <xs:element name="vectorSpatialRepresentation" type="gmd:MD_VectorSpatialRepresentation_Type" minOccurs="0"/>
                        <xs:element name="gridSpatialRepresentation" type="gmd:MD_GridSpatialRepresentation_Type" minOccurs="0"/>
                    </xs:choice>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="spatialRepresentation" type="esponMD:spatialRepresentation_Type" substitutionGroup="gmd:AbstractMD_SpatialRepresentation"/>
    <!-- ..................................................................... -->
    <xs:complexType name="spatialRepresentation_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:spatialRepresentation"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ========================================================================= -->
    <xs:complexType name="indicatorIdentificationType_Type">
        <xs:annotation>
            <xs:documentation>Use this at INDICATOR level instead of gmd:MD_DataIdentification to give the minimum but necessary informations
            - name, classification, code, etc., are mandatory elements;
             </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:AbstractMD_Identification_Type">
                <xs:sequence>
                    <!--  INHERITED AND MANDATORY ELEMENTS-->
                    <!--
                    <xs:element name="citation" type="gmd:CI_Citation_PropertyType"/>
                    <xs:element name="abstract" type="gco:CharacterString_PropertyType"/>
                    -->
                    <xs:element name="code" type="gco:CharacterString_PropertyType"/>
                    <xs:element name="unitOfMeasure" type="gco:CharacterString_PropertyType"/>
                    <xs:element name="indicatorMethodology" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                    <xs:element name="language" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                    <!--  NOT EDIT : ENGLISH -->
                    <xs:element name="characterSet" type="gmd:MD_CharacterSetCode_PropertyType" minOccurs="0" maxOccurs="1"/>
                    <!--  NOT EDIT : UTF8 -->
                    <xs:element name="classification" type="esponMD:classificationType_PropertyType" maxOccurs="unbounded"/>
                    <xs:element name="extent" type="gmd:EX_SpatialTemporalExtent_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                    <!-- set extent.minOccurs to 0 for IMPORT step -->
                </xs:sequence>
            </xs:extension>
            <!-- gmd:MD_DataIdentification_Type  -->
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="indicatorIdentificationType" type="esponMD:indicatorIdentificationType_Type" substitutionGroup="gmd:AbstractMD_Identification"/
>
    <!-- gmd:MD_DataIdentification  -->
    <!-- ..................................................................... -->
    <xs:complexType name="indicatorIdentificationType_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:indicatorIdentificationType"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ========================================================================= -->
    <xs:complexType name="scopeType_Type">
        <xs:annotation>
            <xs:documentation>Give the scope of the information given for quality :
            - can be the whole coverage of the indicator (many years, the full study area),
            - either a small part (one year, just one geographic unit by example)
            Label is mandatory and unique for one indicator.
            Informations givens for the smaller extent prevail above the wider extent.
            </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:DQ_Scope_Type">
```

```
                <xs:sequence>
                    <!--  INHERITED AND MANDATORY ELEMENTS-->
                    <!--
                    <xs:element name="level" type="gmd:MD_ScopeCode_PropertyType"/> -->
                    <!-- XSL FILL gmd:level with a foo value such as series -->
                    <xs:element name="spatioTemporalExtent" type="gmd:EX_SpatialTemporalExtent_PropertyType" minOccurs="0"/>
                    <!-- replaced gmd:EX_Extent_PropertyType by gmd:EX_SpatialTemporalExtent_PropertyType, and set it as optional (for IMPORT step)
-->
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:complexType name="scopeType_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:scopeType"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="scopeType" type="esponMD:scopeType_Type" substitutionGroup="gmd:DQ_Scope"/>
    <!-- ===================================================================== -->
    <xs:complexType name="sourceCitationType_Type">
        <xs:annotation>
            <xs:documentation>Standardized sourceCitation reference (extends the gmd:CI_Citation_Type)
            The new and mandatory element is the date of the data acquisition ;
            and you can precise the dataSetURI (optional) which indicates the location (web link) of the data to download
            as well as the URI of the provider, and the name of the source (file or database name) that the provider uses.
            You MUST add the elements of gmd:CI_Citation_Type and date (of extraction)
            </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:CI_Citation_Type">
                <xs:sequence>
                    <!--  INHERITED AND MANDATORY ELEMENTS-->
                    <!--
                    <xs:element name="title" type="gco:CharacterString_PropertyType"/>
                    <xs:element name="date" type="gmd:CI_Date_PropertyType" maxOccurs="unbounded"/>
                    -->
                    <xs:element name="dataSetURI" type="gco:CharacterString_PropertyType" minOccurs="0"/><!--  if available --
>
                    <!-- Extension : 07 february 2011 (CP on demand of ESPON 2013 project -->
                    <xs:element name="providerURI" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                    <xs:element name="sourceName" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="sourceCitationType" type="esponMD:sourceCitationType_Type" substitutionGroup="gmd:CI_Citation"/>
    <!-- ..................................................................... -->
    <xs:complexType name="sourceCitationType_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:sourceCitationType"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ===================================================================== -->
    <xs:complexType name="sourceType_Type">
        <xs:annotation>
            <xs:documentation>Espon source gives a provider name (the title of citation), and optionnal provider URI, a date of acquisition, a source
name and source URI
            </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gco:AbstractObject_Type">
                <xs:sequence>
                    <!-- name of the provider is filled with the sourceCitationType.title instead -->
                    <!--  <xs:element name="newSource" type="gco:Boolean_PropertyType" minOccurs="1" maxOccurs="1"/> -->
                    <xs:element name="sourceCitation" type="esponMD:sourceCitationType_PropertyType"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="sourceType" type="esponMD:sourceType_Type"/>
    <!-- ..................................................................... -->
    <xs:complexType name="sourceType_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:sourceType"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ===================================================================== -->
    <xs:complexType name="methodologyType_Type">
        <xs:annotation>
            <xs:documentation>
            This new element defines the process steps linked to an indicator in a simplified way.
            You can :
            - link one to many files with the metadata file, documenting the process steps
            - express the formula you could have applied on the data to compute them from source data.
            - describe in a text the process steps that lead to this indicator
            - list of components separated by a semi-colon separator
```

```
        </xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gco:AbstractObject_Type">
            <xs:sequence>
                <xs:element name="description" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                <xs:element name="formula" type="gco:CharacterString_PropertyType" minOccurs="0"/>
                <xs:element name="file" type="gco:CharacterString_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="methodologyType" type="esponMD:methodologyType_Type"/>
<!-- ..................................................................... -->
<xs:complexType name="methodologyType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:methodologyType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================== -->
<xs:complexType name="lineageType_Type">
    <xs:annotation>
        <xs:documentation>Espon lineage is made of 2 parts :
        - the source,
        - the methodology,
        This element extends the ancient gmd:LI_Lineage_Type but if forces some new mandatory simplified elements
        </xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gmd:LI_Lineage_Type">
            <xs:sequence>
                <xs:element name="source" type="esponMD:sourceType_PropertyType"/>
                <xs:element name="methodology" type="esponMD:methodologyType_PropertyType"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="lineageType" type="esponMD:lineageType_Type" substitutionGroup="gmd:LI_Lineage"/>
<!-- ..................................................................... -->
<xs:complexType name="lineageType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:lineageType"/>
    </xs:sequence>
    <xs:attributeGroup ref="gco:ObjectReference"/>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================== -->
<xs:simpleType name="qualityLevelCode_Type">
    <xs:annotation>
        <xs:documentation>Give a human estimate level of the indicator quality : high is the best quality level</xs:documentation>
    </xs:annotation>
    <xs:restriction base="xs:string">
        <xs:enumeration value="high"/>
        <xs:enumeration value="medium"/>
        <xs:enumeration value="low"/>
        <xs:enumeration value="no opinion"/>
    </xs:restriction>
</xs:simpleType>
<!-- ..................................................................... -->
<xs:element name="qualityLevelCode" type="esponMD:qualityLevelCode_Type" substitutionGroup="gco:CharacterString"/>
<!-- ..................................................................... -->
<xs:complexType name="qualityLevelCode_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:qualityLevelCode"/>
    </xs:sequence>
    <xs:attribute ref="gco:nilReason"/>
</xs:complexType>
<!-- ========================================================================== -->
<xs:complexType name="reliabilityType_Type">
    <xs:annotation>
        <xs:documentation>This new element could be assimilated to a report : indicates id data are issues from an official source
        (in an document, ESPON will provide a list of official providers (or considered as)),
         if data have been estimated, and the quality level that the provider estimates </xs:documentation>
    </xs:annotation>
    <xs:complexContent>
        <xs:extension base="gco:AbstractObject_Type">
            <xs:sequence>
                <xs:element name="official" type="gco:Boolean_PropertyType" minOccurs="0"/>
                <!--  official is not mandatory at IMPORT step : will be computed through a rule -->
                <xs:element name="estimation" type="gco:Boolean_PropertyType"/>
                <xs:element name="qualityLevel" type="esponMD:qualityLevelCode_PropertyType"/>
            </xs:sequence>
        </xs:extension>
    </xs:complexContent>
</xs:complexType>
<!-- ..................................................................... -->
<xs:element name="reliabilityType" type="esponMD:reliabilityType_Type"/>
<!-- ..................................................................... -->
<xs:complexType name="reliabilityType_PropertyType">
    <xs:sequence minOccurs="0">
        <xs:element ref="esponMD:reliabilityType"/>
    </xs:sequence>
```

```
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ============================================================================ -->
    <xs:complexType name="dataQualityType_Type">
        <xs:annotation>
            <xs:documentation>
            Quality is extended with a reliability element and a constraint element,
            and the scope element is extended to include an SpatioTemporalExtent element
            Informations givens for the smaller extent prevail above the wider extent.
            The constraint at value level allow user for a finest control of data dissemination rights.
            Indeed, constraints are semantically linked with the lineage of the data : if data are extracted from a source
            that doesn't allow public data dissemination, this fact can be expressed here.
            </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:DQ_DataQuality_Type">
                <xs:sequence>
                    <!-- INHERITED MANDATORY
                    <xs:element name="scope" type="gmd:DQ_Scope_PropertyType"/>
                    but esponMD:scopeType_PropertyType can be used instead of gmd:DQ_Scope_PropertyType
                    -->
                    <!--
                    label points on a set of values inside the dataset
                    scope is mandatory, but can be typed with scopeType_Type instead of DQ_Scope_Type
                    For the scope, only a label should be provided at IMPORT, valued with "series" or "dataset" value
                    lineage is mandatory: gives the source, the methodology, the temporal lineage, the spatial lineage (NUTS version, NUTS level)
                    constraint is mandatory: gives the access rights, copyrights of data, and metadata access
                    (a metadata access to false would mean that the existence of this value should be hidden to public)
                    -->

                    <xs:element name="label" type="gco:CharacterString_PropertyType" />
                    <!-- <xs:element name="scopeType" type="esponMD:scopeType_PropertyType" minOccurs="0"/> -->
                    <xs:element name="lineage" type="esponMD:lineageType_PropertyType"/>
                    <xs:element name="reliability" type="esponMD:reliabilityType_PropertyType"/>
                    <xs:element name="constraints" type="esponMD:constraintsType_PropertyType"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="dataQualityType" type="esponMD:dataQualityType_Type" substitutionGroup="gmd:DQ_DataQuality"/>
    <!-- ..................................................................... -->
    <xs:complexType name="dataQualityType_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:dataQualityType"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
    <!-- ============================================================================ -->
    <!-- not used for the moment -->
    <xs:complexType name="indicatorMetadata_Type">
        <xs:annotation>
            <xs:documentation>
            Use this element for INDICATOR level. This forces the quality and new indicatorIdentification elements to be mandatory.
            But contact and dataStamp are still mandatory
            User could add as many elements as wished (this gives more flexibility to the extension)
            The extension keeps all the ancient elements, in order to allow for the copy-paste of metadata set information
            This would be usefull for example when extracting a set of indicators coming from various datasets
            </xs:documentation>
        </xs:annotation>
        <xs:complexContent>
            <xs:extension base="gmd:MD_Metadata_Type">
                <xs:sequence>
                    <!-- INHERITED MANDATORY
                    <xs:element name="contact" type="gmd:CI_ResponsibleParty_PropertyType" maxOccurs="unbounded"/>
                    <xs:element name="dateStamp" type="gco:Date_PropertyType"/>
                    -->
                    <xs:element name="indicatorIdentification" type="esponMD:indicatorIdentificationType_PropertyType" maxOccurs="unbounded"/>
                    <!-- quality can be specified either at dataset level, either on each indicator
                        The quality element will be mandatory at EXPORT step, but can be skipped at IMPORT step, if the user fill quality elements at
Dataset level.
                    <xs:element name="dataQualityInfo" type="esponMD:dataQualityType_PropertyType" minOccurs="0" maxOccurs="unbounded"/>
                     -->

                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
    <!-- ..................................................................... -->
    <xs:element name="indicatorMetadata" type="esponMD:indicatorMetadata_Type" substitutionGroup="gmd:MD_Metadata"/>
    <!-- ..................................................................... -->
    <xs:complexType name="esponMetadata_PropertyType">
        <xs:sequence minOccurs="0">
            <xs:element ref="esponMD:indicatorMetadata"/>
        </xs:sequence>
        <xs:attributeGroup ref="gco:ObjectReference"/>
        <xs:attribute ref="gco:nilReason"/>
    </xs:complexType>
</xs:schema>
```

Structure d’un fichier DSD en SDMX

```
<!DOCTYPE root>
<root>
    <Structure xmlns="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message"
        xmlns:message="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message SDMXMessage.xsd
        http://www.SDMX.org/resources/SDMXML/schemas/v2_0/structure SDMXStructure.xsd">

        <!-- l'entête (Header) permet d'identifier cette structure de données,
        à laquelle on attache un identifiant unique (ID), un nom (Name) ainsi qu'une date (Prepared)
        Le créateur de cette structure de données (Sender) peut être référencé dans une liste de codes
        pré-établis dans un des registres SDMX : il est alors identifié par son id dans cette liste.-->
        <Header>
            <ID>IREF000506</ID>
            <Test>false</Test>
            <Name>ECB structural definitions</Name>
            <Prepared>2006-10-25T14:26:00</Prepared>
            <Sender id="4F0"/>
        </Header>

        <!-- Chaque élément Concept définit un des concepts utilisés pour identifier et décrire des données. -->
        <Concept agencyID="ECB" id="COLLECTION">
            <Name xml:lang="fr">Collection d'indicateurs</Name>
        </Concept>

        <!-- Chaque élément CodeList définit une liste de codes (et leur signification) qui seront utilisés pour valuer un attribut.
            Chaque élément CodeList  contient au moins 2 attributs:
            - L'ID de l'organisme (agency) responsable de cette liste de code (dans l'exemple, "ECB")
            - L'ID de cette liste de code (dans l'exemple, "CL_EXR_SUFFIX").-->
        <CodeList agencyID="ECB" id="CL_EXR_SUFFIX">
            <Name xml:lang="en">Liste des codes associés à la variation des taux de change</Name>
            <Code value="A">
                <Description xml:lang="en">Moyenne (ou mesure standardisée) sur la période</Description>
            </Code>
            <Code value="E">
                <Description xml:lang="en">Valeur en fin de période</Description>
            </Code>
        </CodeList>
        <CodeList agencyID="ECB" id="CL_EXR_TYPE">
        ...
        </CodeList>

        <!-- L'élément KeyFamily définit ensuite la structure du document qui portera les données
            - L'ID de l'organisme (agency) responsable de cette structure est obligatoire (dans l'exemple, "ECB").
            - l'ID de ce jeu de données (dans l'exemple, "ECB_EXR1")
            - l'URI de l'espace de nommage de ce jeu de données (dans l'exemple, "http://www.ecb.int/vocabulary/stats/exr/1")
            - le nom du jeu de donnée dans une langue (dans l'exemple, "Taux d'échanges") -->
        <KeyFamily agencyID="ECB" id="ECB_EXR1"
            uri="http://www.ecb.int/vocabulary/stats/exr/1">
            <Name xml:lang="fr">Taux d'échanges</Name>
            <Components>
                <!-- Listes des dimensions utilisées pour descrire les valeurs statistiques-->
                <Dimension conceptRef="FREQ" codelist="CL_FREQ" isFrequencyDimension="true"/>
                <Dimension conceptRef="CURRENCY" codelist="CL_CURRENCY"/><!-- la devise au numérateur -->
                <Dimension conceptRef="CURRENCY_DENOM" codelist="CL_CURRENCY"/> <!-- la devise au numérateur -->
                <Dimension conceptRef="EXR_TYPE" codelist="CL_EXR_TYPE"/> <!-- Type de taux d'échange -->
                <Dimension conceptRef="EXR_SUFFIX" codelist="CL_EXR_SUFFIX"/> <!-- Précision sur la mesure donnée du taux d'échange -->
                <TimeDimension conceptRef="TIME_PERIOD"/> <!-- période de validité de l'information -->

                <!-- Dimensions associées au niveau du groupe (Group) -->
                <Group id="Group">
                    <DimensionRef>CURRENCY</DimensionRef>
                    <DimensionRef>CURRENCY_DENOM</DimensionRef>
                    <DimensionRef>EXR_TYPE</DimensionRef>
                    <DimensionRef>EXR_SUFFIX</DimensionRef>
                </Group>
                <!-- PrimaryMeasure indique quel concept porte la valeur mesurée pour chaque unité - Par convention, c'est le concept OBS_VALUE -->
                <PrimaryMeasure conceptRef="OBS_VALUE"/>

                <!-- liste des attributs -->
                <Attributes>
                    <!--  l'attribut TIME_FORMAT est défini au niveau des Series, est obligatoire,
                    est une chaîne de caractère de longueur 3, au format de la norme ISO8601 -->
                    <Attribute conceptRef="TIME_FORMAT" attachmentLevel="Series"
                        assignmentStatus="Mandatory" isTimeFormat="true">
                        <TextFormat textType="String" maxLength="3"/>
                    </Attribute>
                    <!--  l'attribut OBS_STATUS est défini au niveau des Observations, est obligatoire, est un code,
                    défini dans la liste  CL_OBS_STATUS -->
                    <Attribute conceptRef="OBS_STATUS" attachmentLevel="Observation"
                        assignmentStatus="Mandatory" codelist="CL_OBS_STATUS"/>
                    <!--  l'attribut DECIMALS est défini au niveau du Group, est obligatoire, est un code, défini dans la liste CL_DECIMALS -->
                    <Attribute conceptRef="DECIMALS" attachmentLevel="Group"
                        assignmentStatus="Mandatory" codelist="CL_DECIMALS">
                        <AttachmentGroup>Group</AttachmentGroup>
                    </Attribute>
                </Attributes>
            </Components>
        </KeyFamily>
    </Structure>
</root>
```

Instanciation du profil esponMD de la norme ISO 19115

```
        <esponMD:dataQualityInfo>
         <esponMD:dataQualityType>
           <gmd:scope>
             <esponMD:scopeType>
               <gmd:level>
                 <gmd:MD_ScopeCode codeList="http://www.isotc211.org/2005/
                     resources/codeList.xml#MD_ScopeCode" codeListValue="
                     series"/>
               </gmd:level>
             </esponMD:scopeType>
           </gmd:scope>
           <esponMD:label>
             <gco:CharacterString>1</gco:CharacterString>
           </esponMD:label>
           <esponMD:lineage>
             <esponMD:lineageType>
               <esponMD:source>
                 <esponMD:sourceType>
                   <esponMD:sourceCitation>
                     <esponMD:sourceCitationType>
                       <gmd:title>
                         <gco:CharacterString>SHRINKING</gco:CharacterString>
                       </gmd:title>
                       <gmd:date>
                         <gco:DateTime>2011-02-16T00:00:00</gco:DateTime>
                       </gmd:date>
                       <esponMD:dataSetURI>
                         <gco:CharacterString>http://www.europarl.europa.eu/
                             activities/committees/studies/download.do?
                             language=en&amp;file=22350#search=%20SHRINKING
                             %20</gco:CharacterString>
                       </esponMD:dataSetURI>
                     </esponMD:sourceCitationType>
                   </esponMD:sourceCitation>
                 </esponMD:sourceType>
               </esponMD:source>
               <esponMD:methodology>
                 <esponMD:methodologyType>
                   <esponMD:description gco:nilReason="missing">
```

```
                <gco:CharacterString/>
              </esponMD:description>
              <esponMD:file>
                <gco:CharacterString>http://www.ums-riate.fr/documents/
                    Shrinking_Study_EN.pdf</gco:CharacterString>
              </esponMD:file>
            </esponMD:methodologyType>
          </esponMD:methodology>
        </esponMD:lineageType>
      </esponMD:lineage>
      <esponMD:reliability>
        <esponMD:reliabilityType>
          <esponMD:estimation>
            <gco:Boolean>true</gco:Boolean>
          </esponMD:estimation>
          <esponMD:fiabilityLevel>
            <esponMD:qualityLevelCode>high</esponMD:qualityLevelCode>
          </esponMD:fiabilityLevel>
        </esponMD:reliabilityType>
      </esponMD:reliability>
      <esponMD:constraints>
        <esponMD:constraintsType>
          <esponMD:dataConstraints>
            <esponMD:dataConstraintsType>
              <esponMD:copyright>
                <gco:CharacterString>Reproduction and translation, for
                    non-commercial purposes, are authorised provided the
                     source is acknowledged and the publisher is given
                    prior notice and sent a copy.</gco:CharacterString>
              </esponMD:copyright>
              <esponMD:freeUse>
                <gco:Boolean>true</gco:Boolean>
              </esponMD:freeUse>
            </esponMD:dataConstraintsType>
          </esponMD:dataConstraints>
          <esponMD:metadataConstraints>
            <esponMD:metadataConstraintsType>
              <esponMD:readRights>
                <gco:Boolean>true</gco:Boolean>
              </esponMD:readRights>
            </esponMD:metadataConstraintsType>
          </esponMD:metadataConstraints>
        </esponMD:constraintsType>
      </esponMD:constraints>
    </esponMD:dataQualityType>
  </esponMD:dataQualityInfo>
```

Listing 8.1 – Exemple de code XML produit.

Traduction du profil esponMD vers SDMX

TABLE 8.1 – Correspondances entre les éléments de la norme ISO 19115 utilisés dans le profil *esponMD* et les Concepts définis par SDMX.

Attribut utilisé dans le profil *esponMD*	Concept(s) SDMX	Commentaire
MD_Metadata		
fileIdentifier	ID	Identifiant du fichier de métadonnées, attribué automatiquement
dateStamp	META_UPDATE	Date de rédaction des métadonnées
CI_ResponsibleParty		
individualName	-	Nom d'un contact
organisationName	CONTACT_ORGANISATION, ORGANISATION_UNIT	Organisation que le contact représente
positionName	CONTACT_FUNCT	Fonction dans l'organisation
role	-	Rôle de ce contact par rapport au jeu de données
CI_Contact		
CI_Telephone		
voice	CONTACT_PHONE	Téléphone fixe
facsimile	CONTACT_FAX	Fax
CI_Adress	CONTACT_MAIL	
electronicMailAddress	CONTACT_EMAIL	Adresse e-mail
deliveryPoint	CONTACT_MAIL	Type de voie (rue, avenue, etc.), numéro de voie et voie.
city	CONTACT_MAIL	Ville
administrativeArea	CONTACT_MAIL	Cedex
postalCode	CONTACT_MAIL	Code postal

Suite sur la page suivante

TABLE 8.1 – Suite de la page précédente

Attribut utilisé dans le profil *esponMD*	Concept(s) SDMX	Commentaire
country	CONTACT_MAIL	Pays
MD_Identification		
abstract	DATA_PRES, DATA_DESCR, DISS_DET	Résumé décrivant le jeu de données
CI_Citation.title	DSI	Titre du jeu de données
CI_Citation.date	DATA_UPDATE	Date d'acquisition du jeu de données
DataQuality		
label	-	Étiquette
level	-	Niveau de détail de l'élément DataQuality - renseigné par "*tile*"
SourceCitation		
dataSetURI	ORIG_DATA_ID	URI de la source de données
sourceName	ONLINE_DB, DISS_OTHER, PUBLICATIONS	Nom du support de la source de données
date	-	Date de publication des données de cette source (millésime)
title	COMPILING_ORG	Nom du fournisseur de données
Methodology		
description	COLL_METHOD, DATA_COMP, DATA_REV, ADJUSTMENT, RECORDING	Description libre de la méthode de calcul ou mesure du groupe de valeurs
formula	IND_TYPE, RECORDING	Formule de calcul ou de transformation des données
file	DOC_METHOD	Document externe décrivant les méthodes de calcul et de transformation des données
Reliability		
estimation	OBS_STATUS	Vrai si la valeur est une estimation
qualityLevel	QUALITY_ASSMNT	Valeur énumérée sur le niveau de qualité
Contraints		
DataConstraints	CONF	Restrictions légales sur la diffusion des données

Suite sur la page suivante

TABLE 8.1 – Suite de la page précédente

Attribut utilisé dans le profil *esponMD*	Concept(s) SDMX	Commentaire
copyright	CONF_POLICY, REL_POL_US_AC	Mention juridique relative aux droits d'usage de ces données
freeUse	CONF_STATUS_OBS, REL_POL_US_AC	Données en accès libre
MetadataConstraints	-	Restrictions légales sur la diffusion des métadonnées
readRight	-	Métadonnées en accès libre
MD_Distribution	ACCESSIBILITY, DISS_ORG, DISS_FORMAT, REL_POL_LEG_ACTS	Conditions d'obtention de la donnée
MD_Maintenance	M_AGENCY	
IndicatorIdentification		
abstract	COMMENT_TS, STAT_CONC_DEF	Résumé décrivant l'indicateur
code	-	Code de l'indicateur
unitOfMeasure	CURRENCY, DECIMALS, UNIT_MULT, UNIT_MEASURE	Unité de mesure de l'indicateur
indicatorMethodology	COLL_METHOD, DATA_COMP, DATA_REV, ADJUSTMENT, RECORDING	Méthodologie pour le calcul de cet indicateur
CI_Citation.title	TITLE	Nom de l'indicateur
CI_Citation.date	DATA_UPDATE	Date d'acquisition de l'indicateur (identique à celle du jeu de données)
Classification	CLASS_SYSTEM	Classification des données
thesaurus	-	Nom du thésaurus
topicCategory	-	Code du thème sélectionné dans le thésaurus
keyword	-	Mot clé

TABLE 8.2 – Structure des concepts présents dans le modèle SDMX du profil esponMD.

Concept	Statut	Description	Signification
Niveau : Dataset			
ID	A	text	Identifiant du fichier de métadonnées.
META_UPDATE	A	Date	Date de rédaction des métadonnées.
CONTACT_PERSON	A	text	Nom d'une personne qui sert de contact pour ce jeu de données.
CONTACT_ORGANISATION	A	text	Organisation que le contact représente.
CONTACT_FUNCT	A	text	Fonction du contact dans l'organisation.
CONTACT_ROLE	A	CL_ROLES	Rôle d'un contact.
CONTACT_PHONE	A	text	Téléphone fixe.
CONTACT_FAX	A	text	Fax.
CONTACT_MAIL	A	text	Adresse postale.
CONTACT_EMAIL	A	text	Adresse e-mail.
DATA_DESCR	A	text	Résumé décrivant le jeu de données.
DSI	D	text	Titre du jeu de données.
DATA_UPDATE	D	Date	Date d'acquisition du jeu de données.
Niveau : Obs			
STAT_UNIT	D	CL_AREA	Code de l'unité statistique territoriale.
QUALITY_VALUE_ID	A	text	Etiquette attachée à un groupe de valeur ayant la même provenance, pour lesquels la qualité, ainsi que les contraintes légales de diffusion sont identiques.
QUALITY_VALUE_SCOPE	A	CL_SCOPES	Niveau du groupe de valeur - de valeur constante « *tile* ».
ORIG_DATA_ID	D	text	URI de la source de données.
PUBLICATIONS	A	text	Nom du support de la source de données.
PUBLI_DATE	D	Date	Date de publication des observations (millésime).
COMPILING_ORG	D	text	Nom du fournisseur de données.
COLL_METHOD	A	text	Description libre de la méthode de calcul ou mesure du groupe de valeurs.

Suite sur la page suivante

TABLE 8.2 – Suite de la page précédente

Concept	Statut	Description	Signification
RECORDING	A	text	Formule de calcul ou de transformation des données.
DOC_METHOD	A	text	Document externe décrivant les méthodes de calcul et de transformation des données.
OBS_STATUS	A	CL_OBS_STATUS	La valeur est-elle normal, une prévision, une estimation, etc. ?
QUALITY_ASSMNT	A	CL_QUALITY_LEVEL	Valeur énumérée sur le niveau de qualité
CONF_POLICY	A	text	Mention juridique relative aux droits d'usage de ces données
CONF_STATUS_OBS	A	CL_CONF_STATUS	Droit d'accès sur les données.
METADATA_ACCESS_RIGHT	A	booléen	Droit d'accès aux métadonnées (libre ou non).
Niveau : Series			
INDICATOR_ID	D	text	Code de l'indicateur.
TIME_PERIOD	D	Date	Date ou période de validité de l'indicateur.
TITLE	A	text	Nom de l'indicateur
DATA_UPDATE	A	Date	Date d'acquisition de l'indicateur (identique à celle du jeu de données)
STAT_CONC_DEF	A	text	Résumé décrivant l'indicateur.
UNIT_MEASURE	A	text	Unité de mesure de l'indicateur
DECIMALS	A	CL_DECIMALS	Nombre de décimaux après la virgule.
UNIT_MULT	A	CL_UNIT_MULT	Exposant de 10, multipliant la valeur
CURRENCY	A	CL_CURRENCY	Devise de la variable - Optionnel
COLL_METHOD	A	text	Méthodologie pour le calcul de cet indicateur.
THESAURUS	A	text	Nom d'un thésaurus.
THEME_CODE	A	text	Code d'un thème sélectionné dans le thésaurus pour décrire l'indicateur.
KEYWORD	A	text	Mot-clé décrivant l'indicateur.

Rappels de statistiques

La recherche de valeurs exceptionnelles repose généralement sur l'emploi de méthodes mathématiques issues de la statistique et de la théorie des probabilités [Saporta 06], dont les résultats sont présentés à l'utilisateur sous une forme graphique lisible et explicite. Ce chapitre présente un rappel des notions de base de la statistique descriptive utiles à la compréhension de méthodes plus sophistiquées.

La démarche générale d'une étude statistique consiste à :
- observer la distribution des données pour chaque variable,
- remarquer si les variables présentent des liaisons,
- émettre des hypothèses portant sur la forme des lois de distribution à employer, sur l'existence de relations d'interdépendance,
- déterminer le test de significativité le plus pertinent.

Les quelques explications qui suivent montrent que cette phase d'étude préliminaire à l'usage de méthodes statistiques plus avancées ne doit pas être négligée.

Une variable X peut être :
- *quantitative* : numérique, à valeur dans l'ensemble des réels $\mathbb{R}$, comme par exemple, l'âge, la taille, le poids, le nombres d'heures, le Produit Intérieur Brut, les revenus, le nombre de chômeurs, etc.
- *qualitative* : à valeur dans un ensemble quelconque ou muni d'une structure d'ordre (qualitative ordinale), comme par exemple, le genre sexuel des personnes (féminin ou masculin) ou bien les Professions et Catégories Socioprofessionnelles (PCS) (cadre d'entreprise, agriculteur exploitant, artisan, ouvrier qualifié, etc.).

Ensuite, en traitant des variables quantitatives, et lorsqu'on s'intéresse à la distribution des valeurs, il s'agit d'abord de déterminer si la variable est discrète ou continue.
- La variable est *discrète* si elle ne prend qu'un nombre fini de valeurs. Ses valeurs (ou modalités) sont notées x_i. Par exemple, le numéro qui peut être obtenu suite à un jet de dé à 6 faces est une variable discrète, qui prend ses valeurs dans l'ensemble $1, 2, 3, 4, 5, 6$.
- La variable est *continue* si elle prend un nombre infini de valeurs. Par exemple, le revenu est une variable continue.

La définition d'une variable continue peut être précisée, sur la base d'une *loi de probabilité*. En théorie des probabilités, un *évènement* se décrit comme un sous-ensemble « mesurable » d'un « univers des possibles ». Les évènements sont des objets auxquels sont associées des probabilités P. On nomme alors probabilité une application d'un ensemble d'évènements à valeurs dans le segment $[0, 1]$, et l'image par cette application d'un évènement est appelée probabilité de l'évènement. La somme des probabilités de tous les évènements possibles est toujours égale à 1. La probabilité s'interprète comme une *mesure de l'incertitude* sur la réalisation d'un évènement, et la répétition d'une expérience doit conduire à vérifier qu'en général tel ou tel évènement est plus ou moins fréquent, et qu'une loi de probabilité peut en être

déduite [5]. Par exemple, si X est l'ensemble des revenus des ménages d'une population, le fait que X soit inférieur à 1500 euros par mois est un évènement, dont on voudrait estimer la probabilité de réalisation dans la population étudiée à partir des mesures x_i du revenu dont on dispose : on cherche à évaluer $P(X < 1500)$. La *fonction de répartition* d'une variable aléatoire X est l'application F de $\mathbb{R}$ dans $[0, 1]$ définie par :

$$F(x) = P(X < x). \tag{8.1}$$

Les variables discrètes ou continues ont une fonction de répartition, voir figure 8.3.

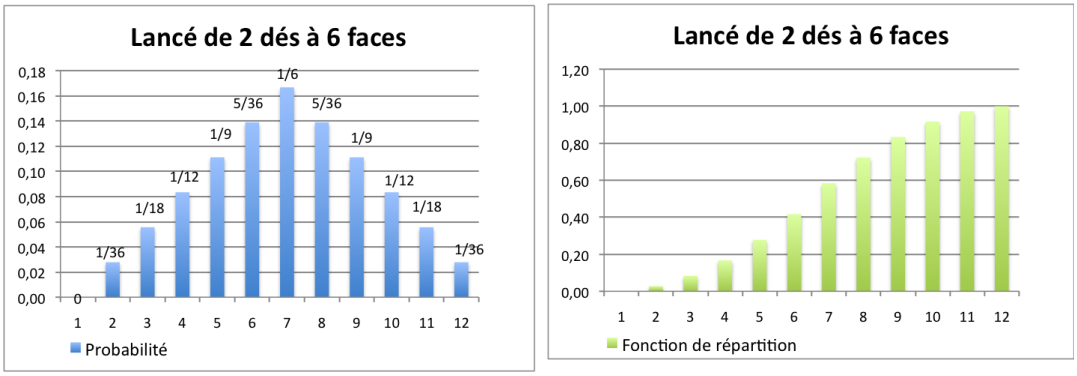


FIGURE 8.3 – Loi de probabilité et fonction de répartition d'une variable discrète.

Cependant, seules les variables continues, ou plus exactement absolument continues, admettent une *densité de probabilité*, notée f. F est alors dérivable, et admet f pour dérivée. La probabilité de réalisation de X dans l'intervalle $[a, b]$ est définie par :

$$P(a < X < b) = \int_a^b f(x).dx = F(b) - F(a). \tag{8.2}$$

Cette probabilité correspond à la surface sous la fonction de densité prise entre les bornes a et b, voir 8.4

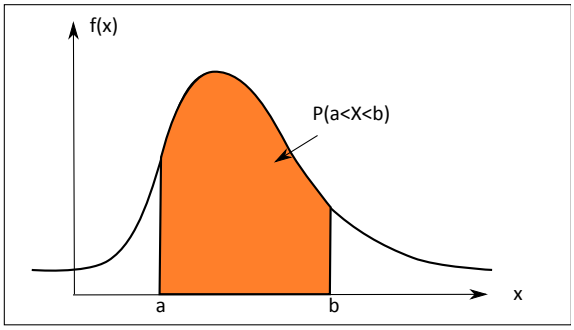


FIGURE 8.4 – Probabilité et fonction de densité d'une variable continue.

5. Ces hypothèses sont issues de la théorie fréquentiste, le courant historiquement dominant en statistiques, bien qu'à l'heure actuelle, des approches basées sur le modèles bayésiens soient également promues avec succès.

Une distribution X de valeurs quantitatives peut être caractérisée par les valeurs suivantes :
- le nombre de valeurs dans X, ou *effectif*, noté n
- la *moyenne* arithmétique, notée $\bar{x}$, est fonction de toutes les observations mais est sensible aux valeurs extrêmes :

$$\bar{x} = \frac{1}{n} \sum_i x_i. \tag{8.3}$$

- la *médiane*, un paramètre de position qui permet de couper la population étudiée en deux groupes contenant le même nombre d'individus : $F(M) = 0.5$. La médiane est un indicateur de position insensible aux variations des valeurs extrêmes.
- le *minimum* x_{min} : $\forall x_i \in X, x_{min} \leq x_i$
- le *maximum* x_{max} : $\forall x_i \in X, x_{max} \geq x_i$
- l'étendue w est dépendante des valeurs extrêmes, c'est un indicateur instable : $w = x_{max} - x_{min}$
- le premier quartile Q_1 : F(Q1) = 0.25
- le troisième quartile Q_3 : F(Q3) = 0.75
- *l'intervalle interquartile* $|Q_3 - Q_1|$ est un indicateur parfois utilisé pour mesurer la dispersion : il est plus robuste que l'étendue.
- *l'espérance* qui se calcule différemment selon que la variable est dite discrète ou continue.
  - pour une variable discrète, l'espérance est la moyenne arithmétique des différentes valeurs de X, pondérées par leurs probabilités [6] :

$$E(X) = \sum_i x_i P(X = x_i) \tag{8.4}$$

  - pour une variable continue admettant une fonction de densité f :

$$E(X) = \int_{\mathbb{R}} x dP_X(x).dx = \int_{\mathbb{R}} x f(x).dx. \tag{8.5}$$

    Si X est positive l'espérance s'interprète comme l'aire située entre l'horizontale y=1 et la fonction de répartition :

$$E(X) = \int_0^{\infty} (1 - F(x)).dx \tag{8.6}$$

- la *variance* σ^2, qui est le moment centré d'ordre 2 d'une distribution, donne une mesure de la dispersion de X autour de la moyenne $\bar{x}$. L'écart-type, noté σ, en est la racine carrée.
  - pour une variable discrète, la variance est la distance euclidienne des valeurs à la moyenne, pondérées par leurs probabilités [7] :

$$\sigma^2 = \sum_i p_i (x_i - \bar{x})^2 \tag{8.7}$$

  - pour une variable continue admettant une fonction de densité f :

$$\sigma^2 = E[(X - \bar{x})^2] = \int_{\mathbb{R}} (x - \bar{x})^2 dP_X(x).dx \tag{8.8}$$

- le *coefficient de variation* CV exprime en pourcentage le rapport entre l'écart-type et la moyenne. Il n'a de sens que si $\bar{x} > 0$:

$$CV = \sigma / \bar{x} \tag{8.9}$$

6. Si la variable est équiprobable, $P(X = x_i) = 1/n$, alors l'espérance se ramène à la moyenne $\bar{x}$ de l'échantillon.
7. Si la variable est équiprobable, $P(X = x_i) = 1/n$, alors la variance se ramène à $\sigma^2 = 1/n \sum_i (x_i - \bar{x})^2$

– le *coefficient d'asymétrie* γ_1 :

$$\gamma_1 = \sum_i p_i (x_i - \bar{x})^3 / \sigma^3 \tag{8.10}$$

– le *coefficient d'aplatissement* γ_2, qui vaut 3 en théorie dans le cas de la loi normale (ou Laplace-Gauss) :

$$\gamma_2 = \sum_i p_i (x_i - \bar{x})^4 / \sigma^4 \tag{8.11}$$

γ_2 mesure l'importance des « queues » de distribution, c'est-à-dire la quantité de valeurs loin de la valeur moyenne. De plus, $\gamma_2 \geq 1 + \gamma_1^2$

Les coefficients d'asymétrie et d'aplatissement sont des caractéristiques à prendre en compte lors de la réalisation de tests de distribution, surtout lorsque ces tests supposent une distribution gaussienne de la variable, donc avec des queues de distribution faibles et une distribution centrée autour de la moyenne des valeurs, voir figure 8.5. Par exemple, dans le cas d'une distribution montrant une forte asymétrie (γ_1 loin de 0), avec une queue de distribution importante ($\gamma_2 < 3$), l'hypothèse de distribution gaussienne peut amener à considérer comme anormale des valeurs fortes en queue de distribution.

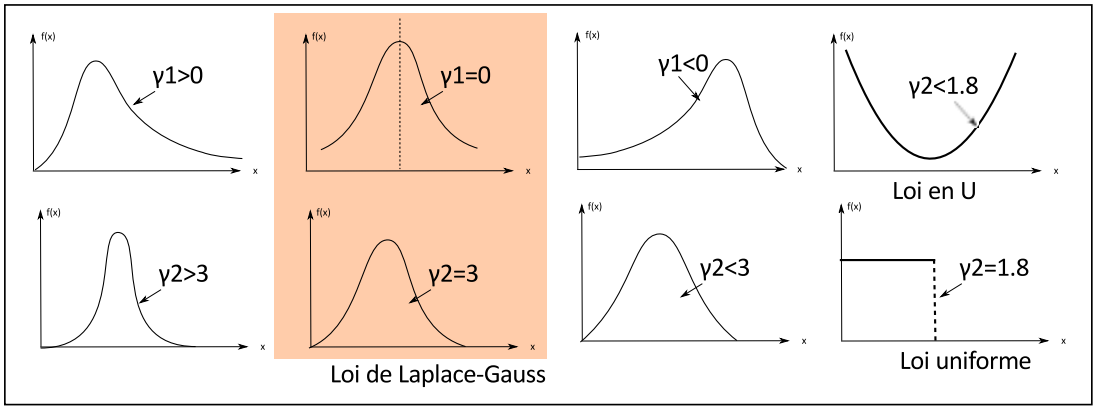


FIGURE 8.5 – Formes de distribution et coefficients de forme associés (γ_1 et γ_2).

Avant d'entamer le calcul d'indicateurs statistiques plus poussés (tels que le calcul de corrélation entre variables), il s'agit de se questionner sur la forme de la distribution. À cet effet, l'histogramme et le polygone de fréquences sont deux représentations de la distribution utiles.

Analogue à la courbe de densité d'une variable aléatoire, un histogramme est une représentation graphique à barres verticales accolées (en tuyaux d'orgue), obtenu après découpage en classes des observations d'une variable continue. Souvent, les « tuyaux » sont accolés pour montrer la continuité de la variable. La hauteur du tuyau est proportionnelle à la fréquence de la classe correspondante. La surface sous l'histogramme vaut toujours 1. La détermination du nombre de classes d'un histogramme est délicate et on ne dispose pas de règles absolues. Un trop faible nombre de classes fait perdre de l'information et aboutit à gommer les différences pouvant exister entre des groupes de l'ensemble étudié. En revanche un trop grand nombre de classes aboutit à des graphiques incohérents : certaines classes deviennent vides ou presque car n est fini.

Le polygone de fréquences est une autre représentation graphique (en ligne brisée) de la distribution de fréquences d'une variable quantitative. Pour tracer le polygone, on joint les points milieu du sommet des rectangles adjacents par un segment de droite. Le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal. Le polygone de fréquences peut être fort utile lorsqu'il s'agit de com-

parer plusieurs populations ou lorsque que le nombre de classes et de sujets est élevé et qu'un certain «polissage» s'avère pertinent.

Avec ces deux représentations (voir figure 8.6), l'utilisateur visualise la distribution des données et peut construire des hypothèses au sujet de sa loi de densité. En particulier, il vérifie si la distribution est uni-modale ou multi-modale, ce qui correspond à l'existence d'un seul ou de plusieurs pics dans la courbe de distribution des données. La plupart des outils que nous présentons n'adressent que des distributions uni-modales. Dans le cas de distributions multi-modales, il faut abandonner, ou s'adresser à un statisticien pour traiter ce cas particulier.

La deuxième chose très importante qu'un histogramme apprend sur les données, c'est si la distribution des données est gaussienne. Il existe dans tous les outils statistiques de nombreux tests permettant de vérifier la normalité [8] des distributions. Par exemple, dans R, qqplot ou qqnorm sont deux méthodes qui permettent de tester la normalité d'une distribution. Également, le test de Shapiro-Wilk teste si un échantillon suit une loi normale ou non : dans R, shapiro.test vérifie l'hypothèse nulle que l'échantillon suit une loi normale, donc si p-value < 0.01, l'échantillon ne suit pas une loi normale. Lorsqu'une distribution est gaussienne, on peut alors calculer les coefficient de corrélation linéaire, sinon, on peut tenter de se ramener à un cas gaussien en prenant le logarithme des données. Enfin, si X est définitivement non-gaussienne, mais monotone, il est possible d'utiliser des méthodes de rangs (et calculer alors le coefficient de Spearman pour vérifier sa corrélation avec une autre variable ordinale).

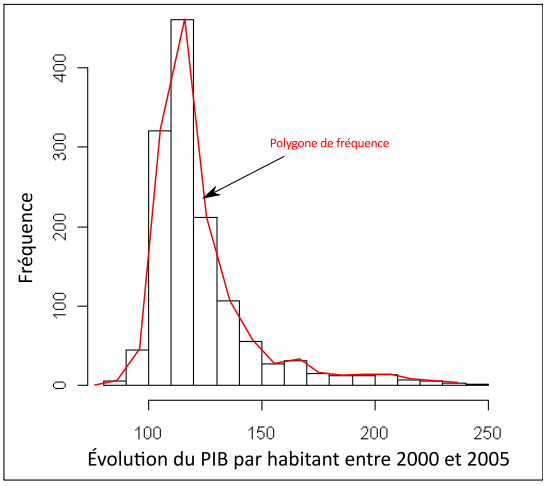


FIGURE 8.6 – Exemple d'histogramme et de polygone de fréquences.

L'observation d'une distribution permet d'émettre des hypothèses concernant le modèle de distribution. Nous rappelons ici qu'un *modèle* est une formulation mathématique d'une relation existant entre au moins deux variables. Les mesures dont l'utilisateur dispose ne concordent peut-être pas exactement avec ce modèle : l'écart à ce modèle est mesuré par les *résidus*. L'objectif d'un "bon" modèle est de réduire au maximum la valeur de ces résidus. Le plus souvent, on cherchera à minimiser la somme du carré de ces écarts, qui reste une valeur toujours positive ou nulle (seulement si le modèle est parfait). La figure 8.7 illustre un modèle très simple (une fonction affine), et la partie colorée en jaune représente la surface à minimiser.

8. En statistiques, « normalité » s'entend au sens de caractère gaussien.

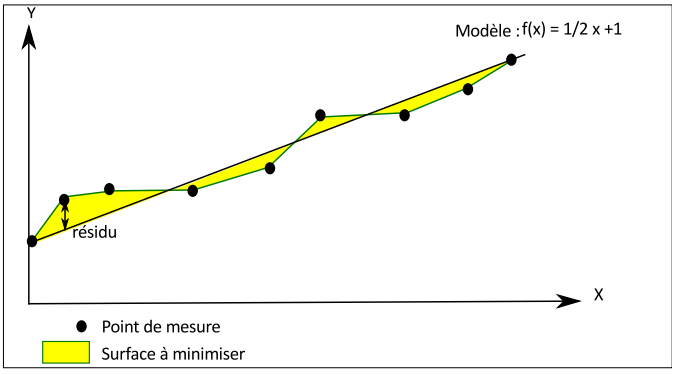


FIGURE 8.7 – Modèle et résidus - une illustration.

La mesure de liaison entre variables est un aspect essentiel de l'analyse statistique. La tendance à évoluer de façon conjointe ou au contraire de façon opposée (c'est-à-dire toute forme de non-hasard) constitue l'objet de l'*étude des corrélations*. Les méthodes et les indices de dépendances varient suivant la nature (qualitative, ordinale, numérique) des variables étudiés. Dans le cadre de ce rappel, certaines de ces méthodes sont expliquées pour le seul cas des variables numériques. Des explications complètes sont disponibles dans [Saporta 06].

On suppose ici qu'on observe pour n individus deux variables X et Y. On a donc n couples $(x_i; y_i)$ ou encore deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^n$, avec :

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Afin d'examiner s'il existe une liaison entre X et Y, chaque observation i est représentée comme un point de coordonnées (x_i, y_i) dans un repère cartésien. On appelle cette représentation « le diagramme de dispersion ». La forme du nuage de points tracé est fondamentale pour la suite (figure 8.8). On dit qu'il y a *corrélation* si il y a dépendance en moyenne : à $X = x$ fixé, la moyenne $\bar{Y}$ est fonction de x. Si cette liaison est approximativement linéaire, on se trouve dans le cas de la *corrélation linéaire*.

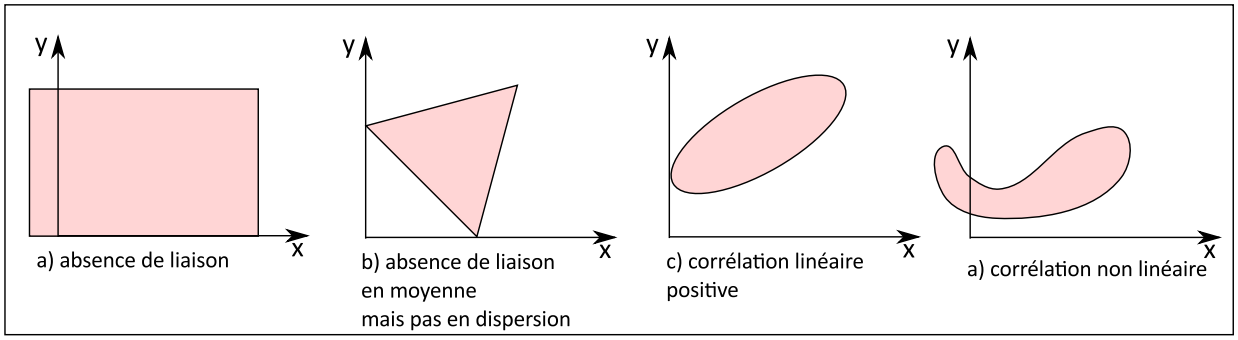


FIGURE 8.8 – Forme du nuage de dispersion et corrélation.

Le *coefficient de corrélation linéaire* ρ, dit de « Bravais-Pearson », mesure exclusivement le caractère plus ou moins linéaire du nuage de points sous l'hypothèse que X et Y sont gaussiennes. Il se calcule comme le quotient de la covariance entre X et Y par le produit de leurs écart-types respectifs :

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - E(X))(Y - E(Y)))}{\sigma_X \sigma_Y} \tag{8.12}$$

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sigma_x \sigma_y} \tag{8.13}$$

Les valeurs limites -1 et +1 sont atteintes si et seulement si il existe une relation linéaire entre X et Y. Pour deux variables indépendantes, ρ = 0, mais la réciproque est fausse. De plus la nullité de ce coefficient exclut la relation linéaire mais n'exclut pas l'existence d'autres relations.

Il existe une interprétation géométrique de ces résultats en se plaçant dans un espace de Hilbert L^2, définit comme l'ensemble de toutes les variables aléatoires définies sur un même univers, muni d'un produit scalaire $< X, Y >= E(X,Y)$ et de la norme $||X|| = (E(X^2))^{1/2}$. L'écart-type σ est donc la norme des variables centrées [9], et la covariance $cov(X,Y)$ le produit scalaire des variables centrées X et Y. On a alors que ρ mesure le cosinus de l'angle entre X-E(X) et Y-E(Y), donc entre les deux vecteurs

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} et \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}.$$

Leur orthogonalité signifie alors l'absence de relation linéaire alors que la colinéarité des vecteurs est synonyme de corrélation linéaire. Ce coefficient ρ est très sensible aux valeurs extrêmes, il n'est donc pas robuste. Pas ailleurs, son usage doit être réservé à des nuages où les points sont répartis de part et d'autres d'une tendance linéaire (cas c) dans la figure 8.8), sinon on s'expose à des interprétations fausses du coefficient. Enfin, la corrélation n'est pas transitive : **x** très corrélé avec **y** et **y** très corrélé avec **z** n'implique pas que **x** soit corrélé avec **z**.

Ces résultats se généralisent à p variables X_j, $j = 1..p$, pour une matrice X à n lignes, et p colonnes, de terme x_i^j (valeur de l'individu i pour la variable j). La *matrice V de variance-covariance* des p variables s'écrit alors :

$$\mathbf{V} = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ . & s_2^2 & \dots & s_{2p} \\ . & & & . \\ . & & & . \\ . & & & s_p^2 \end{bmatrix} \text{ avec } s_{kl} = \frac{1}{n}\sum_{i=n}^{n} x_i^k x_i^l - \bar{x^k}\bar{x^l} \tag{8.14}$$

La matrice regroupant tous les coefficients de corrélation linéaire entre les p variables prises deux à deux est notée **R**.

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ . & 1 & \dots & s_{2p} \\ . & & & . \\ . & & & . \\ r_{p1} & & & 1 \end{bmatrix}$$

9. La variable X est dite centrée si à tous les individus x_i on retire la valeur de la moyenne $\bar{x}$, elle est dite réduite si les valeurs x_i de tous les individus sont divisées par l'écart-type des individus

En posant $D_{1/s}$ la matrice diagonale des inverses des variances :

$$\mathbf{D_{1/s}} = \begin{bmatrix} 1/s_1 & 0 & \dots & 0 \\ 0 & 1/s_2 & \dots & 0 \\ . & & & . \\ . & & & . \\ 0 & & & 1/s_p \end{bmatrix}$$

On a : $\mathbf{R} = \mathbf{D_{1/s}} \mathbf{V} \mathbf{D_{1/s}}$

$\mathbf{R}$ est identique à la matrice de variance-covariance des données centrées et réduites. $\mathbf{R}$ résume la structure des dépendances linéaires ente les p variables. Comme $\mathbf{V}$, $\mathbf{R}$ est une matrice symétrique positive.

Ainsi, en statistiques, l'étude des variables amène souvent l'analyste à mesurer le coefficient de corrélation entre des variables d'intérêt. Cependant, ce n'est qu'en réitérant un nombre élevé de fois cette mesure qu'il peut être un peu certain du caractère non aléatoire des corrélations observées. En effet, il faut bien remarquer que cette mesure est elle-même aléatoire, et on voudrait déterminer la distribution de probabilité de la variable R qui correspond à cet échantillonage. En admettant être dans le cas où la mesure de ρ est justifiée, il s'agit de déterminer la valeur significative de ρ à partir de laquelle les variables sont effectivement corrélées. Lorsque les observations proviennent d'un couple gaussien de variables indépendantes, on peut connaitre facilement la loi de distribution de R et sa fonction de densité. Ainsi au risque 5% on déclarera qu'une liaison est significative sur un échantillons de 30 observations si $|\rho| > 0.36$. Sinon, si ρ n'est pas nul, on utilise des approximations qui mettent en jeu la transformée de Fisher de R. Cette transformation permet de tester des valeurs a priori pour ρ et de trouver des intervalles de confiance pour ρ à partir de R. Lorsque le couple (X, Y) n'est pas gaussien, les résultats précédents restent valables à condition que n soit grand ($n > 30$). Le seuil de signification décroît quand n croît ; cependant, le fait de trouver que ρ diffère significativement de 0 ne garantit nullement que la liaison soit forte. Cette remarque est cruciale et a rapport avec la « *significativité* » du test. Ce problème est souligné par Openshaw [Openshaw 96] qui indique que la grande majorité des consommateurs de statistiques se contentent de ces seuils, sans comprendre les hypothèses sous-jacentes.

Enfin, certaines de ces analyses peuvent être transposées au cas de variables qualitatives, modulo quelques adaptations. Par exemple, la liaison entre des variables ordinales peut être évaluée par le coefficient de Spearman. De même, l'analyse des correspondances permet d'étudier la liaison entre deux variables qualitatives. Sur le plan mathématique, l'analyse des correspondances peut être considérée comme l'analyse en composantes principales, mais avec une métrique spéciale, la métrique du χ^2, [Saporta 06].

Bibliographie

[Abadie 08] Nathalie Abadie & Sébastien Mustière. *Création d'une taxonomie géographique à partir des spécifications de bases de données*. In Colloque International de Géomatique et d'Analyse Spatiale SAGEO'08, Montpellier, France, 2008.

[Aisenor 07] Aisenor. *Reading the INSPIRE Metadata Draft*, April 2007.

[Albrecht 07] Jochen Albrecht. *Dynamic GIS*. In J. Wilson & Stewart A. Fotheringham, editeurs, Handbook of GIScience, pages 436–446. Blackwell, London, 2007.

[Alfred 04] Arta Dilo Alfred & Alfred Stein. *Definition and Implementation of Vague Objects*. Rapport technique, International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, Netherlands, 2004.

[Allen 83] James F. Allen. *Maintaining knowledge about temporal intervals*. Commun. ACM, vol. 26, no. 11, pages 832–843, 1983.

[Allen 84] James F. Allen. *Towards a general theory of action and time.* Artificial Intelligence, vol. 23, pages 123–154, 1984.

[Allen 94] James F. Allen & George Ferguson. *Actions and events in interval temporal logic.* Technical report tr521, Computer Science Department, University of Rochester, 1994.

[Anderson 71] Theodore W. Anderson. The statistical analysis of time series. Wiley, New York, NY, USA, 1971.

[Andrienko 01] Natalia Andrienko, Gennady Andrienko & Peter Gatalsky. *Exploring change in census time series with interactive dynamic maps and graphics*. Computational Statistics, vol. 16, no. 3, pages 417–433, 2001.

[Andrienko 03a] G. Andrienko, N. Andrienko & V. Gitis. *Interactive maps for visual exploration of grid and vector geodata*. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 57, no. 5, pages 380–389, April 2003.

[Andrienko 03b] Natalia Andrienko, Gennady Andrienko & Peter Gatalsky. *Exploratory spatio-temporal visualization : an analytical review*. Journal of Visual Languages & Computing, vol. 14, no. 6, pages 503 – 541, 2003. Visual Data Mining.

[Andrienko 06] Natalia Andrienko & Gennady Andrienko. Exploratory analysis of spatial and temporal data. Springer, 2006.

[Angles 08] Renzo Angles & Claudio Gutierrez. *Survey of graph database models*. ACM Comput. Surv., vol. 40, pages 1 :1–1 :39, February 2008.

[Anselin 89] Luc Anselin. *What is Special About Spatial Data ? Alternative Perspectives on Spatial Data Analysis*. In D.A. Griffith, editeur, Spatial Statistics, Past, Present and Future, pages 63–77. Institut of Mathematical Geography, Ann Arbor, MI, 1989.

[Anselin 92] Luc Anselin. *SPACESTAT : a program for the analysis of spatial data*. Rapport technique, NCGIA, Santa Barbara, California, USA, 1992.

[Anselin 93] Luc Anselin. *Exploratory spatial data analysis and geographic information systems*. In New tools for spatial analysis, pages 45–54. Eurostat, Luxembourg, 1993.

[Anselin 95] Luc Anselin. *Local indicators of spatial association-LISA*. Geographical Analysis, vol. 2, pages 93–115, 1995.

[Anselin 96] Luc Anselin. *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. In Pete F. Fisher, H.J Scholten & D. Unwin, editeurs, Spatial Analytical Perspectives on GIS, pages 111–125. Taylor & Francis, London, 1996.

[Anselin 04] Luc Anselin, Ibnu Syabri & Youngihn Kho. *GeoDa : An Introduction to Spatial Data Analysis*. Geographical Analysis, vol. 38, pages 5–22, 2004.

[Arbia 89] Giuseppe Arbia. *Statistical Effect of Data Transformations : A Proposed General Framework*. In Michael F Goodchild & Sucharita Gopal, editeurs, The Accuracy of Spatial Data Bases, pages 249–259. Taylor & Francis, London, UK, 1989.

[Arel 02] Dominique Arel. *Démographie et politique dans les premiers recensements post-soviétiques : méfiance envers l'État, identités en question*. Population, vol. 57, no. 2, pages 791–820, 2002.

[Armstrong 88] Marc P. Armstrong. *Temporality in Spatial Databases*. In GIS/LIS 88 Proceedings : Accessing the World, Volume II., pages 880–889, Falls Church, Virginia, 1988. American Society for Photogrammetry and Remote Sensing.

[Arnaud 00] Michel Arnaud & Xavier Emery. Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatiques. Hermès, Paris, France, 2000.

[Badard 00] Thierry Badard. *Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques.* PhD thesis, Université de Marne la Vallée, Marne-la-Vallée, France, 2000.

[Badran 97] F. Badran, P. Daigremont & S. Thiria. *Régression par carte topologique*. In Statistiques et méthodes neuronales. S. Thiria et al, 1997.

[Bancillon 91] François Bancillon, Claude Delobel & Paris C. Kanellakis. Building an object-oriented database system : The story of o2. Morgan Kaufmann, 1991.

[Banerjee 09] Sandipto Banerjee & Karen C. Davis. *Modeling Data Warehouse Schema Evolution over Extended Hierarchy Semantics*. J. Data Semantics, vol. 13, pages 72–96, 2009.

[Banos 01] Arnaud Banos. *A propos de l'analyse spatiale exploratoire des données*. Cybergeo : European Journal of Geography, Systèmes, Modélisation, Géostatistiques, no. 197, 2001.

[Barde 05] Julien Barde. *Mutualisation de Données et de Connaissances pour la Gestion Intégrée des Zones Côtières. Application au Projet SYSCOLAG.* PhD thesis, Université Montpellier II, 2005.

[Bastien 01] Christian Bastien & D. L. Scapin. *Évaluation des systèmes d'information et critères ergonomiques.* In C. Kolski, editeur, Systèmes d'information et interactions homme-machine. Environnement évolués et évaluation de l'IHM. Interaction homme-machine pour les SI, volume 2, pages 53–79. Hermès, 2001.

[Bédard 97] Yvan Bédard. *Spatial OLAP*. In 2ème Forum annuel sur la R-D, Géomatique VI : Un monde accessible, Montréal, 13-14 Novembre 1997.

[Bédard 01] Yvan Bédard, T. Merrett & J. Han. Geographic data mining and knowledge discovery., chapitre Fundamentals of spatial data warehousing for geographic knowledge discovery., pages 53–73. Taylor & Francis, London, 2001.

[Bédard 04] Yvan Bédard, Suzie Larrivée, Marie-Josée Proulx & Martin Nadeau. *Modeling Geospatial Databases with Plug-Ins for Visual Languages : A Pragmatic Approach and the Impacts of 16 Years of Research and Experimentations on Perceptory*. In ER (Workshops), pages 17–30, 2004.

[Bel Adj Ali 99] Atef Bel Adj Ali & François Vauglin. *Geometric Matching of Polygons in GISs and assessment of Geometrical Quality of Polygons*. In Michael Goodchild Wenzhong Shi & Peter Fisher, editeurs, International Symposium on Spatial Data Quality'99, pages 33–43, Hong Kong Polytechnic University, 1999.

[Bel Adj Ali 01] Atef Bel Adj Ali. *Qualité géométrique des entités géométriques surfaciques. Application à l'appariement et définition d'une typologie des écarts géométriques.* PhD thesis, Université de Marne la Vallée, IGN, 2001.

[Beller 91] Aaron Beller, Tom Giblin, Khanh V. Le, Steve Litz, Tim Kittel & David Schimel. *A temporal GIS prototype for global change research*. In GIS/LIS'91, volume 2, pages 752–765, 1991.

[Belussi 99] Alberto Belussi, Mauro Negri & Giuseppe Pelagatti. *Management of data changes in geodatabases : time component in GIS*. Geomatics Info Magazine International, vol. 13, no. 7, pages 41–43, 1999.

[Ben Rebah 08] Maher Ben Rebah. *La cartographie dynamique comme outil d'investigation territoriale : le cas du découpage territorial en Tunisie*. PhD thesis, Université de Paris VII, 2008.

[Ben Rebah 11] Maher Ben Rebah, Christine Plumejeaud, Ronan Ysebaert & Didier Peeters. *Towards an approach of time series data issue : from empirical methods to applications*. Technical report tr2, ESPON Monitoring Commitee, 2011.

[Bergeron 92] Marcel Bergeron. *Vocabulaire de la géomatique*, 1992.

[Berry 68] Brian Berry. Spatial analysis : a reader in statistical geography. Prentice-Hall, 1968.

[Bertin 67] Jacques Bertin. Sémiologie graphique. Mouton/Gauthier-Villars, Paris, 1967.

[Billen 02] Roland Billen. *Nouvelle perception de la spatialité des objets et de leurs relations. Développement d'une modélisation tri-dimensionelle de l'information spatiale.* PhD thesis, Université de Liège, 2002.

[Billen 04] Roland Billen & Eliseo Clementini. *Étude des caractéristiques projectives des objets spatiaux et de leurs relations*. Revue Internationale de Géomatique, vol. 14, pages 145–165, 2004.

[Bimonte 07] Sandro Bimonte. *Intégration de l'information géographique dans les entrpôts de données et l'analyse en ligne : de la modélisation à la visualisation.* PhD thesis, Institut National des Sciences Appliquées de Lyon, 2007.

[BIPM 83] BIPM. *Résolution 1 de la 17e réunion de la CGPM (1983)*, 1983.

[Bivand 08] Roger Bivand, Edzer J. Pebesma & Virgilio Gomez-Rubio. Applied spatial data analysis with r. Use R ! Springer, 2008.

[Body 03] Mathurin Body, Maryvonne Miquel, Yvan Bedard & Anne Tchounikine. *Handling Evolutions in Multidimensional Structures*. In IEEE 19th International Conference on Data Engineering (ICDE), 2003.

[Booch 99] Grady Booch, James E. Rumbaugh & Ivar Jacobson. The unified modeling language user guide. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, 1999.

[Box 94] George E.P. Box, Gwilym M. Jenkins & Gregory Reinsel. Time series analysis, forecasting and control. Prentice Hall, 3rd edition edition, 1994.

[Bracken 89] I Bracken & D Martin. *The generation of spatial population distributions from census centroid data*. Environment and Planning A, vol. 21, no. 4, pages 537–543, 1989.

[Brilhante 06] V. Brilhante, A. Ferreira, J. Marinho & J. S. Pereira. *Information Integration through Ontology and Metadata for Sustainability Analysis.* In International Environmental Modelling and Software Society (iEMSs) Third Biannual Meeting "Summit on Environmental Modelling and SoftwareThird Biennial Meeting", Burlington, USA, 2006.

[Brunet 92] Roger Brunet, R Ferras & H Théry. Les mots de la géographie - dictionnaire critique. Reclus-Documentation Française, Paris & Montpellier, 1992.

[Brunet 97] Roger Brunet, Jean-Christophe François & Claude Grasland. *Entretien avec Roger Brunet : La discontinuité en géographie ; origines et problèmes de recherche*. L'Espace Géographique, vol. 4, pages 297–308, 1997.

[Brunsdon 96] Chris Brunsdon & Martin Charlton. *Developing an exploratory spatial analysis system in XLisp-Stat*. In D. Parker, editeur, Innovatons in GIS 3, pages 135–146. Taylor and Francis, 1996.

[Camagni 10] Roberto Camagni, Andreu Ulied, Mark Schucksmith & Franck Bruinsma. *TIPTAP : Territorial Impact Package for Transport and Agricultural Policies*, Fevrier 2010.

[Cauvin 76] Colette Cauvin & Sylvie Rimbert. Les méthodes de la cartographie thématique, fascicule i : La lecture numérique des cartes thématiques. Éditions Universitaires de Fribourg, 1976.

[Cauvin 97] Colette Cauvin. *Au sujet des transformations cartographiques de position*. Cybergeo : Revue européenne de Géographie, no. 15, 1997.

[C.E. 07] C.E. *Commission Européenne - Draft Implementing Rules for Metadata*, December 2007.

[Chareille 04] Pascal Chareille, Xavier Rodier & Elisabeth Zadora-Rio. *Analyse des transformations du maillage paroissial et communal en Touraine à l'aide d'un SIG*. Histoire & mesure, vol. 19, no. 3/4, pages 317–344, 2004.

[Charleux 05] Laure Charleux. *GWR, MAUP et lissage par potentiels*. Revue Internationale de Géomatique, vol. 15, no. 2, pages 195–209, 2005.

[Charre 95] Joel Charre. Statistique et territoire. Espace modes d'emploi. GIP Reclus, 1995.

[Chengalur-Smith 99] InduShobha N. Chengalur-Smith, Donald P. Ballou & Harold L. Pazer. *The Impact of Data Quality Information on Decision Making : An Exploratory Analysis*. IEEE Transactions on Knowledge and Data Engineering, vol. 11, pages 853–864, 1999.

[Chenu 97] A Chenu. *La catégorisation statistique - Présentation du dossier*. Sociétés contemporaines, vol. 26, pages 2–4, 1997.

[Cheylan 93] Jean-Paul Cheylan & Sylvie Lardon. *Towards a Conceptual Data Model for the Analysis of Spatio-Temporal Processes : The Example of the Search for Optimal Grazing Strategies*. In COSIT, pages 158–176, 1993.

[Cheylan 97] Jean-Paul Cheylan, Sylvie Lardon, Helene Mathian & Léna Sanders. *Les problématiques liées au temps dans les SIG*. Revue Internationale de Géomatique, no. 4, pages 287–305, 1997.

[Cheylan 99] Jean-Paul Cheylan, Denis Gautier, Sylvie Lardon, Thérèse Libourel, Helene Mathian, Serge Motet & Léna Sanders. *Les mots du traitement de l'information spatio-temporelle*. Revue internationale de Géomatique, vol. 9, no. 1, pages 11–23, 1999.

[Chignoli 97] Robert Chignoli, Pierre Crescenzo & Philippe Lahire. *Lien entre classes dans les langages à objets*. Rapport de recherche 97-22, Laboratoire I2S, Université de Nice-Sophia-Antipolis, 1997.

[Chrisman 84] Nicholas R. Chrisman. *The role of quality information in the long-term functioning of a geographic information system*. Cartographica, vol. 21, pages 79–87, 1984.

[Christophle 09] Eva Christophle. *Conception d'un catalogue de données pour la mutualisation et le partage de l'information environnementale*. Rapport de stage, LIG, Grenoble, juin 2009.

[Claramunt 95] Christophe Claramunt & Marius Thériault. *Managing Time in GIS : An Event-Oriented Approach*. In Proceedings of the International Workshop on Temporal Databases, pages 23–42, London, UK, 1995. Springer-Verlag.

[Clarke 81] Bowman L. Clarke. *A calculus of individuals based on 'connection'*. Notre Dame Journal of Formal Logic, vol. 22, no. 3, pages 204–218, 1981.

[Clarke 85] Bowman L. Clarke. *Individuals and Points*. Notre-dame journal of formal logic, vol. 26, no. 1, 1985.

[Clarke 95] Dereck G. Clarke & David .M. Clark. *Lineage*. In S.C. Guptill S.C. & Morrison J.L., editeurs, Elements of spatial data quality, pages 13–30. Oxford, Elsevier, 1995.

[Claval 68] Paul Claval. Régions, nations, grands espaces. Genin, Paris, 1968.

[Clementini 93] Eliseo Clementini, Paolino Di Felice & Peter van Oosterom. *A Small Set of Formal Topological Relationships Suitable for End-User Interaction*. In SSD, pages 277–295, 1993.

[Clementini 01] Eliseo Clementini & Paolino Di Felice. *A spatial model for complex objects with a broad boundary supporting queries on uncertain data*. Data and Knowledge Engineering, vol. 37, pages 285–305, 2001.

[Clementini 06] Eliseo Clementini & Roland Billen. *Modeling and computing ternary projective relations between regions*. IEEE Transactions on Knowledge and Data Engineering, vol. 18, pages 799–814, 2006.

[Clementini 08] Eliseo Clementini & Robert Laurini. *Un cadre conceptuel pour modéliser les relations spatiales.* Revue des Nouvelles Technologies de l'Information (RNTI), vol. E-14, pages 1–17, 2008.

[Cliff 81] Andrew D. Cliff & Keith Ord. Spatial processes. models and applications. Pion, Londres, 1981.

[Cliff 98] Andrew D. Cliff & Peter Haggett. *On complex geographical space : computing frameworks for spatial diffusion processes*. In PA Longley, SM Brooks, R MacDonnel & B. Macmillan, editeurs, Geocomputation : A primer, pages 231–256. John Wiley, New York, NY, USA, 1998.

[Clifford 97] James Clifford, Curtis Dyreson, Christian S. Jensen & Richard T. Snodgrass. *On the Semantics of "Now" in Databases*. ACM Transactions on Database Systems, vol. 22, pages 171–214, 1997.

[Cohn 96] Anthony G. Cohn & Nicholas Mark Gotts. *The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries.* In Burrough P.A. & Andrew U. Frank, editeurs, Geographic Objects with Indeterminate Boundaries., GISDATA Series, Vol. 2, pages 171–187. Taylor & Francis, 1996.

[Cohn 01] Anthony G. Cohn & Shyamanta M. Hazarika. *Qualitative spatial representation and reasoning : An overview.* Fundamenta Informaticae, vol. 46, pages 1–29, 2001.

[Colledge 98] Michael J. Colledge. *Statistical Integration Through Metadata Management*. International Statistical, vol. 67, no. 1, pages 79–98, 1998.

[Comber 05] Alexis Comber, Pete F. Fisher & Richard A. Wadsworth. *A comparison of statistical and expert approaches to data integration*. Journal of Environmental Management, vol. 77, pages 47–55, 2005.

[Comber 10] Alexis Comber, Andy Lear & Richard Wadsworth. *A comparison of different semantic methods for integrating thematic geographical information : the example of land cover*. In AGILE'2010, 2010.

[Cressie 91] Noel Cressie. Statistics for spatial data. Wiley, New York, 1991.

[Cui 93] Zhan Cui, Anthony G. Cohn & David A. Randell. *Qualitative and Topological Relationships in Spatial Databases*. In David J. Abel & Beng Chin Ooi, editeurs, Advances in Spatial Databases, Third International Symposium, SSD'93, volume 692 of *Lecture Notes in Computer Science*, pages 296–315, 1993.

[Daniel 08] Florian Daniel, Fabio Casati, Themis Palpanas, Oleksiy Chayka & Cappiello Cinzia. *Enabling Better Decisions through Quality-aware Reports*. In International Conference on Information Quality (ICIQ), 2008.

[D'Aubigny 94] C. D'Aubigny & Gérard D'Aubigny. *Agrégation spatiale et résumés statistiques*. Revue internationale de Géomatique, vol. 4, no. 3/4, pages 307–336, 1994.

[D'Aubigny 06] Gérard D'Aubigny. *Dépendance spatiale et autocorrélation*. In Jean-Jacques Droesbecke, Michel Lejeune & Gilbert Saporta, editeurs, Analyse statistique des données spatiales, pages 17–45. TECHNIP, 2006.

[DCMI 95] Dublin Core Metadata Initiative DCMI. *Metadata Terms (en ligne sur http ://-dublincore.org/)*, 1995.

[Dean 96] Pat Dean & Bo Sundgren. *Quality Aspects of a Modern Database Service (Position Paper)*. In Per Svensson & James C. French, editeurs, SSDBM, pages 156–161. IEEE Computer Society, 1996.

[Decroly 96] Jean-Michel Decroly & Claude Grasland. *Organisation spatiale et organisation territoriale des comportements démographiques : une approche subjective*. In Bocquet-Appel Jean-Pierre, Daniel Courgeau & Denise Pumain, editeurs, Spatial Analysis of Biodemographic Data. Analyse spatiale de données biodémographiques, pages 131–170. INED, 1996.

[Deichmann 01] Uwe Deichmann, Deborah Balk & Greg Yetman. *Transforming Population Data for Interdisciplinary Usages : From Census to Grid*. Rapport technique, Yale University, New Haven, CT, 2001.

[Dell'Erba 97] Edouard Dell'Erba & Thérèse Libourel. *Temps et évolution d'entités géoréférencées*. In Treizième journée de Bases de Données Avancées (BDA'97), Grenoble, 1997.

[Dempster 67] Arthur P. Dempster. *Upper and lower probabilities induced by a multivalued mapping*. Annales of mathematical statistics, vol. 38, no. 2, pages 325–339, 1967.

[Desconnets 07] Jean-Christophe Desconnets, Stéphane Clerc & Thérèse Libourel. *Cataloguer pour diffuser les ressources environnementales.* In XXVème congrès Inforsid, Perros Guirrec, France, 22-25 Mai 2007.

[Devillers 04] Rodolphe Devillers. *Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales*. PhD thesis, Université de Laval, Québec, 2004.

[Devillers 05] Robert Devillers Rodolphe et Jeansoulin. Qualité de l'information géographique. Hermès Lavoisier, 2005.

[Devogele 97] Thomas Devogele. *Processus d'intégration et d'appariement des bases de données géographiques. Application à une base de données routière multi-échelles.* PhD thesis, Université de Versailles, IGN, 1997.

[Devogele 98] Thomas Devogele, Christine Parent & Stefano Spaccapietra. *On spatial database integration*. International Journal of Geographical Information Science, vol. 12, pages 335–352, 1998.

[Devogèle 02] Thomas Devogèle. *A new Merging process for data integration based on the discrete Fréchet distance*. In D. Richardson et P. van Oosterom, editeur, 10th International Symposium on Spatial Data Handling (SDH), pages 167–181, Ottawa (Canada), 2002.

[Droesbecke 89] Jean-Jacques Droesbecke, Bernard Fichet & Philippe Tassi. Séries chronologiques : théorie et pratique des modèles arima. Economica, Paris, 1989.

[Droesbecke 94] Jean-Jacques Droesbecke, Bernard Fichet & Philippe Tassi. Modélisation arch. théorie statistique et application dans le domaine de la finance. Éditions de l'Université de Bruxelles, Bruxelles et Edition Ellipses, Paris, 1994.

[Droesbecke 06] Jean-Jacques Droesbecke, Michel Lejeune & Gilbert Saporta. Analyse statistique des données spatiales. Technip, 2006.

[Dubrule 83] Olivier Dubrule. *Two methods with differents objectives : splines and kriging*. Mathematical geology, vol. 15, pages 245–255, 1983.

[Dumolard 03] Pierre Dumolard, Nathalie Dubus & Laure Charleux. Les statistiques en géographie. Atouts géographie. Belin, 2003.

[Dupin de Saint-Cyr 08] Florence Dupin de Saint-Cyr, Robert Jeansoulin & Henri Prade. *Fusing Uncertain Structured Spatial Information*. In Sergio Greco & Thomas Lukasiewicz, editeurs, Scalable Uncertainty Management, volume 5291 of *Lecture Notes in Computer Science*, pages 174–188. Springer Berlin, Heidelberg, 2008.

[Egenhofer 89] Max J. Egenhofer. *A formal definition of Binary Topological Relationships.* In Third International Conference on Foundations of Data Organization and Algorithms (FODO)., Paris, France, 1989. Springer-Verlag.

[Egenhofer 91] Max J. Egenhofer & Robert D. Franzosa. *Point-set topological spatial relations*. International Journal of Geographical Information Science, vol. 5, no. 2, pages 161–174, 1991.

[Egenhofer 95] Max J. Egenhofer & Mark David M. *Naive Geography*. In Spatial Information Theory : A Theoretical Basis for GIS - International Conference, COSIT'95., volume 988 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin / Heidelberg, 1995.

[EPSG 85] EPSG. *EPSG Geodetic parameter Registry*, 1985.

[Ester 97] Martin Ester, Hans-Peter Kriegel & Jörg Sander. *Spatial Data Mining : A Database Approach*. In Lecture Notes in Computer Science, editeur, Proc. of the Fifth Int. Symposium on Large Spatial Databases (SSD '97), pages 47–66, Berlin, Germany, 1997. Springer.

[Ester 00] Martin Ester, Alexander Frommelt, Hans-Peter Kriegel & Jöorg Sander. *Spatial Data Mining : Database Primitives, Algorithms and Efficient DBMS Support*. Data Min. Knowl. Discov., vol. 4, no. 2-3, pages 193–216, 2000.

[Euzenat 93] Jérôme Euzenat. *Representation granulaire du temps.* Revue dintelligence artificielle, vol. 7, no. 3, pages 329–361, 1993.

[Filzmoser 04] Peter Filzmoser. *A multivariate outlier detection method.* In S. Aivazian, P. Filzmoser & Yu. Kharin, editeurs, Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, volume 1, pages 18–22, Minsk, 2004.

[Fisher 95] Pete F. Fisher & Mitch Langford. *Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation*. Environment and Planning A, vol. 27, pages 211–224, 1995.

[Flowerdew 89] Robin Flowerdew & Mick Green. *Statistical methods for inference between incompatible zonal systems*. In The Accuracy of Spatial Data Bases. Taylor & Francis, 1989.

[Fonseca 03] Frederico Fonseca, Clodoveu Davis & Gilberto Câmara. *Bridging Ontologies and Conceptual Schemas in Geographic Information Integration*. Geoinformatica, vol. 7, no. 4, pages 355–378, 2003.

[Forbus 84] Kenneth D. Forbus. *Qualitative process theory*. Artificial Intelligence,, vol. 24, pages 85–168, 1984.

[Fotheringham 91] Stewart A. Fotheringham & David W. Wong. *The modifiable areal unit problem in multivariate statistical analysis*. Environment and Planning A, vol. 23, no. 7, pages 1025 – 1044, 1991.

[Fotheringham 00] A. Stewart Fotheringham, Chris Brunsdon & Martin Charlton. Quantitative geography. Sage, London, UK, 2000.

[Fotheringham 02] A. Stewart Fotheringham, Chris Brunsdon & Martin Charlton. Geographically weighted regression : The analysis of spatially varying relationships. Wiley, Chichester, 2002.

[François 02] Jean-Christophe François. *Ressemblances et proximités : un point de vue sur le contexte théorique de la notion de discontinuité géographique*. Cybergeo : Revue européenne de Géographie, no. 214, 2002.

[Frank 92] Andrew U. Frank. *Qualitative spatial reasoning about distances and directions in geographic space.* Journal of Visual Languages & Computing, vol. 3, pages 343–371, 1992.

[Frank 01] Andrew U. Frank, Jean-Paul Cheylan & Jonathan Raper. Life and motion of socio-economic units. European Science Foundation GISDATA. Taylor and Francis, London, May 2001.

[Freska 91] Christian Freska. *Qualitative Spatial Reasoning*. In David M. Mark & Andrew U. Frank, editeurs, Cognitive and Linguistic Aspects of Geographic Space. Kluwer Academic Publishers, Dordrecht, 1991.

[Gallego 10] Francisco Javier Gallego. *A population density grid of the European Union*. Population and Environment, vol. 31, no. 6, pages 460–473, July 2010.

[Galton 04] Antony Galton. *Fields and Objects in Space, Time, and Space-time*. Spatial cognition and computation, vol. 4, no. 1, pages 39–68, 2004.

[Gauthier 02] Jason G. Gauthier. Measuring america : the decennial censuses from 1790 to 2000. U.S. Census Bureau, 2002.

[Gayte 97] Olivier Gayte, Thérèse Libourel, Jean-Paul Cheylan & Sylvie Lardon. Pollen, méthode de conception des systèmes d'information sur l'environnement. Hermès, Paris, 1997.

[Getis 92] Arthur Getis & Keith Ord. *The Analysis of Spatial Association by Use of Distance Statistics*. Geographical Analysis, vol. 24, pages 189–206, 1992.

[Getis 95] Arthur Getis & Keith Ord. *Local spatial autocorrelation statistics : distributional issues and an application*. Geographical Analysis, vol. 27, pages 189–206, 1995.

[Gómez 09] Leticia Gómez, Sophie Haesevoets, Bart Kuijpers & Alejandro A. Vaisman. *Spatial aggregation : Data model and implementation*. Information Systems, vol. 34, no. 6, pages 551–576, September 2009.

[Goodchild 80] Michael F Goodchild & Noel Lam. *Areal interpolation : a variant of the traditionnal spatial problem*. Geo-processing, vol. 1, pages 297–312, 1980.

[Goodchild 93] Michael F Goodchild, Luc Anselin & Uwe Diechman. *A general framework for the areal interpolation of socio-economic data*. Environment and Planning A, vol. 25, pages 383–397, 1993.

[Goodchild 00] Michael F. Goodchild. *Introduction : special issue on uncertainty in geographic information systems*. Fuzzy Sets Syst., vol. 113, pages 3–5, July 2000.

[Google 10] Google. *Google public data explorer - http ://www.google.com/publicdata/explore*, May 2010.

[Gotway-Crawford 05] Carol A. Gotway-Crawford & Linda Young. *Change of support : an interdisciplinary challenge*. Geostatistics for Environmental Applications, pages 1–13, 2005.

[Gotway 02] Carol Gotway & Linda Young. *Combining Incompatible Spatial Data*. Journal of the American Statistical Association, vol. 97, no. 458, pages 632–648, 2002.

[Goux 03] Dominique Goux. *Une histoire de l'Enquête Emploi*. Economie et Statistique, no. 362, 2003.

[Goyal 00] Roop K. Goyal & Max J. Egenhofer. *Consistent Queries over Cardinal Directions across Different Levels of Detail*. In DEXA Workshop, pages 876–880, 2000.

[Grasland 98] Claude Grasland. *Les maillages territoriaux : niveau d'observation ou niveaux d'organisation*. In Actes des entretiens J. Cartier, volume 76-77-78 of *Les découpages du territoire*, pages 115–132. INSEE-METHODES, 1998.

[Grasland 00] Claude Grasland, Hélène Mathian & Jean-Marc Vincent. *Multiscalar analysis and map generalisation of discrete social phenomena : Statistical problems and political consequences*. Statistical Journal of the United Nations Economic Commission for Europe, vol. 17, no. 2, pages 157–188, 2000.

[Grasland 02] Claude Grasland & Alina Potrykowska. *Mesures de la proximité spatiale : les migrations résidentielles à Varsovie*. Espace géographique, vol. 31, 2002.

[Grasland 05a] Claude Grasland, France Guerin-Pace & Christophe Terrier. *La diffusion spatiale, sociale et temporelle des pièces euros étrangères : un problème complexe*. In Actes des journées de Méthodologie Statistique, 2005.

[Grasland 05b] Claude Grasland, Hervé Martin, Jean-Marc Vincent, Jérôme Gensel, Hélène Mathian, Said Ouhalal, Oliver Cuenot, Euloge Edi & Lilianne Lizzi. *Le projet Hypercarte : analyse spatiale et cartographie interactive*. In SAGEO'2005, Avignon, France, 20-23rd June 2005.

[Grasland 06] Claude Grasland & Malika Madelin. Espon 3.4.3 project, chapitre The Modifiable Areas Unit Problem - Final report, page 254. The ESPON Monitoring Committee, Esch-sur-Alzette, Luxembourg, 2006.

[Grasland 09] Claude Grasland & Jérôme Gensel. *ESPON 2013 Database, 1st Interim Report*. http ://www.espon.eu/export/sites/default/Documents/Projects/ScientificPlatform/ESPONDatabase2013/fir_espondb_2013_27-02-09.pdf, 2009.

[Grasland 10a] Claude Grasland. *Spatial analysis of social facts*. In Handbook of Quantitative Geography. Bavaud F. & Mager C., 2010.

[Grasland 10b] Claude Grasland & Jérôme Gensel. *ESPON 2013 Database, 2nd Interim Report*, 2010.

[Grasland 10c] Claude Grasland & Jérôme Gensel. *ESPON 2013 Database, Final Report*, December 2010.

[Gregory 02] Ian Gregory. *Time-variant GIS Databases of Changing Historical Administrative Boundaries : A European Comparison*. Transactions in GIS, vol. 6, no. 2, pages 161–178, 2002.

[Grubbs 69] Franck E. Grubbs. *Procedures for detecting outlying observations in samples.* Technometrics, vol. 11, no. 1, pages 1–21, 1969.

[Grumbach 01] Stéphane Grumbach, Philippe Rigaux & Luc Segoufin. *Spatio-Temporal Data Handling with Constraints*. GeoInformatica, vol. 5, no. 1, pages 95–115, 2001.

[Grzegorczyk 51] Andrzej Grzegorczyk. *Undecidability of some topological theories.* Fundamenta Mathematicae, vol. 38, pages 137–152, 1951.

[Guo 09] Diansheng Guo & Jeremy Mennis. *Spatial data mining and geographic knowledge discovery-An introduction*. Computers, Environment and Urban Systems, pages 403–408, 2009.

[Haggett 73] Peter Haggett. L'analyse spatiale en géographie humaine. Armand Colin, Paris, 1973.

[Haining 03] Robert Haining. Spatial data analysis : Theory and practice. Cambridge University Press, 2003.

[Hall 09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. *The WEKA Data Mining Software : An Update*. SIGKDD Explorations, vol. 11, no. 1, 2009.

[Hangouët 05] Jean-François Hangouët. *Évaluation et documentation de la qualité*. In Robert Devillers Rodolphe et Jeansoulin, editeur, Qualité de l'information géographique, Information géographique et aménagement du territoire, pages 247–272. Hermès Lavoisier, 2005.

[Harris 10] Paul Harris & Martin Charlton. *SPATIAL ANALYSIS FOR QUALITY CONTROL, Phase 1 : The identification of logical input errors and statistical outliers*. Rapport technique, ESPON, 2010.

[Haslett 90] John Haslett, Graham Wills & Antony Unwin. *SPIDER - an interactive statistical tool for the analysis of spatially distributed data.* International Journal of Geographical Information Systems, 1990.

[Hengl 08] Tomislav Hengl, Emiel van Loon, Henk Sierdsema & Willem Bouten. *Advancing Spatio-temporal Analysis of Ecological Data : Examples in R.* In Osvaldo Gervasi, Beniamino Murgante, Antonio Laganà, David Taniar, Youngsong Mun & Marina L. Gavrilova, editeurs, ICCSA (1), volume 5072 of *Lecture Notes in Computer Science*, pages 692–707. Springer, 2008.

[Hernàndez 93] David Hernàndez. *Maintaining Qualitative Spatial Knowledge*. In Andrew U. Frank & Irene Campari, editeurs, Spatial Information Theory : A Theoretical Basis for GIS - European Conference, COSIT'93, numéro 716 in Lecture Notes in Computer Science, pages 36–53. Springer-Verlag, 1993.

[Hofmann 99] Thomas Hofmann. *Probabilistic latent semantic indexing*. In Hearst M., Gey F. & Tong R., editeurs, Proceedings of 22nd International Conference on Research and Development in Information Retrieval, pages 50–57, Berkeley, Californie, US, 1999.

[Hornsby 98] Kathleen Hornsby & Max J. Egenhofer. *Identity-Based Change Operations for Composite Objects*. In T. Poiker & N. Christman, editeurs, Eighth International Symposium on Spatial Data Handling, pages 202–213, Vancouver, Canada, July 1998.

[Howenstine 93] Erick Howenstine. *Measuring Demographic Change : The Split Tract Problem*. The Professional Geographer, vol. 45, no. 4, pages 425–430, 1993.

[IDEE 08] Institut Géographique National d'Espagne IDEE. *Infrastructure de Données Spatiales de l'Espagne*. http ://www.idee.es/show.do ?to=pideep_catalogoIDEE.FR, 2008.

[IGN, France 10] IGN, France. *IGNF-spatialRefSys (http ://lambert93.ign.fr/fileadmin/files/IGNF-spatial_ref_sys.sql)*, 2010.

[ISO 02a] International Organization for Standardisation ISO. *ISO19108 :2002 Geographic information – Temporal schema*. http ://www.iso.org/iso/catalogue_detail.htm ?- csnumber=26013, 2002.

[ISO 02b] International Organization for Standardisation ISO. *ISO19113 :2002 Quality principles*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?-csnumber=26018, 2002.

[ISO 03a] International Organization for Standardisation ISO. *ISO19112 :2003 Geographic information – Spatial referencing by geographic identifiers*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?- csnumber=41126, 2003.

[ISO 03b] International Organization for Standardisation ISO. *ISO19114 :2003 Quality evaluation procedures*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?-csnumber=26019, 2003.

[ISO 03c] International Organization for Standardisation ISO. *ISO19115 :2003 Geographic Information - Metadata*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?-csnumber=26020, 2003.

[ISO 04a] International Organization for Standardisation ISO. *ISO 8601 :2004 Éléments de données et formats d'échange – Échange d'information – Représentation de la date et de l'heure (http ://www.iso.org/iso/fr/catalogue_detail.htm ?csnumber=40874)*. http ://www.iso.org/iso/fr/catalogue_detail.htm ?csnumber=40874, 2004.

[ISO 04b] International Organization for Standardisation ISO. *Technologies de l'information – Registres de métadonnées (RM)*. http ://www.iso.org/iso/fr/catalogue_detail.htm ?- csnumber=31367, 2004.

[ISO 05] International Organization for Standardisation ISO. *ISO17369 :2005 Statistical data and metadata exchange*. http ://www.iso.org/iso/catalogue_detail.htm ?csnumber=40555, 2005.

[ISO 07a] International Organization for Standardisation ISO. *ISO19111 :2007 Geographic information – Spatial referencing by coordinates*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?-csnumber=41126, 2007.

[ISO 07b] International Organization for Standardisation ISO. *ISO19136 :2007, Technical committee 211, Geographic Information – Geography Markup Language (GML)*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm-?csnumber=32554, 2007.

[ISO 09] International Organization for Standardisation ISO. *Information et documentation – L'ensemble des éléments de métadonnées Dublin Core*. http ://www.iso.org/iso/fr/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm ?-csnumber=52142, 2009.

[ISO 10] International Organization for Standardisation ISO. *ISO19142 :2010, Technical Committee 211, Geographic information – Web Feature Service*. http ://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm ?- csnumber=42136, 2010.

[Iwasaki 97] Yumi Iwasaki. *Real-World Applications of Qualitative Reasoning*. IEEE Expert : Intelligent Systems and Their Applications, vol. 12, no. 3, May 1997.

[Jeansoulin 11] Robert Jeansoulin, Odile Papini, Henri Prade & Steven Schockaert. *Introduction : Uncertainty Issues in Spatial Information*. In Methods for Handling Imperfect Spatial Information, volume 256 of *Studies in Fuzziness and Soft Computing*. Springer, 2011.

[Jensen 91] Christian S. Jensen & Richard Thomas Snodgrass. *Temporel specialisation and generalization*. Rapport technique 91-25, Departement of mathematics and computer science, Aalborg University, 1991.

[Jensen 96] Christian S. Jensen & Richard T. Snodgrass. *Semantics of time-varying information*. Information Systems, vol. 21, no. 4, pages 311 – 352, 1996.

[Jensen 98] S. Jensen Christian, Curtis Dyreson, Michael Böhlen & al. *The Consensus Glossary of Temporal Database Concepts*. Temporal Databases : Research and Practice, vol. 1399, pages 367–405, 1998.

[Josselin 00] Didier Josselin, Jérôme Bolot & P. Chatonnay. *Optimisation de découpages territoriaux. Proposition de méthodes d'agrégation spatiale dirigée*. Revue internationale de Géomatique, vol. 10, no. 3-4, pages 383–409, 2000.

[Kauppinen 07] Tomi Kauppinen & Eero Hyvönen. Ontologies : A handbook of principles, concepts and applications in information systems, chapitre Modeling and Reasoning about Changes in Ontology Time Series. Springer-Verlag, New York, NY, USA, rajiv kishore and ram ramesh and raj sharman edition, 2007.

[Kazakos 03] W. Kazakos, A. Akhounov & H. Paoli. *Editing ISO 19115 Compliant Metadata in EUROSION*. In Albrecht Gnauck & Ralph Heinrichs, editeurs, 17th International Conference Informatics for Environmental Protection (EnviroInfo) : The Information Society and Enlargement of the European Union, pages 468–474, 2003.

[Kent 97] Jean-Pierre Kent & Maarten Schuerhoff. *Some Thoughts About a Metadata Management System*. In SSDBM '97 : Proceedings of the Ninth International Conference on Scientific and Statistical Database Management, pages 174–185, Washington, DC, USA, 1997. IEEE Computer Society.

[Kieffer 02] Annick Kieffer, Marco Oberti & Edmond Preteceille. *Enjeux et usages des catégories socio-professionnelles en Europe*. Sociétés contemporaines, vol. 45-46, pages 157–185, 2002.

[Klein 09] Etienne Klein. *Quelle est la forme du temps ? Linéaire ou cyclique ?* http ://www.confinde.com/ ?p=47, 2009.

[Kokolakis 01] George Kokolakis, George Kouvaras & Georgia Panagopoulou. *The Role of Metadata in the Intelligent Use of the Data : The User Approach.* In Proceedings of the New Techniques in Information Technology and Statistics., pages 257–262, Hersonissos Crete, Greece, 2001.

[Kristiansson 00] Göran Kristiansson. *Sweden : National Archival Database (NAD)*, 2000.

[Langford 92] Mitchel Langford & D.J. Unwin. *Generating and mapping population density surface within a GIS*. The Cartographic Journal, vol. 31, pages 21–26, 1992.

[Langran 88] Gail E. Langran. *Temporal design tradeoffs*. In GIS/LIS'88 Proceedings : Accessing the World, Volume II., pages 880–899, Falls Church, Virginia, 1988.

[Langran 92] Gail E. Langran. Time in geographic information systems. Taylor and Francis, Seattle, WA, USA, 1992.

[Langran 98] Gail E. Langran & Nicholas R. Chrisman. *A Framework for Temporal Geographic Information*. Cartographica : The International Journal for Geographic Information and Geovisualization, vol. 25, no. 3, pages 1–14, 1998.

[Lanter 91] David P. Lanter & Rupert Essinger. *User-Centered Graphical User Interface Design for GIS*. Rapport technique 91-6, National Center for Geographic Information and Analysis, 1991.

[Lardon 99] Sylvie Lardon, Thérèse Libourel & Jean-Paul Cheylan. *Concevoir la dynamique des entités spatio-temporelles*. Revue internationale de géomatique, vol. 9, no. 1, pages 67–99, 1999.

[Laura Díaz 07] Cristian Martín Laura Díaz, Michael Gould, Carlos Granell & Miguel Ángel Manso. *Semi-Automatic Metadata Extraction from Imagery and Cartographic Data*. In IGARRS, 2007.

[Le Bras 93] Hervé Le Bras. La planète au village - migrations et peuplement en france. Editions de l'Aube, 1993.

[Le Gléau 99] Jean-Pierre Le Gléau. Un zonage pourquoi faire ? projet d'article pour les annales des Ponts-et-Chaussées, Novembre 1999.

[Lebart 69] Ludovic Lebart. Analyse statistique de la contiguïté, chapitre XVIII, pages 81–112. Publications de l'Institut de Statistique des Universités de Paris, 1969.

[Lee 05] Sang-Il Lee. *Between the quantitative and GIS revolutions : towards an SDA-centered GIScience*. Journal of Geography Education, 2005.

[Levenshtein 65] Vladimir I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals*. Doklady Akademii Nauk SSSR, vol. 4, no. 163, pages 845–848, 1965.

[Ligozat 10] Gérard Ligozat. Raisonnement qualitatif sur le temps et l'espace. Ingénierie des langues. Lavoisier, hermes edition, 2010.

[Locoh 95] Thérèse Locoh & Elisabeth Omoluabi. *Où sont donc passés les 30 millions de nigérians manquants ?* In J. Vallin, editeur, Clins d'œil de démographes à l'Afrique, documents et manuels du CEPED n° 2, pages 57–75. CEPED, 1995.

[Longley 05] P.A. Longley, M. F. Goodchild, D. Maguire & D. Rhind. Geographic information systems and science. Wiley, 2005.

[Lwin 09] KoKo Lwin & Yuji Murayama. *A GIS Approach to Estimation of Building Population for Micro-spatial Analysis*. T. GIS, vol. 13, no. 4, pages 401–414, 2009.

[Mahalanobis 36] Prasanta Chandra Mahalanobis. *On the generalised distance in statistics*. Proceedings of the National Institute of Sciences of India, vol. 2, no. 1, pages 49–55, 1936.

[Manso 04] Miguel .A Manso, Javier Nogueras, Javier Zarazaga & Barnabé Miguel A. *Automatic Metadata Extraction from Geographic Information*. In 7th AGILE Conference on Geographic Information Science, pages 379–385, 2004.

[Marceau 99] Danielle J. Marceau. *The scale issue in social and natural sciences*. Canadian Journal of Remote Sens, vol. 25, no. 4, pages 347–356, 1999.

[Markoff 73] John Markoff & Gilbert Shapiro. *The Linkage of Data Describing Overlapping Geographical Units*. Historical Methods Newsletter, vol. 7, pages 34–46, 1973.

[Martin 03] David Martin. *Extending the automated zoning procedure to reconcile incompatible zoning systems.* International Journal of Geographical Information Science, vol. 17, no. 2, pages 181–196, 2003.

[Matheron 63] Georges Matheron. *Principles of geostatistics*. Economy geology, vol. 58, pages 1246–1268, 1963.

[Mathian 01] Hélène Mathian & Marie Piron. *Echelles géographiques et méthodes statistiques multidimensionnelles*. In Léna Sanders, editeur, Modèles en analyse spatiale. Hermès, 2001.

[McCarthy 82] John L. McCarthy. *Metadata Management for Large Statistical Databases*. In Eigth International Conference on Very Large Data Bases, pages 234–243, Mexico city, Mexico, September 8-10 1982. Morgan Kaufmann.

[McKenzie 91] Edwin McKenzie & Richard Thomas Snodgrass. *An evaluation of relational algebras incorporating the time dimension in databases*. ACM computing surveys, vol. 23, no. 4, pages 501–543, 1991.

[Meyer 04] David Meyer, Torsten Hothorn, Friedrich Leisch & Kurt Hornik. *StatDataML : An XML format for statistical data*. Journal of Computational Statistics, vol. 19, no. 3, pages 493–509, August 2004.

[Miller 00] Harvey J. Miller. *Geographic representation in spatial analysis*. Journal of Geographical Systems, vol. 2, no. 1, pages 55–60, 2000.

[Miquel 03] Maryvonne Miquel & Anne Tchounikine. *Extension du modèle M3 aux évolutions temporelles dans les applications SOLAP*. Revue des Nouvelles Technologies de l'Information (RNTI), 2003.

[Miron 09] Alina Dia Miron. *Découverte d'associations sémantiques pour le Web Sémantique Géospatial : le framework ONTOAST*. PhD thesis, Université Joseph Fourier, Grenoble, 2009.

[Moles 95] Abraham Moles & Elisabeth Moles. Les sciences de l'imprécis. Points Sciences. Seuil, 1995.

[Monmonier 89] Mark Monmonier. *Geographic brushing : enhancing exploratory analysis of the scatterplot matrix.* Geographical Analysis, vol. 21, pages 81–84, 1989.

[Morgan 03] Kevin Morgan. *How Objective 1 arrived in Wales : the political origins of a coup.* Contemporary Wales, vol. 15, no. 1, pages 20–29, 2003.

[Motte 03] Claude Motte, Isabelle Séguy & Christine Thérier. Communes d'hier, communes d'aujourd'hui. les communes de la france métropolitaine, 1801-2001. dictionnaire d'histoire administrative. Institut National d'Études Démographiques, 2003.

[Mugglin 00] Andrew S. Mugglin, Bradley P. Carlin & Alan E. Gelfand. *Fully model based approaches for spatially misaligned data*. Journal of the American Statistical Association, vol. 95, pages 877–887, 2000.

[Mustière 01] Sébastien Mustière. *Apprentissage supervisé pour la généralisation cartographique*. PhD thesis, Thèse de l'Université Paris 6, 2001.

[Nass 92] Clifford Nass & David Garfinkle. *Localized autocorrelation diagnostic statistic (LADS) for spatial models : Conceptualization, utilization and computation.* Regional science and urban economics, vol. 22, pages 333–346, 1992.

[Naumann 04] Felix Naumann, Johann Christoph Freytag & Ulf Leser. *Completeness of integrated information sources*. Inf. Syst., vol. 29, no. 7, pages 583–615, 2004.

[Nordhaus 02] William D. Nordhaus. *Alternative approaches to spatial rescaling*. Rapport technique, Yale University, New Haven, CT, 2002.

[Nordhaus 05] William D. Nordhaus, David Azam Qazi Corderi, Nadejda Makarova Victor, Mohammed Mukhtar & Alexander Miltner. *The G-Econ database on gridded output : Methods and data*. Rapport technique, Yale University, 2005.

[Norman 03] Paul Norman, Philip Rees & Paul Boyle. *Achieving data compatibility over space and time : creating consistent geographical zones*. International Journal of Population Geography, vol. 9, pages 365–386, 2003.

[Nyquist 28] Harry Nyquist. *Certain Topics in Telegraph Transmission Theory*. In Proceedings of the Institute of Electrical and Electronics Engineers, volume 90, pages 280–305, 1928.

[Oakley 05] Graeme Oakley, Alistair Hamilton & Jeremy Michel. *Experiences and Plans of the Australian Bureau of Statistics related to Data and Metadata Exchange*. Rapport technique, Australian Bureau of Statistics, 2005.

[OCDE 08] Organisation pour la Coopération et le Développement Economique OCDE. *HANDBOOK ON CONSTRUCTING COMPOSITE INDICATORS : METHODOLOGY AND USER GUIDE*. Rapport technique ISBN 978-92-64-04345-9, OCDE, 2008.

[OGC 99] Open Géospatial Consortium OGC. *Simple Feature Access - Part 2 : SQL Option*. http ://www.opengeospatial.org/standards/sfs, 1999.

[Oliveau 04] Sébastien Oliveau. *Modernisation villageoise et distance à la ville en Inde du Sud.* PhD thesis, Université Paris 1 Panthéon-Sorbonne, thèse de doctorant en géographie, 2004.

[Olteanu-Raimond 09] Ana-Maria Olteanu-Raimond, Sébastien Mustière & Anne Ruas. *Fusion des connaissances pour apparier des données géographiques*. Revue Internationale de Géomatique, vol. 19, no. 3, pages 321–349, 2009.

[Openshaw 77] Stan Openshaw. *A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling*. International Journal of Geographical Information Science, 1977.

[Openshaw 79] Stan Openshaw & P.J. Taylor. *A million or so correlation coefficients : Three experiments on the modifiable areal unit problem.* In N. Wrigley, editeur, Statistical methods in the spatial sciences, pages 127–144. Pion, London, 1979.

[Openshaw 81] Stan Openshaw. *Le problème de l'agrégation spatiale en géographie*. L'Espace Géographique, vol. 1, pages 15–24, 1981.

[Openshaw 87] Stan Openshaw, Martin Charlton, Colin Wymer & Craft Alan W. *A mark I geographical analysis machine for the automated analysis of point data sets*. International Journal of Geographical Information Science, vol. 1, pages 335–358, 1987.

[Openshaw 88] Stan Openshaw. *Building an automated modelling system to explore a universe of spatial interaction models*. Geographical Analysis, no. 20, pages 31–46, 1988.

[Openshaw 94] Stan Openshaw. *Two exploratory space-time-attribute pattern analysers relevant to GIS*. In A. Stewart Fotheringham & P Rogerson, editeurs, Spatial Analysis and GIS, pages 83–105. Taylor & Francis, London, UK, 1994.

[Openshaw 96] Stan Openshaw. *Developing GIS-relevant zone-based spatial analysis methods*. In P.A. Longley & Michael Batty, editeurs, Spatial analysis : Modelling in a GIS environment, pages 55–73. GeoInformation International, Cambridge, UK, 1996.

[Oregon 09] Oregon. *Department of Revenue - Boundary Change Information (www.state.or.us/DOR/PTD/docs/local-b/504-405.pdf)*, 2009.

[Ott 01] Thomas Ott & Franck Swiaczny. Time-integrative geographic information systems. management and analysis of spatio-temporal data. Springer, 2001.

[Pan 04] Feng Pan & Jerry R. Hobbs. *Time in OWL-S*. In Proceedings of the AAAI Spring Symposium on Semantic Web Services, pages 29–36. Stanford University, 2004.

[Paramo 05] Fidel Paramo & Oscar Gomez. *Environmental Accounting. Methodological guidebook. Data processing of land cover flows.* Internal report, European Topic Centre on Terrestrial Environment, Barcelona, Spain, 2005.

[Parent 06] Christine Parent, Stefano Spaccapietra & Esteban Zimányi. Conceptual modeling for traditional and spatio-temporal applications : The mads approach. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[Parlement européen 03] Parlement européen. *Établissement d'une nomenclature commune des unités territoriales statistiques (NUTS)*. Journal officiel, vol. L 154, no. 0041, 2003.

[Parlement européen 07] Parlement européen. *directive 2007/2/CE établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE) (http ://inspire.jrc.ec.europa.eu/)*, 2007.

[Parlement européen 10] Parlement européen. *implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services*. Journal officiel européen, Décembre 2010.

[Pattuelli 03] Maria Cristina Pattuelli, Stephanie W. Haas, Stephanie W. Haas & Jesse Wilbur. *The GovStat Ontology*. In Proceedings of the 2003 annual national conference on Digital Government research (DG.O), 2003.

[Pebesma 10] E. J. Pebesma, Dan Cornford, Gregoire Dubois, Gerard B.M. Heuvelink, Dionisis Hristopoulos, Jürgen Pilz, Ulrich Stöhlker, Gary Morin & O. Skoien Jon. *INTAMAP : the design and implementation of an interoperable automated interpolation web service*. Computers & Geosciences, vol. 2, pages 12–30, 2010.

[Pedersen 01] Torben Bach Pedersen, Christian S. Jensen & Curtis E. Dyreson. *A Foundation for Capturing and Querying Complex Multidimensional Data*. Information Systems, vol. 26, pages 383–423, 2001.

[Peuquet 94] Donna Peuquet. *It's About Time : A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems.* Annals of the Association of American Geographers, vol. 83, no. 3, pages 441–461, 1994.

[Peuquet 95] Donna Peuquet & Niu Duan. *An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data*. International Journal of Geographical Information Science, vol. 9, no. 1, pages 7–24, 1995.

[Peuquet 02] Donna Peuquet. Representations of time and space. Guildford Press, New York, NY, USA, 2002.

[Piron 93] Marie Piron. *Changer d'échelle : une méthode pour l'analyse des système multi-échelles*. L'espace géographique, no. 2, 1993.

[Pison 11] Gilles Pison, Hélène Mathian, Christine Plumejeaud & Jérôme Gensel. *Exploring world demography on line*. In International Cartographic Conference, Paris, July 2011.

[Plumejeaud 09a] Christine Plumejeaud, Jerome Gensel, Marlène Villanova-Oliver, Maher Ben Rebah & Guillaume Vergnaud. *Modélisation de hiérarchies territoriales multiples - Vers la gestion d'informations spatio-temporelles évolutives*. In Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2009), Paris, France, 2009.

[Plumejeaud 09b] Christine Plumejeaud, Julie Prud'homme, Paule-Annick Davoine & Jérôme Gensel. *Etude de méthodes de transfert d'indicateurs associés à différents découpages du territoire - Application à la ville de Grenoble*. In SAGEO 2009, Paris, France, 2009.

[Plumejeaud 10] Christine Plumejeaud, Ronan Ysebaert & Maria-José Ramos. *Acquisition and Storage of data and metadata in ESPON 2013 DB*. Technical report tr1, ESPON, 2010.

[Plumejeaud 11] Christine Plumejeaud & Jérôme Gensel. *Complexité liée à la variabilité sémantique des statistiques socio-économiques*. In Extraction et Gestion des Connaissances 2011 - Atelier sur la Fouille de Données Complexes, 2011.

[Proisy 11] Alexandre Proisy & Juan Carlos Rodado. *Fonds structurels européens : quel bilan à mi-parcours pour les NEM ?* Flash économie - recherche économique, no. 113, 2011.

[Pumain 97] Denise Pumain & Thérèse Saint-Julien. L'analyse spatiale - 1. localisations dans l'espace. collection Cursus. Armand Colin, Paris, 1997.

[Pumain 02] Denise Pumain & Marie-Claire Robic. *Le rôle des mathématiques dans une "révolution" théorique et quantitative : la géographie française depuis les années 1970*. Revue d'Histoire des Sciences Humaines, vol. 1, no. 6, pages 123–144, 2002.

[Pumain 04] Denise Pumain. *Maillage - Hypergéo*, 2004.

[Pumain 10] Denise Pumain. *Les enjeux des zonages*. http ://www.sfds.asso.fr/ressource.php ?fct=ddoc&i=679, Avril 2010.

[Rafanelli 90] Maurizio Rafanelli & Arie Shoshani. *Storm : A Statistical Object Representation Model*. In Zbigniew Michalewicz, editeur, In Proc. of Statistical and Scientific Database Management, 5th International Conference SSDBM, pages 14–29, Charlotte, NC, USA, April 3-5 1990.

[Raffestin 80] Claude Raffestin. Pour une géographie du pouvoir. Litec, Paris, 1980.

[Raimond 07] Ana-Maria Raimond. *Appariement de données géographiques utilisant la théorie des croyances*. Le Monde des cartes, pages 38–45, 2007.

[Randell 92] David A. Randell, Zhan Cui & Anthony G. Cohn. *A spatial logic based on regions and Connection*. In Morgan Kaufmann, editeur, 3rd Int. Conf. on Knowledge Representation and Reasoning, 1992.

[Raper 95] Jonathan Raper & David Livingstone. *Development of a Geomorphological Spatial Model Using Object-Oriented Design*. International Journal of Geographical Information Systems, vol. 9, no. 4, pages 359–383, 1995.

[Rase 01] Wolf-Dieter Rase. *Volume-preserving interpolation of a smooth surface from polygon-related data*. Journal of Geographical Systems, vol. 3, no. 2, pages 199–213, 2001.

[Raynal 96] Laurent Raynal, Pierre Dumolard, Gérard d'Aubigny, Christiane Weber, Philippe Rigaux, M. Scoll & D. Larcena. *Gérer et générer des données spatiales hiérarchisées*. Revue internationale de Géomatique, vol. 6, no. 4, pages 365–382, 1996.

[Reibel 07] Michel Reibel & Aditya Agrawal. *Areal Interpolation of Population Counts Using Pre-classified Land Cover Data*. Population Research and Policy Review, vol. 26, no. 5-6, 2007.

[Renolen 96] Agnar Renolen. *History graphs : Conceptual modelling of spatiotemporal data.* In International Cartographic Association, editeur, GIS Frontiers in Business and Science, Brno, Czech Republic, 1996.

[Rensik 02] Ronald A. Rensik. *Change detection.* Annual Review of Psychology, vol. 53, pages 245–277, 2002.

[Renz 98] Jochen Renz & Bernard Nebel. *Spatial Reasoning with Topological Information*. In Spatial Cognition, volume 1404 of *Lectures Notes in Computer Sciences*, pages 351–371. Springer, 1998.

[Rigaux 95] Philippe Rigaux & Michel Scholl. *Multi-Scale Partitions : Application to Spatial and Statistical Databases*. In SSD '95 : Proceedings of the 4th International Symposium on Advances in Spatial Databases, numéro 3-540-60159-7, pages 170–183, London, UK, 1995. Springer-Verlag.

[Roberts 10] Jason J. Roberts, Benjamin D. Best, Daniel C. Dunn, Eric A. Treml & Patrick N. Halpin. *Marine Geospatial Ecology Tools : An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++*. Environmental Modelling & Software, vol. 25, pages 1197–1207, 2010.

[Robertson 04] Stephen Robertson. *Understanding inverse document frequency : On theoretical arguments for IDF*. Journal of Documentation, vol. 60, no. 503-520, page 2004, 2004.

[Rolland-May 84] Christiane Rolland-May. *Notes sur les espaces géographiques flous*. Bulletin de l'Association de Géographes Français, vol. 502, 1984.

[Rossini 07] Anthony Rossini, Luke Tierney & Na Li. *Simple parallel statistical computing in r.* Journal of Computational & Graphical Statistics, vol. 16, no. 2, June 2007.

[Rousseeuw 96] Peter Rousseeuw & Annick Leroy. Robust regression and outlier detection. John Wiley & Sons, 1996.

[Rousseeuw 99] Peter Rousseeuw, Ida Ruts & John W. Tukey. *The bagplot : a bivariate Boxplot*. The American Statistician, vol. 53, no. 4, pages 382–387, 1999.

[Ruas 99] Anne Ruas. *Modèle de généralisation de données géographiques à base de contraintes et d'autonomie*. PhD thesis, Université de Marne la Vallée, 1999.

[Ruas 04] Anne Ruas. *Le changement de niveau de détail dans la représentation de l'information géographique.* PhD thesis, Université de Marne-La-Vallée, 2004.

[Sanders 99] Léna Sanders, Denis Gautier & Hélène Mathian. *Les concepts de système spatial et de dynamique, un essai de formalisation*. Revue internationale de Géomatique, vol. 9, no. 1, pages 25–44, 1999.

[Saporta 06] Gilbert Saporta. Probabilités, analyse des données et statistique. TECHNIP, Paris, France, 2006.

[Schmit 06] Christian Schmit, Mark D.A. Rounsevell & Isidore La Jeunesse. *The limitations of spatial land use data in environmental analysis*. Environmental Science & Policy, vol. 9, no. 2, pages 174 – 188, 2006. Assessing Climate Change Effects on Land Use and Ecosystems in Europe.

[Selivanov 09] Victor L. Selivanov. *Undecidability in Some Structures Related to Computation Theory*. J. Log. Comput., vol. 19, no. 1, pages 177–197, 2009.

[Servigne 05] Sylvie Servigne, Nicolas Lesage & Thérèse Libourel. *Composantes qualité et métadonnées*. In Robert Devillers Rodolphe et Jeansoulin, editeur, Qualité de l'information géographique, Information géographique et aménagement du territoire, pages 213–246. Hermès Lavoisier, 2005.

[Shafer 76] Glenn Shafer. A mathematical theory of evidence. Princeton University Press, 1976.

[Sheeren 04] David Sheeren, Sébastien Mustière & Jean-Daniel Zucker. *How to Integrate Heterogeneous Spatial Databases in a Consistent Way ?* In 8th conference on Advanced Databases and Information Systems (ADBIS'04), Budapest, Hungary, 2004.

[Shneiderman 96] Ben Shneiderman. *The Eyes Have It : A Task by Data Type Taxonomy for Information Visualizations*. In Proceedings of the 1996 IEEE Symposium on Visual Languages, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[Shoshani 82] Arie Shoshani. *Statistical Databases : Characteristics, Problems, and some Solutions*. In VLDB '82 : Proceedings of the 8th International Conference on Very Large Data Bases, pages 208–222, San Francisco, CA, USA, 1982. Morgan Kaufmann Publishers Inc.

[Smith 94] Barry Smith. *Fiat Objects*. Topoi, vol. 20, no. 2, pages 131–148, 1994.

[Smith 96] Brian C. Smith. Hacking the world : On the origin of objects. MIT Press, Cambridge, MA, 1996.

[Snodgrass 92] Richard Thomas Snodgrass. *Temporal databases*. In Andrew U. Frank, I. Campari & U. Formentini, editeurs, Proceedings of Theories and Methods of Spatio-temporal reasonning, pages 22–64. Springer-Verlag, New-York, 1992.

[Spaccapietra 02] Stefano Spaccapietra, Christine Parent & Yves Dupont. *Model Independent Assertions for Integration of Heterogeneous Schemas*. Very Large DataBases Journal, vol. 1, no. 1, pages 81–126, 2002.

[Spaccapietra 04] Stefano Spaccapietra, Nadine Cullot, Christine Parent & Christelle Vangenot. *On spatial ontologies*. In 6th Brazilian Symposium On Geoinformatics, Campos do Jordao, Brazil, 22-24 November 2004.

[Spery 01] Laurent Spery, Christophe Claramunt & Thérèse Libourel. *A Spatio-Temporal Model for the Manipulation of Lineage Metadata*. GeoInformatica, vol. 5, no. 1, pages 51–70, 2001.

[Takahara 80] Yasuhiko Takahara, D. Macko & Mihajlo D. Mesarovic. Théorie des systèmes hiérarchiques à niveaux multiples. Economica, 1980.

[Tarski 59] Alfred Tarski. *What is elementary geometry ?* In L. Henkin, P. Suppes & A. Tarsky, editeurs, The axiomatic method with Special Reference to Geometry and Physics, pages 16–29. North-Holland Publishing Company, Amsterdam, 1959.

[Taussi 07] Matti Taussi. *Automatic production of metadata out of geographic datasets*. Master's thesis, Helsinki University of Technology, Espoo, Finlande, 2007.

[Tchounikine 05] Anne Tchounikine, Maryvonne Miquel, Robert Laurini, Taher Ahmed, Sandro Bimonte & Virginie Baillot. Panorama de travaux autour de l'intégration de données spatio-temporelles dans les hypercubes, pages 21–33. Revue des Nouvelles Technologies de l'Information, cépaduès edition, 2005.

[Templ 09] Matthias Templ, Peter Filzmoser & K. Hron. *Robust Imputation of Missing Values in Compositional Data Using the R-package robCompositions*. In Proceedings of the NTTS Conference, 2009. talk : NTTS Conference, Brüssel ; 2009-02-19.

[Terrier 80] Christophe Terrier. *Mirabelle*. Courrier des Statistiques, no. 73, 1980.

[Terrier 98] Christophe Terrier. *Zonage de Pouvoir, Zonage de Savoir*. In Jean-Marc Benoit, Philippe Benoit & Daniel Pucci, editeurs, La France redécoupée, Enquête sur la quadrature de l'hexagone. Belin, 1998.

[Terrier 00] Christophe Terrier. *Les Zonages et l'Europe*. Annales des Ponts et Chaussées, no. 93, pages 68–72, 2000.

[Terrier 05] Christophe Terrier. *Les découpages territoriaux : problèmes épistémologiques et méthodologiques*. In Violette Rey & Thérèse Saint-Julien, editeurs, Territoires d'Europe, la différence en partage. ENS Editions, Lyon, 2005.

[Thériault 99] Marius Thériault & Christophe Claramunt. *La représentation du temps et des processus dans les SIG*. Revue internationale de géomatique, vol. 9, no. 1, pages 67–99, 1999.

[Thomas 05] James J. Thomas & Kristin A. Cook, editeurs. Illuminating the path : The research and development agenda for visual analytics. National Vizualisation and Analytics Center, 2005.

[Tobler 70] Waldo R. Tobler. *A Computer Movie Simulating Urban Growth in the Detroit Region*. Economic Geography, vol. 46, no. 2, pages 234–240, 1970.

[Tobler 79] Waldo R. Tobler. *Smooth pycnopylactic interpolation for geographical regions*. Journal of the American Statistical Association, vol. 74, pages 519–530, 1979.

[Tukey 77] John M. Tukey. Exploratory data analysis. Addison Wesley Longman Publishing Co., Inc., 1977.

[Ubeda 97] Thierry Ubeda. *Contrôle de la qualité spatiale des bases de données géographiques. Cohérence topologique et corrections d'erreurs*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 1997.

[UMS 2414 RIATE 08] UMS 2414 RIATE, UMR 8504 Géographie-cités, LIG, IGEAT, Umeå University, Università l'Orientale & Université "Alexandru Ioan Cuza" Iaşi. *Régions en déclin : un nouveau paradigme démographique et territorial*. Etude pour le Parlement Européen - Structural and cohesion policies - Juillet 2008 IP/B/REGI/IC/2007-044, Parlement européen, Juillet 2008.

[UN/ECE 95] UN/ECE. *Guidelines for the Modelling of Statistical Data and Metadata*. Rapport technique, UN/ECE, New York, Geneva, 1995.

[UN/ECE 00] UN/ECE. *Terminology on statistical metadata*. In Conference of European Statisticians Statistical Standards and Studies, numéro 53, Geneva, 2000.

[Unwin 94] Antony Unwin. *REGARDing Geographic Data*. In P. Dirschedl & R. Ostermann, editeurs, Computational Statistics, pages 315–326. Physica-Verlag, 1994.

[Vaisman 07] Alejandro A. Vaisman. *Data Quality-Based Requirements Elicitation for Decision Support Systems.* In Christian Koncilia & Robert Wrembel, editeurs, Data Warehouses and OLAP : Concepts, Architectures and Solutions, chapitre 3, pages 58–86. IRM press, 2007.

[Vangenot 98] Christelle Vangenot. *Représentation multi résolution, Concepts pour la description de bases de données avec multi-représentation.* Revue internationale de géomatique, vol. 8, no. 1-2, pages 121–147, 1998.

[Vauglin 98] François Vauglin & Atef Bel Adj Ali. *Geometric matching of polygonal surfaces in GISs*. In RTI Annual conference, pages 1511–1516, Tampa, FI(USA), 1998.

[Villeneuve 72] Paul Y. Villeneuve. *Un paradigme pour l'étude de l'organisation spatiale des sociétés*. Cahiers de géographie du Québec, vol. 16, no. 38, pages 199–211, 1972.

[Voisard 01] Agnès Voisard, Michel Scholl & Philippe Rigaux. Spatial databases with application to gis. Morgan Kaufmann, 2001.

[Wachowicz 99] Monica Wachowicz. Object-oriented design for temporal gis. Taylor & Francis, Bristol, PA, USA, 1999.

[Wadsworth 06] Richard A. Wadsworth, Alexis Comber & Pete F. Fisher. *Expert Knowledge and Embedded Knowledge or why long rambling class descriptions are useful.* In Gregory Elmes Andreas Riedl Wolfgang Kainz, editeur, Progress in Spatial Data Handling, Proceedings of SDH, pages 197 – 213. Springer Berlin / Heidelberg, 2006.

[Walter 99] Volker Walter & Dieter Fritsch. *Matching spatial data sets : a statistical approach*. International Journal of Geographical Information Science, vol. 13, no. 5, pages 445–473, 1999.

[Wand 96] Yair Wand & Richard Y. Wang. *Anchoring Data Quality Dimensions in Ontological Foundations*. In Communications of the ACM, pages 86–95, 1996.

[Waniez 10] Philippe Waniez. Cartographie thématique et analyse des données avec philcarto 5.xx pour windows. http ://philcarto.free.fr/Logiciels/DOCdeGRANITn1.zip, 2010.

[Wilde 03] Marion Wilde, Manfred Lange, Hardy Pundt, Nicole Ostländer & Krzysztof Janowicz. *An environmental metadata profile for the EU project MEDIS*. In Albrecht Gnauck & Ralph Heinrichs, editeurs, 17th International Conference Informatics for Environmental Protection, pages 482–489, 2003.

[Wilde 04] Marion Wilde & Hardy Pundt. *Development of an ISO compliant, internet-based metadata editor for the EU project MEDIS*. In AGIT, Symposium und Fachmesse für Angewandte Geoinformatik, Salzburg, Austria, 2004.

[Wilks 39] Samuel S. Wilks. *The rise of modern statistical science*. In MIT Industrial Statistics Conference, pages 283–310, New York, NY, USA, 1939. Pitman Publ. Corp.

[Woodruff 97] Allison Woodruff & Michael Stonebraker. *Supporting Fine-grained Data Lineage in a Database Visualization Environment*. In Proceedings of the Thirteenth International Conference on Data Engineering, numéro Report n °UCB/CSD-97-932, pages 91–102, Birmingham, U.K., April 1997.

[Worboys 92] Michael F. Worboys. *A model for spatio-temporal information*. In 5th International Symposium on Spatial Data Handling, volume 2, pages 602–611, 1992.

[Worboys 98] Michael F. Worboys. *A generic model for spatio-bitemporal geographic information*. In M.J Egenhofer & R.G. Golledge, editeurs, Spatial and Temporal Reasoning in Geographic Information Systems., pages 25–39. Oxford University Press, 1998.

[Worboys 05] Michael F. Worboys. *Event-oriented approaches to geographic phenomena*. International Journal of Geographical Information Science, vol. 19, pages 1–28, 2005.

[Yuan 96] May Yuan. *Temporal GIS and Spatiotemporal Modeling.* In M F Goodchild, editeur, Integrating GIS and Environmental Modeling, 1996.

[Yuan 99] May Yuan. *Use of a three-domain representation to enhance GIS support for complex spatio-temporal queries.* Transactions in GIS, vol. 3, pages 137–159, 1999.

[Zaninetti 05] Jean Marc Zaninetti. Statistique spatiale : méthodes et applications géomatiques. Lavoisier, 2005.

[Zarazaga-Soria 03] F. Javier Zarazaga-Soria, Javier Lacasta, Javier Nogueras-Iso, M. Pilar Torres & P.R. Muro-Medrano. *A Java Tool for Creating ISO/FGDC Geographic Metadata*. In L. Bernard & A. Sliwinski und K. Senkler, editeurs, Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung. Beiträge zu den Münsteraner GI-Tagen, volume 18, pages 17–30. IfGIprints, 2003.

[Zeitouni 00] Karine Zeitouni. *A Survey on Spatial Data Mining Methods Databases and Statistics*. In Point of Views, Information Resources Management Association International Conference (IRMA'2000), Data Warehousing and Mining Track, 2000.

[Zeitouni 01] Karine Zeitouni, Laurent Yeh & Marie-Aude Aufaure. *Join Indices as a Tool for Spatial Data Mining*. In LNCS, editeur, Temporal, Spatial, and Spatio-Temporal Data Mining, Lecture Notes in Computer Science, pages 105–116. Springer, Berlin / Heidelberg, 2001.

[Zhan 98] F.B. Zhan. *Approximate analysis of binary topological relations between geographic regions with indeterminate boundaries.* Soft Computing, vol. 2, pages 28–34, 1998.

Publications

Revues internationales

Christine Plumejeaud, Hélène Mathian, Jérôme Gensel, Claude Grasland, *Spatio-temporal analysis of territorial changes from a multi-scale perspective*, International Journal of Geographical Information Systems. Vol 1-16 Aout 2011

Chapitres de livres

Christine Plumejeaud, Bogdan Moisuc, Sandro Bimonte, Marlène Villanova-Olver, Jérôme Gensel, *An Object-Oriented Model for the Sustainable Management of Evolving Spatio-Temporal Information* In : Geocomputation and Urban Planning edited by Beniamino Murgante, Giuseppe Borruso, Alessandra Lapucci. Series : Studies in Computational Intelligence, Vol. 176, Approx. 280 p., Springer 2009, ISBN : 978-3-540-89929-7

Revues nationales

Christine Plumejeaud, Jérôme Gensel, Marlène Villanova-Oliver, Maher Ben Rebah, Guillaume Vergnaud, *Modélisation des hiérarchies territoriales multiples évolutives*, Revue Internationale de Géomatique, Vol. 21/2, pp. 183-201, ed. Hermès, 2011.

Christine Plumejeaud, Jean-Marc Vincent, Claude Grasland, Jérôme Gensel, Hélène Mathian, Serge Guelton, Joël Boulier, *HyperSmooth : calcul et visualisation de cartes de potentiel interactives*, Colloque International de Géomatique et d'Analyse Spatiale. Numéro spécial de la Revue Nouvelle des Technologies de l'Information. RNTI-E-13, pp 19-42, ed. Cépaduès, 2008.

Conférences internationales

Gilles Pison, Hélène Mathian, Christine Plumejeaud, Jérôme Gensel, *Exploring world demography on line*. 25eme Conférence Cartographique Internationale – Paris, France, July, 2011

Christine Plumejeaud, Julie Prud'homme, Paule-Annick Davoine, Jérôme Gensel, *Transferring Indicators into Different Partitions of Geographic Space*. In : David Taniar and Osvaldo Gervasi and Beniamino Murgante and Eric Pardede and Bernady O. Apduhan (eds) : Computational Science and Its Applications – ICCA, LNCS, Heidelberg : Springer, Vol. 6016, pp. 445 ?460, 2010

Christine Plumejeaud, Jean-Marc Vincent, Claude Grasland, Sandro Bimonte, Hélène Mathian, Serge Guelton, Joël Boulier, Jérôme Gensel *HyperSmooth, a system for Interactive Spatial Analysis via Potential Maps*, 8th international Symposium on Web and Wireless Geographical Information Systems 2008 - W2GIS 2008, Shanghai, December 11-12, China, 2008.

CONFÉRENCES NATIONALES

Christine Plumejeaud, Jérôme Gensel, *Complexité liée à la variabilité sémantique des statistiques socio-économiques* Atelier FDC,EGC 2011, Brest, 25 Janvier, France, 2011 (in French).

Dounia Azzi, Christine Plumejeaud, Marlène Villanova-Oliver, Jérôme Gensel, *Vers un système pour l'évaluation de la qualité de données spatio-temporelles*, SAGEO 2010, Toulouse, November 17-19, France, 2010 (in French).

Christine Plumejeaud, Jérôme Gensel, Marlène Villanova-Oliver, *Opérationnalisation d ?un profil ISO 19115 pour des métadonnées socio-économiques*, INFORSID 2010, Marseille, May 25-28, France, 2010 (in French).

Christine Plumejeaud, Julie Prud'homme, Paule-Annick Davoine, Jérôme Gensel, *Etude de méthodes de transfert d'indicateurs associés à différents découpages du territoire - Application à la ville de Grenoble*, Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2009), Paris, November 25-27, France, 2009 (in French).

Christine Plumejeaud, Jérôme Gensel, Marlène Villanova-Oliver, Maher Ben Rebah, Guillaume Vergnaud, *Modélisation de hiérarchies territoriales multiples - Vers la gestion d'informations spatio-temporelles évolutives*, Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2009), Paris, November 25-27, France, 2009 (in French).

Christine Plumejeaud, Jean-Marc Vincent, Claude Grasland, Jérôme Gensel, Hélène Mathian, Serge Guelton, Joël Boulier, *HyperSmooth : calcul et visualisation de cartes de potentiel interactives*, Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2007), Clermont-Ferrand, June 18-20, France, 2007 (in French).

POSTERS ET COMMUNICATIONS ORALES

Christine Plumejeaud, Dounia Azzi, Marlène Villanova-Oliver, Jérôme Gensel, *Toward an interactive system for checking spatio-temporal data quality*. Mapping global change - Spatial Statistics 2011, 23 Mars 2011, Enschede, Pays-bas.

Christine Plumejeaud, Hélène Mathian, Jérôme Gensel, Claude Grasland, *Visualizing Evolutions of Socio-Economic Disparities with HyperTime* Workshop GeoSpatial Visual Analytics : Focus on Time of AGILE 2010. 11 May 2010, Guimaraes, Portugal.

Résumé :

Cette thèse se situe dans le domaine de la modélisation spatio-temporelle, et nos travaux portent plus particulièrement sur la gestion de l'information statistique territoriale. Aujourd'hui, la mise à disposition d'un grand volume de statistiques territoriales par différents producteurs (Eurostat, l'INSEE, l'Agence Européenne de l'Environnement, l'ONU, etc.) offre la perspective d'analyses riches, permettant de combiner des données portant sur des thématiques diverses (économiques, sociales, environnementales), à des niveaux d'étude du territoire multiples : du local (les communes) au global (les états). Cependant, il apparaît que les supports, les définitions, les modalités de classification, et le niveau de fiabilité de ces données ne sont pas homogènes, ni dans l'espace, ni dans le temps. De ce fait, les données sont difficilement comparables. Cette hétérogénéité est au cœur de notre problématique, et pour lui faire face, c'est-à-dire l'appréhender, la mesurer et la contrôler, nous faisons dans cette thèse trois propositions pour permettre *in fine* une exploitation avisée de ce type de données.

La première proposition a pour cible le support de l'information statistique territoriale, et cherche à rendre compte à la fois de son caractère évolutif et de son caractère hiérarchique. La deuxième proposition traite du problème de variabilité sémantique des valeurs statistiques associées au support, au moyen de métadonnées. Nous proposons un profil adapté du standard ISO 19115, facilitant l'acquisition de ces métadonnées pour des producteurs de données. La troisième proposition concerne la mise à disposition d'outils pour analyser et visualiser ces informations dans un mode interactif. Nous proposons une plate-forme dédiée aux analyses statistiques, visant en particulier à repérer des valeurs exceptionnelles (*outliers* en anglais), et à les mettre en relation avec leur origine, et les modalités de leur production.

L'ensemble de ces propositions a fait l'objet d'une validation avec leur implémentation dans le cadre d'un projet confié par l'observatoire européen du territoire, ESPON, portant sur la constitution d'une base de données multi-niveaux d'indicateurs statistiques sur l'Europe et son voisinage, depuis 1950 à 2050.

Mots clés : Statistiques socio-économiques, Modélisation spatio-temporelle, Changement du support, Métadonnées, Valeurs exceptionnelles, SIG.

Abstract :

In the field of spatiotemporal modelling, this research focuses specifically on the management of territorial statistical information. Today, the availability of large amounts of statistical information by different regional producers (Eurostat, INSEE, the European Environment Agency, the UN, etc..) offers a rich analytical perspective, by the combination of data on various topics (economic, social, environmental), at various levels of study : from the local (municipalities) to global (states). However, it appears that the spatial supports, the definitions, the various classifications and the reliability level of those data are very heterogeneous. This heterogeneity is at the core of our problem. In order to cope with that, that is to say to measure, control and analyse this heterogeneity, we draw three proposals allowing for a wiser exploitation of this kind of data.

The first proposal aims at taking into account the change of the support through the time, modelling both the evolutive aspect of the territories and their hierarchical organisation. The second proposal deals with the semantic variability of the data values associated to this support, through the use of metadata. A profile of the ISO 19115 standard is defined, in order to ease the edition of those metadata for data producers. The last proposal defines a platform dedicated to spatiotemporal data exploration and comparison. In particular, outliers can be discovered through the use of statistical methods, and their values can be discussed and documented further through the use of metadata showing their origin and how they have been produced.

These proposals have been implemented and tested in the framework of a European project funded by the European Spatial Planning and Observatory Network (ESPON), which aims at building a database handling statistical indicators at various levels in Europe and its neighbourhood, from 1950 up to 2050.

Title : Models et methods for evolutive spatio-temporal information.

Keywords : Socio-economic statistics, Spatio-temporal modelling, Change of support, Metadata, Outliers, GIS.