THÈSE EN COTUTELLE
UNIVERSITE PARIS 8 – UNIVERSITE DE TUNIS

Présentée par **Asma GHARBI**

En vue de l'obtention du grade de **Docteur**

Spécialité : **Informatique**

Une approche à base de règles d'association pour l'explication et la prévision de l'évolution territoriale

Présentée le 13/02/2018 devant le jury composé de :

Mme Hanene Azzag	Maîtresse de Conférences, HDR Université Paris 13, France	Rapporteure
Mme Amel Borgi	Maîtresse de Conférences, HDR Université de Tunis El Manar, Tunisie	Rapporteure
Mme Jacqueline Signorini	Professeure des Universités Université Paris 8, France	Examinatrice
M. Cyril De Runz	Maître de Conférences, HDR Université de Reims, France	Examinateur
M. Herman Akdag	Professeur des Universités Université Paris 8, France	Directeur
M. Sami Faiz	Professeur des Universités Université de La Manouba, Tunisie	Co-Directeur

Remerciements

J'aimerais tout d'abord remercier l'ensemble des membres de mon jury pour avoir accepté d'évaluer cette thèse : Mesdames Amel Borgi et Hanene Azzag pour avoir accepté d'en être les rapporteures, ainsi que Madame la Professeur Jacqueline Signorini pour avoir accepté d'examiner ce méemoire.

Je tiens à remercier mes directeurs de thèse Mr Herman AKDAG et Mr Sami FAIZ pour la confiance qu'ils m'ont accordée en acceptant de diriger mes travaux de recherche. J'aimerais également les remercier pour leur disponibilité, écoute, respect, encouragement et compréhension qui m'ont permis de me lever motivée et de continuer mon chemin malgré les moments de doutes.

C'est à M. Cyril De Runz que tourne ma gratitude la plus profonde pour les longues heures d'encadrement qu'il m'a consacrées, sa patience, ses conseils précieux et ses multiples encouragements.

J'adresse également mes remerciements aux chercheurs et personnels de LIASD pour l'accueil et les conditions de travail privilégiées ainsi que pour les séminaires de recherche qui furent pour moi un moyen d'élargir mes connaissances scientifiques. Je ne pourrais pas oublier mes collègues les docteurs et futurs docteurs Malangua, Besma, Akram, Rabah, Saloua pour la très bonne ambiance que j'ai toujours trouvée au bureau, ainsi que pour les nombreuses discussions qui m'ont beaucoup apporté tantôt humainement tantôt scientifiquement.

Ces remerciements ne peuvent se conclure sans exprimer ma gratitude à ma famille pour leur soutien quotidien, leur encouragement récurrent, et leur croyance en moi et en mes capacités. Je t'aime maman, je t'aime papa, je vous aime Zaynab, Ahmed et Hayet.

Enfin, mes remerciements les plus distinguées et ma reconnaissance la plus sincère vont à mon éepoux qui fut pour moi une raison de vivre, me battre et réussir. Je t'aime beaucoup Hilel. Merci pour ton soutien continu, pour tes

encouragements. Merci pour l'amour et la compréhension que tu m'as offert depuis le début de cette épreuve. Je te remercie aussi pour avoir accepté de sacrifier ton quotidien, ta famille, et tes amis en Tunisie pour m'accompagner dans la réalisation de mes ambitions. Je te remercie également pour avoir financé ce parcours sans le moindre regret ni plainte.

Résumé

Dans ce mémoire, nous partons de l'hypothèse que les dynamiques spatiales et les usages des objets géographiques peuvent, en partie, être expliqués voire anticipés par leurs historiques de changements de fonctions et de co-localisations. Ainsi, nous proposons d'exploiter la recherche des motifs fréquents et des règles d'associations pour en extraire des règles régissant ces dynamiques. Ce travail tente également d'adapter le processus de fouille de données pour tenir compte de la spécificité des données spatio-temporelles utilisées, en particulier, leur asymétrie.

Dans ce contexte, notre proposition traite des questions liées à la modélisation des relations spatio-temporelles incorporées dans le jeu de données, la représentation adéquate des données d'apprentissage, pour ainsi, produire des règles adaptées à notre problème de prédiction. La prise en compte de l'asymétrie des attributs d'apprentissage en termes de fréquence est traitée selon deux approches : une approche utilisant plusieurs seuils de support minimum et une approche traitant disjointement les attributs. Dans le cadre de la première approche, deux adaptations de l'algorithme MSApriori ont été proposées pour la définition et l'affectation de ces seuils. Quant à la deuxième approche, nous proposons l'algorithme BERA basé sur des sémantiques des prédicats pour la génération de règles en allant de la construction de la conclusion vers la construction de la prémisse.

Afin de vérifier et évaluer ces différentes propositions, nous proposons un dispositif expérimental (SAFFIET). À l'aide de ce dernier, une étude expérimentale est menée sur différents jeux de test issus des données Corine Land Cover.

Table des matières

Liste des tableaux

Table des figures

Liste des abréviations

— **minsup** : Minimum support
— **MBA** : Modèle à Base d'Agents
— **minconf** : Minimum confiance
— **AC** : Automate Cellulaire
— **ECD** : Extraction de Connaissances à partir des Données
— **KDD** : Knowledge Discovery in Database
— **ADN** : Acide désoxyribonucléique
— **conf** : confiance
— **CVC** : Chauffage, Ventilation et Climatisation
— **SGBD** : système de gestion de base de données
— **BERA** : Backward Extraction Rule Algorithm
— **SAFFIET** : Spatial And Functional Frequent Itemset Extraction Tool
— **SIG** : Système d'Information Géographique
— **CLC** : Corine Land Cover
— **RAC** : Règle d'Association Classifiante
— **MCD** : Modèle Conceptuel de données
— **SPF** : *Sequence of Past Functions* – Séquence des précédentes fonctions
— **SMI** : Support Minimum d'Item
— **SQL** : Structured Query Language
— **OST** : Objet Spatio-Temporel
— **OGC** : Open Geospatial Consortium
— **EPL** : Event Pattern Language
— **CDL** : Change Description Language
— **LUT** : Land-Use Transport Models

rajoute la liste des abréviations.

Chapitre 1

Introduction générale

1.1 Introduction

1.1.1 Fouille de données spatio-temporelles

L'information digitalisée a rendu possible la collecte et l'archivage d'une énorme quantité de données opérationnelles et facilement accessibles. Bien que ces entrepôts de données ne cessent de croitre principalement en raison de la disponibilité des systèmes de bases de données qui sont de plus en plus puissants et efficaces, leur richesse en connaissance incorporée semble toujours sous exploitée [Witten and Frank, 2005]. Pour combler ce besoin, plusieurs méthodes d'analyse et de traitement des données ont été proposées [Zaki and Wagner Meira, 2014]. Celles-ci visent essentiellement à découvrir et à extraire les connaissances cachées ou non-triviales qui utilisées à bon escient, serviront pour une deuxième phase d'exploitation de ces données : la modélisation. Ce sont ces modèles produits qui induiront différentes hypothèses et postulats à vocation explicative et/ou prédictive qui seront employés pour répondre à un problème donné. En effet le terme modèle représente, dans le domaine de la fouille de données, les relations entre les variables ou les valeurs décrivant des individus dans un jeu de données [Kantardzic, 2003]. Un modèle n'émane pas d'une théorie mais plutôt d'un bon ajustement aux données. Il tend à fournir toutes les règles de fonctionnement et peut même aider à prédire le fonctionnement futur du système étudié. Au fil des ans, différentes techniques de fouille de données à l'instar des modèles statistiques linéaires et non linéaires, des réseaux de neurones, des algorithmes génétiques, des arbres de décision, des méthodes de partitionnement, les règles d'association, *etc.* ont

été proposées afin de définir des modèles permettant d'identifier des caractéristiques des données, des patrons, ou des règles opérationnelles pouvant être employés pour diverses applications pratiques [Bharati and Ramageri, 2010].

Les techniques de fouille de données ont initialement ciblé des ensembles de données simples et structurés telles que les bases de données relationnelles ou des entrepôts de données structurées. Plus important encore, ces algorithmes ont été utilisés pour analyser des données qui se produisent dans une même période de temps ou, en d'autres termes, des données faiblement évolutives/dynamiques, *i.e.,* qui ne devraient pas subir beaucoup de changements au fil du temps. Cependant, le progrès en technologies du matériel, de l'information et de la communication, principalement l'internet, ont conduit à des méthodes de collecte et d'acquisition d'information plus sophistiquées (e.g. les capteurs, scanners, GPS, télémètre laser, *etc.*). Les données ainsi récoltées sont plus complexes à l'instar des données multimédias et des données spatio-temporelles [Longley et al., 2015].

L'information spatiale et temporelle est implicitement présente dans la plupart des bases de données. En effet, quel que soit le domaine d'application, chaque entité physique ou morale peut très souvent être associée à une localisation dans l'espace et certains de ses attributs peuvent varier avec le temps. Par conséquent, il est utile de développer des techniques qui résument efficacement ces données et qui découvrent leurs tendances spatio-temporelles, dans le cadre d'un modèle qui, ainsi, aide à la prise de décision [Cheng et al., 2014]. Ces modèles doivent saisir, entre autres, le comportement évolutif de ces entités au fils du temps et donc fournir un aperçu utile pour le suivi et la prédiction d'éventuelles occurrences d'évènements qui lui sont liés. Ces évènements peuvent alors être liés à des entités géoréférencées et correspondre aux changements d'états de ces dernières [Langran, 1992, Hornsby and Egenhofer, 2000, Spéry et al., 2001a]. L'observation de l'ensemble des évènements des différentes entités constituant un système (e.g. les quartiers (entités) d'une ville (système)) peut permettre de détecter un phénomène et d'informer sur sa dynamique. Par exemple, le phénomène de l'évolution « physique » d'une ville est observable à partir des différents changements fonctionnels (vocation) ou spatiaux (forme) que subissent les bâtiments ou les zones géographiques qui la constituent. Par conséquent, les techniques d'extraction de connaissances peuvent modéliser ce phénomène et fournir donc des postulats pour expliquer ou même prédire ses dynamiques. Finalement, les phénomènes spatio-temporels représentent des processus de changement exprimés par des séries ou séquences d'évène-

ments [Liu et al., 2016]. Ces évènements qui sont définis par des changements de caractéristiques des entités, peuvent ainsi être détectables à partir des différentes versions temporelles (ou historiques) de la base de données qui les décrivent. L'objectif de cette thèse est de chercher à détecter ces évènements et les explorer pour en déduire un modèle spatio-temporel capturant, si possible, l'essence du phénomène dans lequel ils s'inscrivent.

1.1.2 Application dans le suivi de changements d'occupation/usage du sol

Les centres urbains ne cessent de croître. Selon L'ONU-Habitat, il est prévu qu'en 2030, 60% du monde sera urbain [UN-Habitat, 2011]. Cette urbanisation du paysage s'exprimant par un changement de la couverture terrestre, s'accompagne de conséquences sur la faune, la flore et même les ressources naturelles indispensables pour l'existence du genre humain. Le développement durable et l'environnement sont devenus des enjeux existentiels qui requièrent le suivi, l'analyse et la modélisation de ces changements. Les méthodes telles que la fouille de données, offrent un immense potentiel de gestion, et d'analyse des données acquises au fil du temps pour décrire les dynamiques de ce genre de phénomène et dévoiler leurs facteurs moteurs. Ce genre de méthodes vise, donc, à aider les décideurs à proposer des modèles de développement et/ou d'utilisation des terres rendant l'exploitation des ressources, des infrastructures et des services publics la plus efficace possible avec un impact limité sur la faune et la flore. Un large panel de travaux, se basant sur la fouille de données spatio-temporelles pour traiter la question des changements d'occupation/usage du sol, ont été proposés [Jenerette and Wu, 2001, Yang et al., 2008, Charif et al., 2012, Malek et al., 2015, Qiang and Lam, 2015]. Ces travaux visent essentiellement à identifier et à caractériser ces changements ainsi qu'à découvrir les relations qu'ils ont avec les différents variables naturels et anthropogénique [Turner et al., 2007]. Dans la plupart de ces travaux, chaque phénomène spatio-temporel est étudié à travers le suivi des changements d'un type particulier de couverture (occupation) du sol, et chaque type de couverture est étudié par des variables qui lui sont spécifiques. Par exemple, pour étudier la déforestation, Mithal et al. [Mithal et al., 2011], se sont concentrés sur les changements de la couverture végétale et ont pour cela utilisé l'indice de

végétation amélioré [1] et la fraction de rayonnement solaire absorbée par les plantes [1]. Dans d'autres exemples tels que le suivi du bâti pour la modélisation de l'étalement urbain, des variables biophysique caractérisant le type de sol ou des variables socioéconomiques peuvent être exploitées. Bien qu'efficace, cette approche, spécifique au type de couverture suivi, se concentre plus sur les caractéristiques internes des zones étudiées et néglige l'effet des relations spatiales et temporelles des données (*i.e.* une telle propriété apparait à une telle co-localisation à un tel moment). Dans l'objectif de remédier à ces problématiques, nous proposons dans ce mémoire une approche générique (suivi de changement d'occupation du sol en général) qui se concentre sur les composantes spatiales et temporelles des données pour trouver des règles (modèle) régissant ce phénomène. Dans cette approche les données d'études sont représentées par un ensemble d'entités. Une entité est définie par une occupation du sol, associée à une zone géographique, et est valide pendant un certain intervalle de temps. La transition (antécédent/successeur) d'une occupation du sol à une autre représente une relation temporelle (de type séquence ou série). La proximité spatiale entre deux entités représente une relation spatiale de co-localisation (voisinage). C'est à partir de ces relations spatiales et temporelles que nous tentons, à l'aide des techniques de fouille de données, de définir un modèle explicatif et éventuellement prédictif des changements d'occupation du sol.

Une des méthodes prometteuses pour la découverte des relations entre les différentes variables stockées dans une base de données, est l'analyse des associations [Maragatham and Lakshmi, 2012]. Cette technique permet d'extraire des motifs qui sont soit des ensembles de variables connexes appelés *itemsets*, soit des règles qui associent l'occurrence des valeurs d'une variable à l'occurrence des valeurs d'autres variables. Dans le cas des données spatio-temporelles, le modèle produit consiste en un ensemble de motifs décrivant les liaisons entre quelques variables de localisation et quelques variables de période temporelle. Ce modèle permet, ainsi, de capturer les caractéristiques spatiales et temporelles des données, essentielles pour étudier l'évolution temporelle d'un ensemble de propriétés décrivant une zone.

1. respectivement, en anglais « Enhanced Vegetation Index » (EVI) et « Fraction of Photosynthetically Active Radiation » (FPAR), ce sont deux indicateurs de la concentration et la « verdure » de la végétation à un endroit donné.

1.2 Problématiques

Bien que la fouille de données soit fondamentalement conçue pour explorer des larges jeux de données (*i.e.* suffisamment grand pour être difficilement appréhendable par l'humain), il existe toujours des cas d'applications avec une petite quantité de données. Dans les domaines d'applications traditionnelles de l'informatique tels que le secteur bancaire et la gestion de relation avec les clients, la quantité et la complexité des données recueillies par les entreprises sont intrinsèquement importantes et fortement croissantes. Dans ce genre d'application, les données sont archivées principalement pour être analysées et pour en extraire des connaissances en utilisant des méthodes classiques telles que les statistiques, les voisinages et le clustering. Malheureusement, cette hypothèse de disponibilité des bases de données énormes et prêtes pour être analysées n'est pas toujours vraie. Dans ces cas, même si la base a une bonne quantité de données stockées, les données finales, qui peuvent être efficacement utilisées pour l'opération mentionnée ci-dessus, pourraient être beaucoup moins nombreuses. Cette problématique se pose communément dans les applications visant à fouiller dans des structures des données incorporés dans la base de données étudiée et non pas les données de la base en tant que telles. Par exemple, le jeu de données spatio-temporel peut être considérablement réduit si on cherche à fouiller dans les séquences d'événements ou les structures décrivant la configuration spatiale des entités géographiques qui y sont incorporées.

Plus que la quantité de données, nous devons également traiter la question de la qualité de ces données. Si nos données sont déséquilibrées, erronées ou sont sans rapport avec la résolution du problème traité (en terme de pertinence des caractéristiques), la taille de celles-ci n'a, donc, qu'une importance relative car elle serait réduite suite aux processus indispensables de nettoyage et de normalisation des données. Dans cette optique, il est nécessaire de faire face à des défis tels que :

— Données Asymétriques : dans certains cas, les données spatio-temporelles peuvent biaiser plus une propriété qu'une autre. Par exemple, l'imagerie par satellite à haute résolution fournit généralement une information spatiale abondante, mais pourrait ne pas enregistrer l'aspect temporel, comme le changement d'image en fonction du moment de la journée. Dans un autre cas, les capteurs stationnés pour la surveillance des événements donnent des informations précises et détaillées par rapport au temps des événements, mais offrent peu

d'informations sur les relations spatiales entre les capteurs répartis.
— Autocorrélation spatiale : cette situation se pose quand la valeur d'une variable à une localisation donnée est liée aux valeurs de la même variable dans les localisations voisines. C'est, notamment, le cas des données géoréférencées qui présentent une interdépendance spatiale. En d'autres termes les attributs des localisations voisines s'influencent entre eux et tendent d'avoir un plus grand degré de similarité. Par exemple, le type de couverture de sol tend à être plus similaire à une distance proche de quelques mètres qu'à une distance lointaine de quelques kilomètres.
— Prétraitement pour un apprentissage pertinent : dans quelques applications, dont la nôtre, on vise à fouiller des structures spécifiques incorporées dans la base de données et non pas les données en tant que telles. Tels sont les cas des séquences de changements d'occupation du sol ou bâtiments (entités géographique) ou des structures décrivant la configuration spatiale de ceux-ci. Ainsi, il devient nécessaire de mener un processus d'identification des relations spatiales (e.g. voisinage) et temporelles (évènement de changement d'occupation du sol détectables à partir des différentes versions temporelles de la base de données qui les décrivent).
— Scalabilité : l'approche d'apprentissage proposée doit tenir compte de la grande quantité de données à éventuellement traiter. Cependant, ceci fait pluôt partie des perspectives de cette thèse.

1.3 Contributions

Bien que la communauté scientifique se soit, de plus en plus, intéressée à explorer la répercussion des caractéristiques spatio-temporelles des données sur l'apprentissage, les relations spatiales existantes entre les entités étudiées et leurs impacts sur l'analyse des données restent, néanmoins, mal explorés.

Dans ce travail, nous nous sommes focalisés sur les relations spatiales et temporelles entre les données et leurs impacts sur l'opération d'apprentissage. Dans cette optique, nous avons proposé une approche à base de règles d'association dont les contributions se résument comme suit :

a) Aspect méthodologique

— Les relations spatiales et temporelles pour l'apprentissage : une entité est définie par une occupation du sol, associée à une zone géographique, et valide pendant un certain intervalle de temps. La transition d'une étiquette (description, sémantique ou fonction) à une autre représente une relation temporelle. La proximité spatiale entre deux entités représente une relation spatiale de co-localisation (voisinage). Afin d'identifier ces relations nous sommes passés d'un modèle en *snapshot* – *i.e.* un ensemble de cartes géographiques indépendantes décrivant un territoire géographique à des dates consécutives – à un modèle se basant sur le paradigme identitaire permettant le suivi des relations spatiales et temporelles d'une entité au fil du temps. C'est à partir de ces relations que nous tentons, à l'aide d'une technique de fouille de données, de définir un modèle explicatif et éventuellement prédictif des changements d'occupation du sol.
— Un pré-traitement des données adéquat au problème d'apprentissage : cette étape consiste à proposer, d'abord, une forme de règle jugée adéquate pour notre problème de prédiction, puis un format d'organisation des données d'apprentissage défini de façon à permettre la génération de ce type de règle. Dans notre cas, une règle est dite intéressante pour la prédiction si elle se base sur des items correspondant, respectivement, aux relations spatiales de voisinages et aux relations temporelles de succession d'états (fonction) des objets prédécesseurs pour prédire un item représentant la fonction d'un objet successeur.
— Des méthodes de génération de règles prenant en compte la nature déséquilibrée des données : nous proposons deux méthodes, une première consistant à adapter apriori en proposant d'utiliser plusieurs supports minimums définis selon deux propositions : méthode par analyse statistique et méthode par groupement [Gharbi et al., 2016] ; Une deuxième basée sur un algorithme définissant une sémantique particulière des prédicats pour genérer les règles en allant de la construction de la conclusion vers la construction de la prémisse.

b) Aspect logiciel

— L'outil *SAFFIET* : Nous présentons l'outil SAFFIET (Spatial And Functional Frequent Itemset Extraction Tool) qui exploite des algorithmes de recherche de motifs fréquents et de règles d'associations, pour extraire des règles d'évolution explicatives et prédictives régissant

les dynamiques spatiales. SAFFIET s'appuie sur les fonctionnalités de solutions logicielles existantes telles que *PostGIS* pour le stockage et la manipulation de l'information spatiale, *Weka* pour l'apprentissage, et *Qgis* pour la visualisation des résultats d'un prochain module de prédiction.

— L'autonomie : notre outil est conçu de manière à effectuer, d'une façon autonome, les différentes étapes méthodologiques exprimées ci-dessus. A partir d'une série temporelle de cartes géographiques décrivant un certain territoire, cet outil crée, automatiquement, un modèle de données identifiant les objets du territoire et traçant leurs évolutions, forme, à partir de ces données, une base d'apprentissage respectant un format prédéfini jugés adéquat à la génération de règles d'évolution, applique un algorithme de recherche de règles d'association afin d'en extraire ces règles-là.

c) Aspect applicatif

— Bien que les modèles prédictifs à base de règles d'associations ont été utilisés dans plusieurs domaines d'applications spatio-temporelles (le climat, e.g., le changement global ; la santé publique, e.g., la détection des zones chaudes ou avec un niveau élevé de criminalité ; la sécurité publique, e.g., suivi et prévision de la propagation des épidémies ; le commerce mobile, e.g., les services géo-dépendants ; *etc.*), peu de travaux utilisant ces techniques ont été proposés pour la modélisation explicative et prédictive des dynamiques territoriales. S'inscrivant dans ce domaine d'application, nous proposons une approche à base de règles d'association qui consiste à fouiller dans des relations spatiales et temporelles incorporées dans les données d'étude. En effet, nous considérons en particulier les relations temporelles de succession d'états et les relations spatiales de voisinage qui sont, respectivement issues d'une pré-étude des différentes évolutions précédentes (changements de fonctions) que subissent les objets géographiques et les configurations spatiales dans lesquelles ils se situent.

1.4 Organisation du manuscrit

Après l'introduction, la première partie de cette thèse, composée de trois chapitres, est consacrée à une étude critique de l'état de l'art. Le premier cha-

pitre (chapitre 2) de cette partie présente les différentes approches existantes pour modéliser des données temporelles, spatiales et spatio-temporelles. Le second chapitre introduit les méthodes informatisées de la modélisation qui vont d'une simple description et traçage de l'information spatio-temporelle décrivant un certain phénomène, à une explication de celui-ci afin d'assurer une assistance analytique, simulatrice et prédictive à leurs utilisateurs. Le troisième chapitre se concentrera sur les méthodes de fouille de données, et en particulier, la recherche de motifs fréquents et de règles d'association au coeur de nos contributions.

Dans la deuxième partie de cette thèse, nous présenterons nos propositions ainsi que les résultats qui en ont découlé. Ainsi, dans un premier chapitre (chapitre 5), nous explicitons la vision méthodologique de nos contributions et dans un deuxième chapitre (chapitre 6) nous décrivons le dispositif expérimental conçu pour les tester en analysant les résultats.

Enfin, nous concluons cette thèse en commentant l'état des lieux des contributions scientifiques. Nous discutons des limites de celles-ci, et nous proposons également quelques pistes qui font l'objet de perspectives à explorer, pour cette thèse.

Chapitre 2

Dynamiques spatio-temporelles : définitions et représentations

2.1 Introduction

Dans ce chapitre, nous présentons les notions de temps et de l'espace, l'espace/temps, ainsi que les différents formalismes et représentations qui leurs sont associés, du point de vue aussi bien conceptuel qu'informatique. Ainsi, nous introduisons, tout d'abord, la notion de temps, celle de l'espace et leurs différents modes de représentations dans les systèmes d'information. Ensuite, nous parcourons les principales études qui associent le temps et l'espace pour la modélisation des processus liés à la distribution, l'évolution et la diffusion de phénomènes spatio-temporels.

2.2 Temps

2.2.1 Définitions

Étant indissociable de l'expérience humaine, le temps a toujours été l'objet de nombreuses interrogations quant à sa nature et sa perception, sa représentation et ses mesures. En effet, les sciences humaines, physiques et mathématiques ont tenté d'apporter des explications à ce phénomène complexe et plusieurs définitions en sont découlées. Dans ce contexte, on distingue deux

principales conceptualisations du temps :

— Le temps subjectif ou psychologique qui selon Kant est une donnée subjective dépendante de l'homme et de son estimation de durée [Janiak, 2016] :

— Le temps objectif qui représente une mesure abstraite, est étudié par la science physique pour répondre à des questions liées aux lois de la nature. Étant défini, par Aristote comme une mesure du mouvement, le temps s'est trouvé lié depuis lors aux notions de changement, de durée et de matière en mouvement. Il a été originellement présenté par Galilée, comme une grandeur physique quantifiable pouvant lier des expériences par le biais d'opérations mathématiques et par Newton comme universel, absolu, uniforme et neutre ou externe aux phénomènes se produisant en son sein [Newton et al., 1759].

> « Le temps absolu, vrai et mathématique, sans relation à rien d'extérieur, coule uniformément, et s'appelle durée. » [Newton et al., 1759, p. 8].

Bien qu'il soit réel et quantifiable, le temps est un phénomène complexe dont la nature est insaisissable comme l'a exprimé Saint Augustin [Augustin d'Hippone, 1864] :

> « Qu'est-ce donc que le temps ? Si personne ne m'interroge, je le sais ; si je veux répondre à cette demande, je l'ignore. » [Augustin d'Hippone, 1864, Confessions, livre XI, p.479].

Par conséquent, les physiciens, selon Klein [Klein, 2009], se sont focalisés sur comment représenter et mesurer au mieux le temps plutôt qu'à répondre à la question de la nature du celui-ci.

2.2.2 Représentation de la notion du temps

Selon ces différentes cultures et les expériences qui leur sont associées, l'homme a envisagé le temps de diverses manières comme l'expriment Mercure D. et G. Pronovost [Mercure and Pronovost, 1989] :

> « Chaque type de société développe sa propre culture du temps. » (Mercure D. et G. Pronovost, Temps et Société, p.10)

Parmi celles-ci, les plus rependues sont (cf. figure 2.1) :

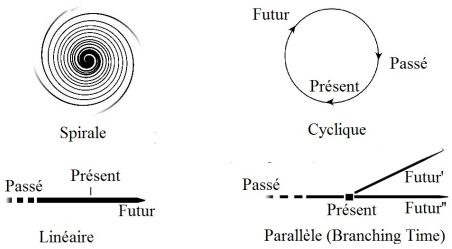


FIGURE 2.1 – Les différentes conceptions du temps.

— L'aspect cyclique du temps : basé sur la récursivité et où le temps est toujours tourné vers le passé. Dans cette conception, le passé, le présent et le futur s'interpénètrent en mouvement cyclique. En effet, les fidèles de cette vision se réfèrent à l'alternance des événements naturels tels que le jour et la nuit et les saisons, etc.
— L'aspect spiral, surtout adopté par la religion indienne qui consiste un retour circulaire avec une déviation linéaire. En d'autres termes, les événements se reproduisent mais pas exactement de la même façon.
— L'aspect linéaire du temps : le temps s'écoule du passé vers le futur en passant par le présent ce qui s'est inspiré des phénomènes tels que la transformation irréversible des créatures avec l'âge.

Contrairement au temps cyclique qui revient vers le passé, le temps linéaire est illustré en ligne continue et s'oriente vers le futur. Il est susceptible d'évoluer vers une conception parallèle si à un ou à plusieurs moments donnés il se divise pour renvoyer sur plusieurs scénarios d'avenir [Nain and Vardi, 2007].

De ce fait, vient l'orientation de la physique moderne vers l'adoption du temps linéaire qui vérifie la causalité et qui représente une manière de rangement ordonné des évènements où tout évènement est l'effet de la cause (évènement) qui l'a précédé. Dans cette optique deux principales approches pour la représentation du temps ont été proposées :

— L'approche newtonienne qui définit le temps comme une grandeur mesurable et quantifiable en dates ce qui permet de décrire le mouvement d'un corps par l'ensemble de ses positions à des dates successives exprimées en différentes unités (années, mois, heure, second, etc).
— L'approche de Leibniz qui à l'encontre de la précédente, dénie le caractère absolu et infini du temps et l'envisage comme un certain ar-

rangement entre des événements. En d'autres termes, une succession d'événements liés entre eux par des relations d'antériorité, de postériorité, et de simultanéité. Les structures quantitatives et qualitatives respectivement émanant de l'approche newtonienne et de celle de Leibniz [Thériault and Claramunt, 1999] donnent, en se combinant, une représentation hybride de façon à ce qu'une mesure de temps soit établie (temps mesuré et positionné selon sa coordonnée sur l'axe temporel) et une algèbre définissant les relations topologiques entre ces mesures (événements placés sur une échelle ordinale selon leurs positions relatives) soit appliquée.

Il est important de noter que dans ce travail, ce sont la conception linéaire parallèle (time-branching) et l'approche de représentation du temps de Leibniz qui sont employées. En effet, nous percevons le temps comme étant une ligne orientée du passé vers le futur sur laquelle nous plaçons les fonctions des objets selon leurs dates d'observation. À un point donné cette ligne peut présenter une disjonction en cas de division d'une entité impliquant la présence de deux fonctions sur la même empreinte spatiale, et une conjonction en cas de fusion de deux entités pour former une seule présentant une fonction unique.

Évènements et topologie Dans la conception linéaire, le temps est représenté comme une ligne droite sur laquelle des états temporels sont placés sous forme de points si on suppose qu'ils sont instantanés sans durée, ou sous forme d'intervalles si on leur associe une durée de validation. Formaliser ainsi le temps revient aussi à définir la notion de l'évolution temporelle ou le changement d'état temporel qui pour Allen [Allen, 1984]correspond principalement à trois notions : l'évènement, l'occurrence, et le processus. Il propose que les occurrences peuvent être désignées comme des évènements si, en se produisant, ils impliquent un résultat ou une culmination comme l'évènement énoncé dans la phrase « j'ai conduit ma voiture jusqu'à mon travail ». Le processus désigne une activité n'impliquant ni résultat ni culmination comme l'exemple exprimé dans la phrase « je suis en train de marcher ». La notion d'évènement, comme énoncée par Galton [Galton, 2000], peut consister en des occurrences discrètes d'un phénomène avec un instant de début et un instant de fin. Selon d'autres définitions [Cheylan et al., 1999], cette notion est liée aux objets pouvant être tangibles (ex : zone géographique) ou non (ex : communautés), et donc correspond à un changement d'états ou de

propriétés de cet objet valide sur un intervalle de temps. Cette conception des évènements peut correspondre donc à des fonctions allant d'une propriété d'un objet à une autre entre les instants t et $t+n$ [Higginbotham et al., 2000].

Dans d'autres conceptualisations, l'évènement est présenté comme la transition (en elle-même) de l'état de l'objet vers un autre ce qui implique sa considération comme instantanée ou brève sur l'échelle du temps [Kim, 1976]. En effet, les discussions à propos de la propriété continue ou discrète des évènements ont été l'objet de vifs débats depuis longtemps. Dans l'approche basée sur la propriété instantanée des évènements, l'écoulement du temps est représenté comme un ensemble d'instants de temps, chacun correspondant à un état d'un objet avec une relation binaire de priorité entre eux (précédence). Cette représentation n'est souvent pas adéquate pour modéliser les évènements du monde réel ayant une durée tels que les évènements exprimés dans la phrase « la construction de la présente basilique Saint-Pierre a pris plus de 120 ans » ou la phrase « il m'a fallu une demi-heure pour finir mon dîner avant de pouvoir regarder mon film pendant le restant de la soirée ». En outre, l'approche à base d'intervalles est ontologiquement plus riche car elle présente plus de relations possibles entre les évènements valides sur des intervalles plutôt que sur des instants telles que :

— Relation de précédence $e1 < e2$
— Relation d'inclusion $e1 \subseteq e2$
— Relation de chevauchement $e1\ o\ e2$
— etc.

Dans [Allen, 1984, Allen and Ferguson, 1994], Allen, qui suppose que les évènements ne peuvent pas avoir des durées nulles et donc doivent être représentés par des intervalles temporels, a proposé une algèbre pour traiter les relations entre ces intervalles. Il a proposé 13 relations topologiques illustrées dans la figure 2.2. L'algèbre d'Allen a connu plusieurs extensions telles que celle proposée par Ladkin [Ladkin, 1987] qui consiste à introduire 5 qualificatifs des relations entre les intervalles convexes qui sont : Souvent (mostly), Toujours (always), Partiellement (partially), Parfois (sometimes), D'une façon disjointe (disjointly) (e.g. I_1 et I_2 deux intervalles, la relation de précédence entre eux peut s'énoncer comme suit : I_1 est souvent avant I_2).

Une autre généralisation de l'approche proposée par Freksa [Freksa, 1992] considère les semi-intervalles comme une tentative pour la représentation de certaines informations temporelles qui sont importantes mais incomplètes. Par exemple, nous pouvons connaître la date de naissance d'une personne mais pas sa date de décès.

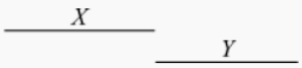
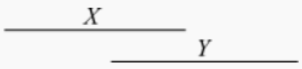
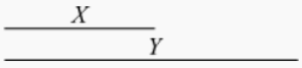
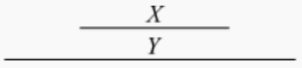
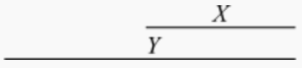
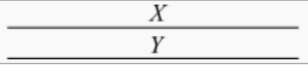
Relation	Illustration	Interpretation
$X < Y$ $Y > X$		X takes place before Y
$X\ \mathbf{m}\ Y$ $Y\ \mathbf{mi}\ X$		X meets Y (*i* stands for *inverse*)
$X\ \mathbf{o}\ Y$ $Y\ \mathbf{oi}\ X$		X overlaps with Y
$X\ \mathbf{s}\ Y$ $Y\ \mathbf{si}\ X$		X starts Y
$X\ \mathbf{d}\ Y$ $Y\ \mathbf{di}\ X$		X during Y
$X\ \mathbf{f}\ Y$ $Y\ \mathbf{fi}\ X$		X finishes Y
$X = Y$		X is equal to Y

FIGURE 2.2 – Illustration des relations entre les intervalles selon l'algèbre temporelle d'Allen [Allen, 1983].

Une autre conception des évènements a été proposée par Quine [Quine, 1953]. Elle consiste à considérer un évènement comme un objet ayant quatre dimensions spatio-temporelles (définis sur 3 coordonnées d'espace et une coordonnée de temps), et que cet évènement occupe pleinement la région spatiotemporelle en question. Ceci implique que deux évènements ne peuvent pas avoir lieu dans une même région spatio-temporelle et donc il est impossible qu'un évènement puisse se répéter comme on ne peut pas déplacer l'espace-temps [Livet, 2008]. En revanche, un évènement peut en inclure un autre : une région spatiotemporelle (qui correspond à l'évènement) peut inclure une ou plusieurs sous-régions.

En effet, le choix de la représentation des évènements (instant ou intervalle) est principalement dépendant des évènements ou phénomènes à représenter, les données disponibles et les relations à représenter correspondant aux besoins d'utilisation. Par exemple, si la requête porte sur la présence

ou non d'un bâtiment dans une zone urbaine à une date bien déterminée, la représentation avec des instances est favorisée, tandis que si elle porte sur la durée durant laquelle un bâtiment avait eu comme fonction « bâtiment résidentiel », une représentation par des intervalles est à privilégier.

2.2.3 Représentation du temps dans les systèmes d'information

Bien qu'il soit défini comme un phénomène continu, le temps ne peut être transcrit sur un support informatique que de façon discrète comme il est rare, voire impossible de disposer de toutes les données (sans interruption) sur une période donnée. Par conséquent, ceci implique le choix d'un niveau de résolution appelé aussi la granularité. La granularité ou la résolution temporelle est définie comme le temps d'observation qui se répète régulièrement et qui est réparti sur la durée totale du déroulement d'un phénomène. Dans les systèmes informatiques, la représentation du temps peut se faire suivant plusieurs niveaux de granularité selon les données stockées et leurs contextes (ex : l'utilisation de l'année pour la datation des naissances pour la réalisation d'une statistique sur la croissance populaire, dans un pays donné, pendant la dernière décennie) Par exemple, dans la norme ISO8601 [ISO, 2004], une représentation numérique acceptée à l'échelon internationale, six niveaux de granularité (année, mois, jour, heure, minute, seconde) sont définis.

Dans les bases de données, la composante temporelle des données est considérée suivant essentiellement trois points de vue :

- Le temps de validité : correspond à la durée de validité d'une donnée attachée à une entité du monde réel. Il est souvent représenté par un intervalle de validité.
- Le temps de transaction : représente l'instant où un évènement est enregistré dans la base de données. Il peut aussi correspondre à l'intervalle pendant lequel un fait est stocké dans la base.
- Le temps utilisateur : peut correspondre au temps perçu, dont la sémantique est seulement connue par l'utilisateur [Guida et al., 1999], ou le temps d'usage par une application (temps d'utilisation d'une donnée pour réaliser un certain traitement).

La distinction entre temps de validité et temps d'enregistrement sert à différencier les bases de données en quatre catégories [Snodgrass and Ahn, 1985, Paque, 2004, Ott and Swiaczny, 2001] :

— Base de données statique : l'historicité n'est pas prise en compte dans la base de données statique. Les nouvelles données remplacent les précédentes.
— Base de données rollback : les données possèdent un ou des attributs de temps de transaction. Chaque modification des données est enregistrée.
— Base de données historique : les données possèdent un ou des attributs de temps de validité. Elles peuvent être modifiées plusieurs fois, une seule version par temps de validité est conservée.
— Base de données bi-temporelle : les données possèdent des attributs de temps de transaction et de temps de validité. En effet, tout est conservé tout en distinguant les faits qui sont valides de celles qui sont erronées.

Plusieurs travaux ont proposé une conception bi-temporelle des données pour prendre en compte, à la fois, le temps de validité et le temps de transaction des données, [Snodgrass, 1992, Claramunt and Thériault, 1995a, Worboys, 1998]. Conserver le temps de transaction est crucial pour certaines applications qui reposent essentiellement sur les notions de versionnement et d'historisation des données : par exemple, la publication, par l'institut géographique nationale, des cartographies et d'indicateurs sujets à des révisions.

Cette conception (bi-temporelle) permet, donc, une vision plus complète et plus précise des données en répondant à des requêtes sur les données les plus précises possibles selon nos actuels critères ; les données comme on les a aperçues à n'importe quel moment ; et quand et pourquoi les données les plus précises ont changé.

Ces efforts de modélisation de la composante temporelle des données ont été accompagnés par plusieurs extensions des langages d'interrogation déjà existants tels que SQL pour le modèle relationnel. TSQL2 permet l'incorporation de l'information temporelle au niveau des tuples ou ce qu'on appelle l'estampillage des tuples. Elle a également rendu possible la gestion du temps de validité, du temps de transaction, et de l'aspect bi-temporel des données selon aussi bien une représentation en intervalles (period-stamped time) qu'une représentation en points temporels (point-stamped time). Ultérieurement, TSQL2 a été, en partie, incorporé dans le standard suivant SQL:1999 (SQL3) pour la définition d'un sous-standard temporel le SQL/Temporal avec principalement une modification du type de données *PERIOD*. Un autre standard appelé SQL:201 a été publié en décembre 2011. Ses principales différences vis-à-vis de l'approche TSQL2 consistent en :

— L'abandon du concept d'attributs cachés, *i.e.* les attributs d'horodatage [1] sont anonymes dans les tables, et donc cachés à l'utilisateur.
— Le remplacement des préfixes par des prédicats temporels à cause de l'incohérence logique qu'ils induisent [Darwen and Date, 2006].
— L'introduction d'un nouveau type de données PERIODFOR utilisé pour désigner l'information durée (intervalle de temps) au lieu de prévoir deux colonnes de date pour le stockage [Kulkarni and Michels, 2012].

Plusieurs SGBD ont proposé des modules pour gérer l'aspect temporel des données. Parmi ceux-ci, citons :

— Teradata qui a fourni deux produits. Teradata 13.10 et Teradata 14 qui présentent des caractéristiques temporelles, intégrées dans la base de données et basées sur TSQL2.
— Oracle Workspace Manager de Oracle qui offre la possibilité de gérer des versions actuelles et historiques de données dans une même base de données.
— IBM DB2: La version 10 a ajouté une caractéristique appelée « *time travel query* » qui se base sur les capacités temporelles (capabilities) offerts par le standard SQL:2011.
— Microsoft SQL Server qui a introduit dans SQL Server 2016 la caractéristique « *Temporal Tables* » [Lien1] [2].
— MarkLogic qui a introduit la caractéristique bitemporelle des données dans sa version 8.0 dans le module « *Temporal Collections* » [Lien2] [3].
— PostgreSQL : à partir de sa version 9.2, étendu (temporal 0.7.1.) [4] pour supporter et gérer les données temporelles. Ainsi, il offre d'autres types de données (ex : timestamp, date, time, interval) supportés par différentes fonctions et opérations autochtones ainsi que par des nouvelles fonctions temporelles (localtime, make_date, etc.). L'extension Temporal Table disponible sur PGXN [5], supporte la plupart des caractéristiques fondamentales du standards SQL:2011. Selon sa documentation, PostgreSQL est conforme à 160 des 179 caractéristiques de

1. L'horodatage (en anglais timestamping) est un mécanisme qui consiste à associer une date et une heure à un événement, une information ou une donnée informatique.
2. http://channel9.msdn.com/Shows/Data-Exposed/Temporal-in-SQL-Server-2016
3. http://docs.marklogic.com/guide/temporal
4. http://www.pgxn.org/dist/temporal/
5. PGXN (PostGresql Extension Network) est un système central de distribution pour les bibliothèques d'extension de PostgreSQL open-source.

base de ce standard. D'ailleurs, aucun des SGBD existants ne prétend être conforme à la totalité des caractéristiques de base du standard SQL:2011.

Dans cette section, nous avons introduit la notion de temps et ses différents modes de représentations dans les systèmes d'information. Notre sujet portant sur la fouille de données spatio-temporelles, nous abordons, dans la section suivante, la question de l'espace.

2.3 Espace

2.3.1 Définitions

Depuis qu'on l'a considéré comme un concept central de la géographie scientifique, les définitions de l'espace se sont multipliées, allant d'une conception simpliste le présentant comme une étendue délimitée de la terre, à d'autres conceptions plus élaborées tenant compte des objets et les relations qui sont en leur sein ainsi que de l'empreinte culturelle et sociale que laisse l'homme en agissant sur lui. Selon son angle de perception, l'espace est aussi classifiable en catégories (espace vécu, espace de vie, espace perçu, espace représenté, espace produit, espace social, etc.) [Di Méo, 1985] qui représentent, comme l'exprime Di Méo :

> « Simplement des modalités différentes de sa prise en compte : modalité de l'action pour l'espace produit par les sociétés, modalité de la connaissance ou de la cognition (faculté pour l'esprit humain d'enregistrer des informations) pour l'espace perçu et représenté, modalité de l'existence humaine pour l'espace vécu. » (Di Méo, Les formations socio-spatiales ou la dimension infra-régionale en géographie, p.27)

En effet, les géographes différencient entre l'espace concret qui consiste en un ensemble des lieux (espace physique de faible étendue) souvent portant des toponymes[6] et l'espace géographique ou l'espace formel [Pierre and Verger, 1970b] qui correspond à des mesures, des relations, des qualifications (ex : industriel, productif, européen) et des représentations

6. Un nom visant à identifier très précisément un détail géographique localisé. Il n'est pas affecté arbitrairement mais de manière à décrire le paysage et évoquer les activités que les habitants y exerçaient.

cartographiques. Cet espace géographique est envisagé, principalement, selon deux perspectives :

— L'espace comme support ou cadre de référence, où on localise des objets et où l'on envisage leurs relations sur, principalement, le critère de distance. Dans cette conception l'espace est indépendant des facteurs externes (absolu), et ses propriétés (ex : conditions géo-climatiques) sont homogènes et isotropes (sont les mêmes dans toutes les directions).
— L'espace relatif qui constitue le fruit d'interactions entre différents objets et phénomènes. Il se décrit non seulement avec des coordonnées mais avec des propriétés variables dans le temps et dans l'espace. Dans cette conception l'espace est à la fois le contenant et le contenu [Bailly et al., 2016].

C'est cette espace (relatif) qui fait de plus en plus l'objet d'études dans le domaine de la géographie ou dans ceux qui lui sont connexes. Il représente une extension plus riche de la conception absolue de l'espace car il la complémente en rajoutant des attributs de contenu (population, activités, climat, etc.). Dans ce travail, principalement visant à étudier l'évolution d'un espace géographique, nous adoptons la conceptualisation de l'espace comme étant absolu. Conséquemment, le modéliser revient à définir et à étudier les objets ou entités qui sont en son sein. Nous nous intéressons aux relations topologiques entre les objets afin de construire des relations de voisinage et de succession temporelle. Cependant, bien que nous nous basons sur un espace absolu, nous nous positionnons dans l'hypothèse géographique des dynamiques, c'est-à-dire que l'évolution d'une entité dépend des échanges et des rapports qu'elle a avec les autres entités à proximité [Pumain and Saint-Julien, 1997]. Notre travail exploite l'une des lois fondamentales de la géographie édictée par [Tobler, 1970] qui s'énonce comme suit :

> « Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés. »(Tobler, A Computer Movie Simulating Urban Growth in the Detroit Region, p.236).

2.3.2 Représentation de l'espace dans les systèmes d'information géographique

Au niveau informatique, représenter l'espace revient à représenter les entités spatiales comme des types de données spatiaux dont les relations sont calculées par des opérateurs spatiaux (e.g. les opérateurs topologiques : intersecte, croise, à l'intérieur, etc.). Ces entités, localisables sur la surface de la terre, sont associées à des informations dites géographiques et qui, chacune, possède une composante graphique et une composante attributaire (cf. figure 2.3). La composante graphique consiste en une description de la forme de l'entité ainsi qu'en sa localisation dans un référentiel cartographique, tandis que la composante attributaire représente les caractéristiques décrivant l'entité, telles que la description géométrique et les caractéristiques thématiques.

Un système d'information géographique (SIG) est défini selon le comité fédérale de Coordination inter-agences pour la Cartographie Numérique aux États Unis (1988) comme un système informatique de matériel, de logiciel et de processus dont la vocation est de collecter, gérer, manipuler, analyser, modéliser et afficher les données spatialement référencées afin de résoudre des problèmes complexes d'aménagement et de gestion. Les SIG sont de plus en plus développés et utilisés dans de nombreuses applications telles que : la topographie, la cartographie, la planification urbaine, l'occupation du sol, la gestion des événements d'urbanisme, la gestion des risques, la gestion des sources d'eau, etc. Une gamme de SIG libres et gratuits, est disponible (Quantum GIS QGIS [7], Geographic Resources Analysis Support System GIS (GRASS [8]), OrbisGIS [9], gvSIG [10], GeoTools [11], etc.). L'augmentation constante du volume des produits géographiques numériques est accompagnée de challenges tels que le maintien des bases de données spatiales « fraîches » à faible coût et à haute fréquence. C'est, effectivement, ce point qui représente l'un des plus grands problèmes et obstacles à l'application et à la promotion des SIG [Walter and Fritsch, 2000, Ramirez, 1998, Shi and Shibasaki, 2000].

7. http://www.qgis.org
8. http://grass.itc.it/index.php
9. http://orbisgis.org/
10. http://www.gvsig.com/
11. http://www.geotools.org/

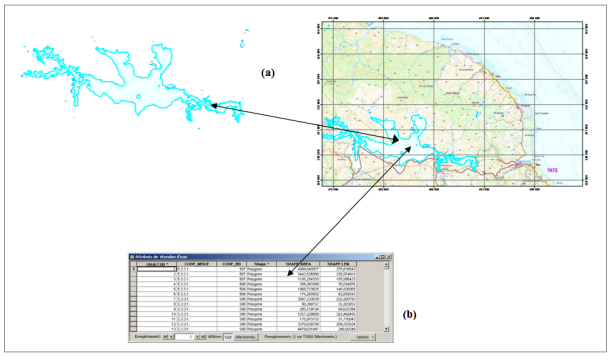


FIGURE 2.3 – Illustration des composantes graphiques (a) et attributaires (b) de l'information géographique.

2.3.2.1 Représentation des données géographiques

Un SIG stocke les informations géographiques sous la forme de couches thématiques reliées les unes aux autres grâce au géoréférencement des données. La représentation informatique de l'information géographique, dans un SIG, peut se faire selon deux modes : Raster ou Vecteur (cf. figure 2.4).

Dans le mode raster, la réalité (l'espace) est représentée sous forme d'une grille régulière. Chaque cellule de ce maillage ou cette grille couvre une aire géographique et à laquelle une et une seule valeur, correspondant à la dimension étudiée, est assignée (e.g. occupation du sol, pollution, altitude). Les cellules ou pixels sont caractérisés par une couleur ou intensité de gris et une taille qui correspondent, respectivement, à la valeur de l'attribut étudié et la résolution spatiale. Plus la taille de la cellule est petite plus la résolution est grande et plus l'information est précise. Le mode raster représente une perception de l'espace comme un continuum ou se produisent des phénomènes sociaux. Ainsi, un objet géographique n'est pas explicitement représenté mais défini par la juxtaposition des cellules similaires. Cependant, dans la représentation vectorielle, une forme constitue, elle seule, un objet.

En effet, le mode vecteur correspond à une représentation conceptuelle de l'espace par un ensemble de primitives géométriques, telles que : les points, les lignes, les courbes et les polygones. Les points définissent les localisations d'éléments séparés pour des phénomènes géographiques trop petits pour être représentés par des lignes ou des surfaces qui n'ont pas de surface réelle

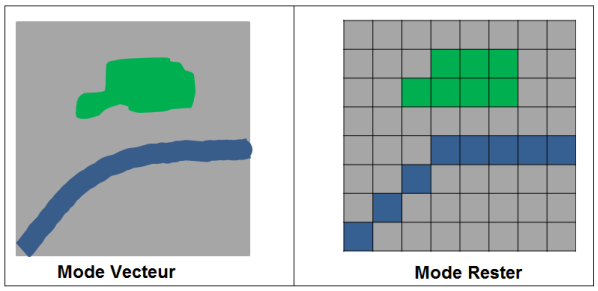


FIGURE 2.4 – Les modes raster et vecteur de représentation de l'information géographique [Caloz and Collet, 2011]

comme les points cotés. Les lignes représentent les formes des objets géographiques trop étroits pour être décrits par des surfaces (ex : rue ou rivières) ou des objets linéaires qui ont une longueur mais pas de surface comme les courbes de niveau. Les polygones représentent la forme et la localisation d'objets homogènes comme des pays, des parcelles, des types de sols, etc. Ces primitives correspondent à des entités ou des objets géographiques qui sont chacun dotés d'un identifiant permettant de les relier à une table de leurs attributs.

Outre ces primitives traditionnelles, les SIG proposent d'autres primitives pour la modélisation de l'information spatiale. Ainsi, MapInfo utilise le type region pour décrire une collection de polygones. Oracle Spatial qui n'utilise pas de type particulier pour les collections de polygones, distingue en revanche le rectangle (rectangle) comme un type particulier de polygone. De même, dans ArcGIS, les types PolygoneM et MultiPatch sont proposés pour représenter, respectivement, les collections de polygones et les collections d'objets de différents types. Afin de faire face à ce problème d'hétérogénéité, le modèle le Simple Features Specification [OGC, 1999] a été proposé par l'Open Geospatial Consortium (OGC [12]) en 1999. Celui-ci a, par la suite, évolué pour donner lieu à deux normes, les normes ISO 19107-Spatial Schema, et ISO 19123-Schema for coverage geometry and functions. Ainsi, les SIG et les bases de données spatiales peuvent s'appuyer aujourd'hui sur

12. Un consortium industriel international de plus de 521 entreprises, organismes gouvernementaux et universités. Sa mission consiste à de faire progresser le développement et l'utilisation des normes internationales et des services de soutien qui favorisent l'interopérabilité géospatiale

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, il existe une alternative pour les SIG qui n'utilisent pas ces normes pour la représentation interne des données. C'est celle d'exporter et d'échanger les données dans un format commun. À cette fin, le langage GML (*Geography Markup Language*) qui est une implémentation en XML des modèles standards de l'ISO 19107, a également été défini 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, 2007].

2.3.2.2 Représentation des relations spatiales

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. Afin de standardiser cette représentation, des normes ont été produites par des organismes comme l'OGC. En ce qui concerne la représentation des relations spatiales (recouvrement, disjonction, inclusion, etc.), des modèles tels que le modèle 9-intersection et la méthode dimensionnelle étendue (dimension extended method - DEM) ont été adoptés par des organismes de standardisation tels que l'OGC. En effet, le modèle 9-intersection a servi de support aux spécifications de l'OGC et du DGIWG [13], qui elles-mêmes sont intégrées dans les SGBD spatiaux qui proposent des fonctions prédéfinies pour calculer la topologie d'entités spatiales. Dans le SGBD PostgreSQL et sa cartouche spatiale PostGIS, les opérations topologiques proposées intègrent les modèles CBM et 9-i [Egenhofer and Franzosa, 1991, Clementini et al., 1993].

2.4 Espace-Temps

Un phénomène géographique ou spatio-temporel est un phénomène qui implique un changement dans l'espace et dans le temps. La littérature géographique abonde en études qui associent le temps et l'espace pour l'analyse ou l'explication des processus liés à la distribution, l'évolution et la diffusion de ces phénomènes. Selon Peuquet [Peuquet, 1994], généralement, de manière

13. DGIWG est l'organisme multi-national responsable de la normalisation géospatiale pour les organisations de défense des pays membres.

similaire à la représentation du temps ou de l'espace, deux vues ontologiques ont été adoptées pour la définition des structures espace-temps : l'approche absolue qui identifie l'espace comme une collection de points et le temps comme un ensemble d'instants qui existent par eux-mêmes ; et l'approche relative qui se focalise sur les entités du monde réel et leurs relations mutuelles pour définir une toile espace-temps subjective.

En se basant sur ces deux paradigmes de modélisation complémentaires, l'ultime objectif de la communauté scientifique est de réaliser un système d'information géographique temporel dit idéal. Celui-ci permettrait d'étudier les phénomènes du monde réel tout en étant capable de tracer les changements dans une zone à travers le stockage et la gestion des données géographiques historiques et anticipées [Beller et al., 1991, Langran, 1992, Frank, 1994, Flewelling et al., 1992, Claramunt and Theriault, 1996]. Peuquet [Peuquet, 1994], postule que disposer de ce système qui, efficacement et simultanément, fonctionne sur les vues absolue et relative de l'espace et le temps, implique que les composantes géographiques (où), temporelles (quand) et thématiques (quelles) qui définissent les dynamiques spatiales soient implémentées en utilisant un modèle de données homogènes. En d'autres termes, les dynamiques spatiales sont représentées dans les SIG selon ces trois dimensions, dont les combinaisons permettent de répondre à des questions liées à l'évolution, la distribution et la diffusion du phénomène étudié telles que :

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

Bien qu'elle permette de répondre aux différents requêtes permettant de tracer les changements, cette triade est incapable de produire des descriptions liées au processus et évènements responsables de ces changements. Celles-ci correspondent aux réponses à la question ajoutée par Thériault et Claramunt [Thériault and Claramunt, 1999] pour compléter la triade de Peuquet : « Comment tel objet s'est trouvé à telle localisation à tel moment? » (cf. figure 2.5).

La triade de Peuquet ainsi que sa version améliorée par Thériault constitue un socle commun aux différents modèles proposés dans la littérature que l'on peut organiser selon trois catégories : les modèles basés sur une représentation implicite des changements et des phénomènes spatio-temporels, les modèles basés sur une représentation explicite des changements et les modèles basés sur une représentation explicite des changements ainsi que des

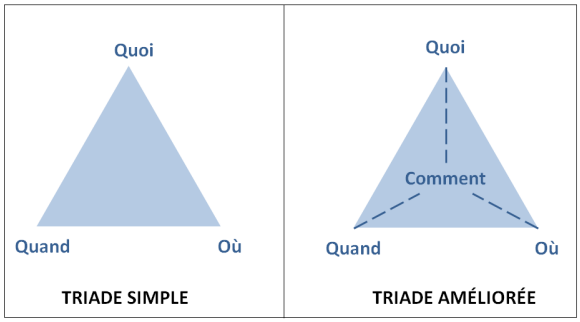


FIGURE 2.5 – Triade spatio-temporelle simple [Peuquet, 1994], puis complétée [Thériault and Claramunt, 1999].

phénomènes qui en sont à l'origine.

2.4.1 Modèles basés sur une représentation implicite des changements et des phénomènes spatio-temporels

Les modèles les plus connus de cette catégorie ont été inspirés par les premières tentatives pour intégrer la dimension temporelle dans les bases de données relationnelles pour créer ainsi les bases de données temporelles. Dans les SIG, ces premiers efforts consistèrent à attribuer une étiquette temporelle, aux couches, aux attributs, et aux objets spatiaux. Ces trois solutions d'étiquetage temporel correspondent respectivement aux modèles à base de succession d'états, appelés aussi modèles en snapshot [Armstrong, 1988], les modèles composites spatio-temporels [Langran and Chrisman, 1988] et les modèles basés sur le paradigme identitaire [Worboys, 1998].

2.4.1.1 Modèles à base de succession d'état

2.4.1.1.1 Modèle de superposition des couches datées Le modèle de superposition de couches datées, dit aussi « en snapshot » [Armstrong, 1988], représente un modèle de type ad-hoc qui vise à décrire l'évolution d'un phénomène spatio-temporel à travers la succession d'états de son support spatial (cf. figure 2.6). Selon ce modèle, la dimension temporelle est intégrée dans le SIG à travers la datation d'un ensemble de

couches raster superposées décrivant, chacune, l'état instantané de l'espace géographique étudié. Ainsi, ce modèle adopte une représentation de temps linéaire et orthogonale au plan spatial et suppose que le support spatial est invariant.

Le modèle en snapshot présente plusieurs avantages qui sont principalement : sa simplicité, son intuitivité, sa compatibilité avec les données raster (ex. satellitaires) qui, grâce au progrès technologique, sont massivement disponibles, son adéquation au mode d'approvisionnement des données dans les SIG et surtout sa capacité à récupérer, intuitivement, l'état d'un territoire à n'importe quel moment appartenant à l'intervalle temporel étudié.

Alors que déterminer l'état de la cellule e_i à t_i est facile, déterminer comment celui-ci a évolué entre t_i et t_{i+1} ou qu'elle est la fréquence de ce changement est plus complexe. C'est-à-dire que ce modèle permet seulement d'enregistrer les états discrets d'un espace géographique et, donc, néglige le stockage explicite du changement en eux-mêmes et encore moins les évènements qui en sont responsables. En outre, le modèle en snapshot est incapable de fournir une représentation explicite des versions des objets géographiques (ensemble de cellules) ; bien qu'il soit possible de déterminer visuellement les différentes versions d'un objet géographique, en faisant défiler les différents snapshots, il n'est, en aucun cas, possible d'enregistrer leurs topologies et structures temporelles (*i.e.* spécifier la version antérieure ou successeure d'un objet géographique).

Se basant sur une représentation inchangée du support spatial, cette modélisation du spatio-temporel est très simpliste et peu réaliste. D'ailleurs, le nombre des entités géographiques, leurs formes et leurs superficies, qui ensemble, constituent le support spatial, sont réellement variables.

Mis à part, ces lacunes conceptuelles, le modèle en snapshot présente un inconvénient computationnel : il est redondant et donc très couteux en mémoire car, pour chaque date d'étude, toute la couche est enregistrée dans la base de données ce qui entraîne une duplication des données inchangées (portions d'espace qui n'ont pas changé d'états).

Pour faire face à ces problèmes, spécifiquement la redondance et l'invariabilité du support spatial, plusieurs extensions à ce modèle ont été proposées, à savoir les modèles à base de composites espace-temps, et les modèles inspirés de la programmation orientée objet.

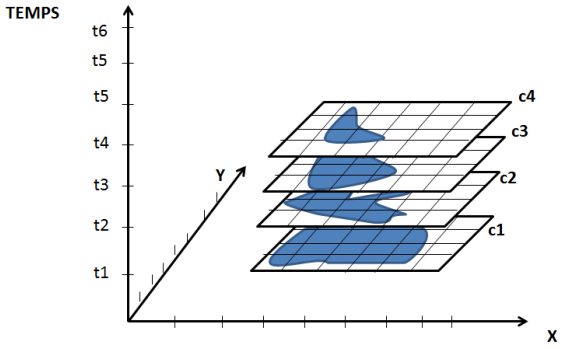


FIGURE 2.6 – Illustration du modèle en snapshot. C_i représentent les différents snapshots, t_i représentent les dates.

2.4.1.1.2 Modèles à base de composites spatio-temporel Visant essentiellement à traiter les problèmes de la variabilité du support spatial et la redondance des données, cette approche représente une extension « vectorielle » aux modèles en snapshots. Le point commun entre les modèles de cette approche consiste en la définition d'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.

Le modèle proposé par Langran et Chrisman [Langran and Chrisman, 1988] et baptisé « modèle à composition spatio-temporelle », représente un précurseur de cette catégorie de modèles. Au lieu d'être rattaché aux localisations, comme les modèles en snapshot, ce modèle raisonne, plutôt, sur les entités et plus spécifiquement sur un attribut de celles-ci.

Ainsi, le monde est perçu comme une collection d'entités qui évoluent, individuellement et à des rythmes différents, dans le temps. L'idée est d'introduire successivement les couches dans l'ordre chronologique, les superposer et n'enregistrer que les portions d'espaces nouvelles (l'empiètement créé lors d'un changement de l'emprise spatiale d'un objet) ou modifiées (modification de l'attribut thématique). En se basant sur la couche initiale (*i.e.* couche référentielle fixe qui décrit l'espace géographique à la date initiale) et les enregistrements datés des changements, ce modèle permet de construire l'état de l'espace à n'importe quelle date (composite spatio-temporel à cette date) en rajoutant à la carte initiale l'accumulation des éléments enregistrés

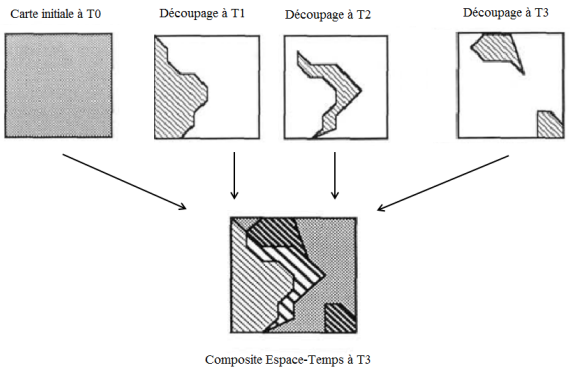


FIGURE 2.7 – Utilisation du modèle composites spatio-temporels pour la reconstitution d'un territoire à une date donnée T_3.

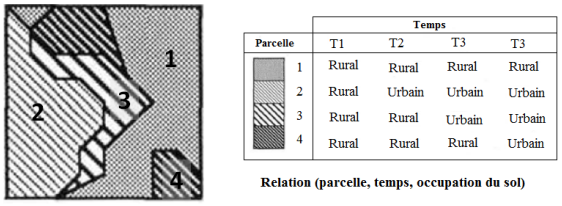
jusqu'à la date choisie, comme illustré dans la figure 2.7.

Tout en réduisant l'espace de stockage, ce premier modèle se base sur la carte de base et les différents composites spatio-temporels pour répondre à des requêtes liées à l'histoire des changements d'une parcelle donnée, telles que :

— Quelle était l'état des données à la date T_i?
— Qu'est ce qui a changé entre les dates T_i et T_j?
— Quelles sont les versions correspondantes à un tel objet et quand est ce que celui-ci a muté?
— Quelle est la fréquence des changements?

Ce modèle a été amélioré, par la suite, par Belussi et al. [Belussi et al., 1999] qui procèdent à la construction et le stockage d'une seule couche historique au lieu de conserver les changements dans les versions successives. Celle-ci intègre toutes les mises à jour historiques dans une même couche fusionnée par l'intersection de toutes les couches étudiées (cf. figure 2.8). Ainsi, l'espace mémoire est économisé car seules la couche initiale de référence, et la couche historique sont stockées.

Bien que l'approche à base de composites spatio-temporels permette de déterminer les versions d'un objet à une date postérieure, cette approche ne propose pas de réponse à la question « Comment un tel objet a évolué entre t_i et t_j? ». Ceci est principalement dû à la rupture de l'identité d'un l'objet initiale lors de son évolution. En effet, les versions postérieures de cet objet sont stockées dans la base sous des identifiant différents.



	Temps			
Parcelle	T1	T2	T3	T3
1	Rural	Rural	Rural	Rural
2	Rural	Urbain	Urbain	Urbain
3	Rural	Rural	Urbain	Urbain
4	Rural	Rural	Rural	Urbain

Relation (parcelle, temps, occupation du sol)

FIGURE 2.8 – Illustration de la couche historique dans le modèle à base de composites spatio-temporels adaptés.

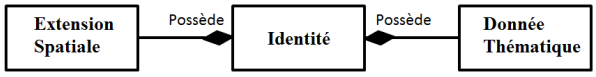


FIGURE 2.9 – Structuration de l'entité géographique [Cheylan et al., 1997].

C'est dans ce contexte que des modèles se basant sur un paradigme identitaire ont été proposés.

2.4.1.1.3 Modèles basés sur le paradigme identitaire Cette catégorie de modèle se base essentiellement sur le paradigme identitaire proposé par Cheylan [Cheylan and Lardon, 1993] pour modéliser le domaine espace-temps. Ce paradigme consiste à accorder une identité propre à chaque zone du support de l'information thématique. Les modèles spatio-temporels basés sur ce paradigme, notamment celui de [Worboys, 1998], se focalise donc sur l'objet géographique (entité géographique) et le définissent comme étant constitué de trois parties qui évoluent indépendamment aux cours du temps : une extension spatiale, une donnée thématique ou attributaire et une identité (cf. figure 2.9).

En effet, Worboys propose, d'identifier les attributs spatiaux, sémantiques et thématiques comme étant des atomes qui composent une entité géographique. Ainsi, le monde est perçu comme étant un ensemble d'atomes tridimensionnels et non un ensemble d'entités géographiques globales. Chaque atome couvre une portion d'espace 2D définie par les valeurs des propriétés thématiques et spatiales de l'entité qui sont valides et stables pendant un intervalle de temps fixé. Le temps, conceptualisé comme orthogonal au

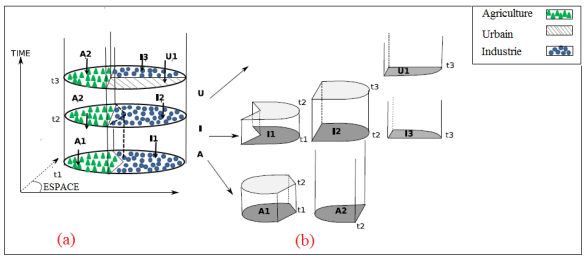


FIGURE 2.10 – Décomposition des objets espace-temps en atomes espaces-temps [Worboys, 1998]. (a) : les objets espace-temps modélisant les changements de support (U, I, A). (b) : les atomes espaces temps correspondants à ces objets (A_1, I_1, I_2, A_2, I_3, U_1).

plan spatial, représente la troisième dimension (cf. figure 2.10). Ce modèle permet, donc, la reconstitution d'une couche géographique tout en faisant l'assemblage des atomes ayant un intervalle de validité temporelle commun (une projection des atomes sur le plan spatial).

Dans les tentatives d'implémentation des modèles basés sur le paradigme identitaire, les chercheurs [Dell'Erba and Libourel, 1997] se sont basés sur l'approche orientée objet du développement informatique qui définit, l'atome comme objet et lui accorde comme propriétés, ses attributs thématiques et spatiaux et une identité. L'identité est selon cette approche intrinsèque et indépendante des autres attributs évolutifs. D'ailleurs, c'est cette identité qui permet de reconnaitre et de localiser dans le temps une entité ayant évolué suite aux différents changements liés à ses attributs.

En effet, dans ces modèles le lien de composition entre l'entité géographique et ses atomes est conservé via l'attribution d'un identifiant unique à l'entité comme aux atomes. Par conséquent, le suivi de l'évolution d'une entité géographique est désormais possible à travers la projection de ses atomes sur l'axe temporel. Autrement dit, la question de la triade de Peuquet améliorée, « Comment un tel objet a évolué entre t_i et t_j ? », peut finalement trouver une réponse. C'est, effectivement, la contribution majeure de ce type de modèle par rapport à ceux mentionnés précédemment. Cependant, la définition de l'identité qui est au coeur de cette contribution a toujours été controversée [Roshannejad and Kainz, 1986, Khoshafian and Copeland, 1986, Hallot and Billen, 2016, Hornsby and Egenhofer, 2000]. Celle-ci doit être

définie selon des critères objectifs afin d'assurer la continuité dans le temps (*i.e.* impossibilité de localiser une entité géographique dans le temps si son identité change lors de l'évolution). Une hypothèse fréquente est de placer le traceur de l'identité sur l'un des attributs comme proposé dans les travaux de Worboys [Worboys, 1992] et Kauppinen [Kauppinen and Hyvönen, 2007], qui, respectivement, ont employé le type d'usage du sol et l'empreinte spatiale comme identifiants. Dans le cadre de l'identification des entités géographiques physiquement existantes, l'empreinte spatiale est, très souvent, prise comme marqueur de l'identité, parfois d'une façon implicite [Kauppinen and Hyvönen, 2007]. Toutefois, utiliser un seul attribut de l'entité géographique comme marqueur de l'identité est insuffisant pour une modélisation réaliste. Par exemple, lorsqu'on identifie une entité par son nom celle-ci est considéré morte dès qu'elle change de nom alors qu'en réalité il s'agit de la même entité avec les mêmes attributs mais avec une appellation différente. Un exemple réel de ceci est le cas de Saint-Petersbourg qui a changé de nom en Petrograd, puis en Leningrad pour revenir actuellement à Saint-petersbourg. Théoriquement, ceci veut dire qu'on a à faire à trois entités avec des historiques différents et indépendants alors qu'en réalité ce ne sont que des atomes de la même entité qui sont supposés lui être associés dans un même trajet de vie. De même, placer le traceur de l'identité, seulement, sur l'empreinte spatiale est susceptible d'empirer la situation. En réalité si la surface d'une entité s'étend l'objet reste le même alors que, théoriquement, ceci implique la rupture de l'identifiant (changement de l'empreinte spatiale) et donc implique la mort de cette entité et la création d'une nouvelle. Également, dans le cas où deux versions d'un même zonage portent une différence légère dans l'empreinte spatiale d'une entité celles-ci peuvent être considérées comme deux entités indépendantes alors, qu'en réalité, elles représentent la même entité géographique.

Comparé aux modèles précédemment explicités, notamment les modèles à base de composites espace-temps et les modèles snapshots, les modèles basés sur les notions d'identité et d'atomes présentent une cohérence temporelle, de façon à ce que les requêtes concernant les évolutions d'un objet donné puissent trouver des réponses. En revanche, Yuan [Yuan, 1999] argumente que cette cohérence, étant liée à une position spatiale, ne permet pas de répondre à des requêtes de type mouvements. En outre, les requêtes temporelles sont restreintes aux attributs possédant un attribut temporel ce qui implique que les requêtes simplement temporelles ou simplement spatiales sont plus compliquées. Un inconvénient supplémentaire de ces modèles est qu'ils ne

représentent que les changements soudains au travers desquels il est difficile d'identifier des processus tels que le mouvement ou le changement d'une entité de l'environnement géographique. En effet, une représentation explicite de ces derniers, selon plusieurs chercheurs dont Langran [Langran, 1992], est au coeur d'un SIG temporel, d'où l'émergence d'une autre classe de modèle dite « modèles à base de changements ».

2.4.2 Modèles basés sur une représentation explicite des changements

Plutôt que de mettre l'emphase sur la succession d'états de l'espace géographique d'étude, ces modèles l'ont mise sur les changements que ses objets constituants ainsi que leurs attributs et leurs relations peuvent subir. Langran [Langran, 1992] postule que ces changements, leurs localisations, et leurs datations devraient être explicitement représentés dans un tel SIG temporel. En effet, les premières tentatives [Hornsby and Egenhofer, 1998] ont porté sur les changements liées à l'existence ou la non-existence des objets spatiaux. Dans ce contexte Hornsby [Hornsby and Egenhofer, 1998] a proposé d'attribuer à chaque objet un identificateur unique et persistant ainsi qu'un état variable en fonction des changements qu'il subit. L'historique de vie d'un objet peut, ainsi, être déterminé à travers les différents états qui lui sont associés, au fil du temps. Un langage visuel CDL (Change Description Language) a été introduit pour concrétiser cette modélisation des changements. Ce dernier modélise un changement par une transition décrivant la progression d'un état d'identité vers un autre et le représente par une flèche orientée. L'origine de celle-ci représente l'état de l'identité pré-changement et sa destination représente son état post-changement. Un objet peut prendre un des trois états suivants (cf. figure 2.11) : objet existant avec histoire, objet inexistant sans histoire, objet inexistant avec histoire.

En se basant sur les transitions et les états, représentant les primitives de ce modèle, deux familles de changements peuvent être distinguées. Les changements d'identité impliquant un seul objet et ceux qui en impliquent plusieurs. La figure 2.11 proposée dans [Hornsby and Egenhofer, 2000] illustre les neufs changements produits par la transition entre deux états d'un seul objet, dont la suppression, la création et l'incarnation.

La figure 2.12 illustre deux exemples de changement impliquant plus qu'un objet. L'exemple (a) représente deux objets A et B où l'objet B est créé à

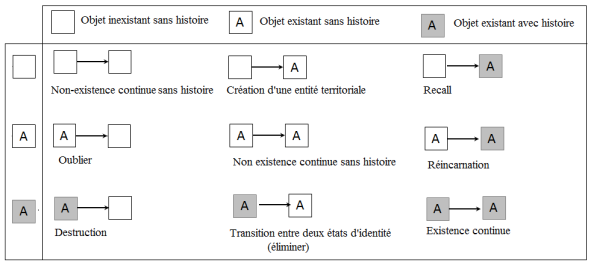


FIGURE 2.11 – Les neufs changements produits par la transition entre deux états d'un seul objet [Hornsby and Egenhofer, 2000].

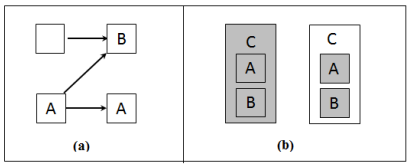


FIGURE 2.12 – Exemples de changements impliquant plusieurs objets.

partir de l'objet A. l'exemple (b) implique trois objets et représente une relation de composition entre l'objet C et les objets A et B. Dans le contexte de l'évolution territoriale, ces deux exemples peuvent correspondre, respectivement à la création d'une ville à partir d'une autre et à la modification de l'organisation hiérarchique des zonages (e.g. la disparition de l'URSS mais la survie des pays qui la composent – la Russie, la Lituanie, la Géorgie, etc.). Bien qu'intéressante, la représentation des changements à travers les transitions reste insuffisante. Elle ne permet pas de représenter avec une même transition un changement touchant à plusieurs objets. De plus, elle se focalise, exclusivement, sur les transitions entre les états des identités et manque, donc, une modélisation plus concrète de l'information spatiale et thématique et des entités géographiques. D'ailleurs, même la notion d'identité d'un objet géographique, demeure, ici, assez confuse (*i.e.* liée au nom, à l'emprise spatiale, à d'autres critères, ou n'est définie par aucune règle ?). En outre, que le changement temporel soit représenté implicitement en supposant un axe de temps ordonné, l'information sur sa durée est négligée.

Dans des contributions ultérieures, notamment celles de l'équipe de

Claramunt [Claramunt and Thériault, 1995b, Claramunt et al., 1997b, Claramunt et al., 1997a, Sriti et al., 2005] des représentations plus fines des changements ont été proposées. Claramunt et Thériault [Claramunt and Thériault, 1995b] ont, d'abord, introduit une typologie des changements spatio-temporels plus variée. S'inscrivant toujours dans l'approche identitaire, les entités, dans ce modèle, évoluent sous l'effet de trois familles de changements. Les changements sur une seule entité indépendante – la disparition, l'expansion et la rotation, etc. –, ceux liés aux relations fonctionnelles entre un certain nombre d'entité interdépendantes – transmission, permutation, etc.– et ceux représentant des restructurations territoriales impliquant, également, plusieurs unités géographiques – fusion, réallocation, division, etc. Il convient de noter que plusieurs autres travaux se basant, également, sur la modélisation explicite des changements spatio-temporels, ont abouti à des types de changements similaires, à savoir : la création et la cessation dans le modèle de graphe historique de Renolen [Renolen, 1996], la division et la fusion dans le modèle de changement de régions de Kauppinen *et al.* [Kauppinen et al., 2008] ; et la fusion, la division, l'extraction, l'intégration, la rectification et l'expropriation dans celui de Spery [Spéry et al., 2001b].

Le modèle de Claramunt et Thériault [Claramunt and Thériault, 1995b] a été, ensuite, étendu par Thériault et al [Thériault et al., 1999] en proposant une taxonomie des évolutions liées à un ensemble d'entités et non pas à des entités considérées d'une façon individuelle. Cependant, l'identification et la qualification des changements décrits dans ces deux modélisations restent insuffisantes pour découvrir les relations de causalités qui produisent les évènements [14] observés. Dans ce contexte plusieurs travaux ont tenté de modéliser les entités et les changements liés à leurs évolutions. Par exemple, Claramunt et al [Claramunt and Theriault, 1996] ont choisi d'utiliser un langage – Event Pattern Language (EPL) – pour définir des changements composites à partir des changements basiques. Worboys [Worboys, 2005] a également proposé un formalisme et un langage basés sur les évènements pour représenter les changements mais cette fois-ci focalisés sur l'espace. Ce travail a été par la suite étendu par Jiang et Worboys [Jiang and Worboys, 2009] pour différencier et spécifier les différents types d'évènements spatiaux, qualifiés d'évènements topologiques. Pour chaque snapshot, les évènements sont modélisés sous forme d'arbre et la modification spatiale d'un évènement est caractérisée par les

14. Evènement ici désigne les changements observés

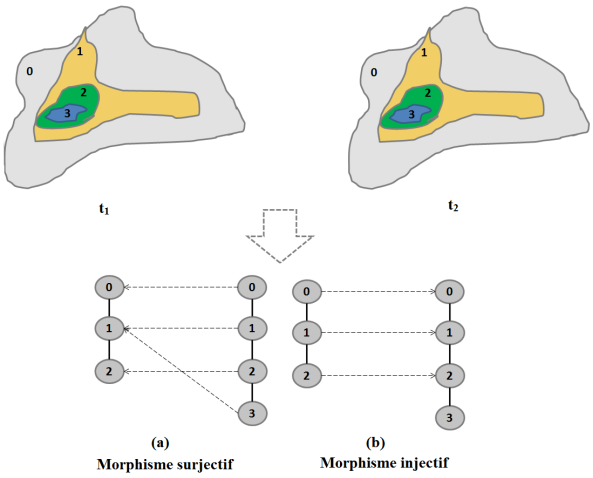


FIGURE 2.13 – Modélisation des changements sous la forme de morphismes d'arbres [Jiang and Worboys, 2009] : (a) cas d'une séparation ; (b) cas d'une insertion.

morphismes entre les arbres (cf. figure 2.13).

Afin de fournir un mécanisme de représentation et d'analyse des réseaux qui caractérisent et relient les évolutions entre entités spatiales, des modèles se basant sur les théories des graphes ont été proposés, à savoir les modèles basées sur les graphes de connectivité [Jiang et al., 2000, Jiang and Claramunt, 2004]; les modèles basées sur les graphes spatio-temporels de filiation [Spéry et al., 2001b] et leurs extensions proposées dans [Sriti et al., 2005] et [Stell, 2003].

En somme, l'idée directrice de ces modèles consiste à représenter, explicitement, les processus de changement et les modéliser en tant que causes de transformation de l'espace géographique d'étude. Les deux principaux atouts de cette approche sont d'une part la construction d'une généalogie ou graphe historique des unités qui constituent le support [Spéry et al., 2001b, Renolen, 1996], et d'autre part l'adoption d'une structure linéaire de la dimension temps avec une topologie (évènements relativement ordonnés) qui assure une vision à plusieurs niveaux de granularité de cette dimension. C'est pourquoi cette dernière tendance correspond à une vraie intégration du temps dans un espace 4-D (trois dimensions pour l'espace et une dimension, à part, pour le temps).

Bien qu'ils permettent une représentation explicite des changements, ces modèles sont incapables de fournir une représentation explicite des phénomènes qui les ont causés. Dans ce contexte, une autre tendance de modélisation s'est fondée sur l'idée que ces phénomènes, désignés comme évènements et processus, peuvent être représentés par des entités « occurrentes ». D'ailleurs c'est cette réflexion sur les entités occurrentes, qui est à l'origine de l'émergence d'une troisième famille de modèles : les modèles basés sur les évènements/processus.

2.4.3 Modèles basés sur les évènements/processus

Cette famille de modèle postule que les changements observés au niveau des entités spatio-temporelles sont causés par des phénomènes appelé processus et évènements. Contrairement aux modèles basés sur les changements, ceux-ci sont explicitement modélisés par des entités conceptuelles à part entière ayant leurs propres attributs et qui occupent la dimension temporelle [Galton, 2007]. Deux visions du monde géographique sont considérées par ces modèles :

— La vision introduite par Grenon et Smith [Grenon and Smith, 2004], qui, étant complémentaire aux modèles basés sur les changements, stipule que le monde se compose d'entités « occurrentes » et d'autres dites « continuantes ». Les entités occurrentes représentent les évènements et processus qui existent puis disparaissent tels que les catastrophes naturelles, la construction ou la destruction d'un bâtiment, etc. et les entités continuantes représentent celles qui persistent dans le temps tels que les objets géographiques. Selon [Grenon and Smith, 2004], deux relations possibles entre les objets et les événements peuvent être distinguées : la participation et l'implication – les objets participent à des événements et les événements impliquent des objets. Par exemple, l'événement construction d'un hôtel implique la création d'un objet hôtel, etc.

— La vision introduite par Reitsma [Reitsma, 2005], se contente de modéliser seulement les entités représentant les processus ou les changements, appelés flux. Ainsi ce type de modélisation ne prend pas en compte les objets du monde réel et est donc limité en termes de phénomènes possible à modéliser. En effet, cette vision n'est adaptée qu'à la modélisation des processus physiques tels que l'érosion ou la pluie [Reitsma and Dubayah, 2007]. Par conséquent, la vision de

Grenon et Smith [Grenon and Smith, 2004] est plus répandue dans la littérature.

Outre ces deux visions de représenter les évènements et processus, d'autres débats philosophiques et ontologiques ont porté sur les définitions mêmes d'un processus et d'un évènement, la distinction entre ces deux concepts et les relations qui peuvent exister entre eux. Dans plusieurs travaux [Forbus, 1984, Claramunt et al., 1997b, Thériault et al., 1999, Galton, 2001], le processus est présenté comme l'entité responsable d'un changement remarqué qui est, typiquement, modélisé par une séquence d'états [Worboys, 2001, Yuan, 2001]; et l'évènement comme l'occurrence d'une chose significative qui est d'intérêt pour le domaine étudié [Peuquet, 1994, Galton, 2001, Worboys, 2001, Yuan, 2001]. D'un point de vue temporel, on distingue un processus d'un évènement en ce que le premier correspond à un intervalle temporel ouvert, tandis que le second correspond à un intervalle fermé [Yuan, 2001, Galton, 2006, Galton, 2007]. Dans ce même contexte, d'autres [Grenon and Smith, 2004, Frank, 2008] considèrent qu'un processus est duratif, alors que l'événement est une transition instantanée. Dans [Galton, 2006, Galton, 2007, Mourelatos, 1978], le critère de d'*homogénéité* fut proposé pour la distinction entre les processus et les évènements : les processus sont considérés comme homogènes (*i.e.* leurs parties temporelles sont du même type qu'eux-mêmes) alors que les évènements ne le sont pas. Par exemple, le processus courir consiste à courir dans toutes ses parties temporelles, alors que l'événement excursion peut être composé d'activités de natures différentes (courir, nager, marche, etc.) [Galton, 2007, Galton, 2006].

En ce qui concerne les relations entre ces deux concepts, pour certains, les événements sont composés de processus [Claramunt et al., 1997b, Yuan, 2001, McIntosh and Yuan, 2005, Galton, 2001, Galton, 2004, Galton, 2007] alors que, pour d'autres, les processus sont composés d'événements [Worboys, 2001]. Outre la composition, des chercheurs [Grenon and Smith, 2004, Worboys and Hornsby, 2004, Galton and Worboys, 2005] proposent d'autres relations telle que *causer* (*initier ou faciliter*), *bloquer ou terminer l'existence d'un autre événement.*

2.5 Conclusion

Notre sujet porte sur la fouille de données spatio-temporelles pour l'extraction de règles explicatives et prédictives du phénomène de changement

d'occupation/usage du sol. Dans ce contexte, nous adoptons la conceptualisation de l'espace comme étant absolu et, ainsi, nous le modélisons à travers la définition et l'étude des objets ou entités qui sont en son sein. Bien que nous nous basons sur un espace absolu, nous nous positionnons dans l'hypothèse géographique des dynamiques, c'est-à-dire que l'évolution d'une entité dépend des échanges et des rapports qu'elle a avec les autres entités à proximité [Pumain and Saint-Julien, 1997]. Nous nous intéressons, donc, aux relations topologiques entre ces objets afin de construire des relations de voisinage et de succession temporelle dont l'analyse, lors de l'apprentissage, peut apporter des réponses sur la dynamique du phénomène étudié. Dans ce contexte, nous partons d'un modèle en couches datées indépendantes, nous exploitons le paradigme identitaire pour identifier les objets et leurs relations spatio-temporelles, pour, enfin, construire un modèle qui trace leurs évolutions à travers une représentation explicite de leurs changements au cours du temps.

Dans ce présent mémoire, nous supposons que les données sont acquises et définies à une seule et même échelle spatiale.

Chapitre 3

Méthodes informatisées pour la modélisation des changements d'occupation/usage du sol

3.1 Introduction

Dans ce chapitre, nous présentons les approches informatisées exploitant, en partie, les représentations de l'information spatio-temporelle (chapitre 2) et qui vont au-delà d'une simple description et traçage des changements d'occupation/usage du sol, à une explication de celles-ci permettant une assistance analytique, simulatrice et prédictive de ce phénomène. Nous donnons un aperçu de ces modèles du point de vue de différents travaux de classification (section 3.2) tout en en mettant en exergue les classifications basées sur les approches de modélisation utilisées (section 3.4).

Ces dernières décennies ont vu l'évolution des modèles de changements d'occupation du sol allant de modèles traditionnels et statiques tels que les modèles mathématiques et linéaires, à des modèles plus complexes et dynamiques dont l'apparition été déclenchée par l'application de la modélisation informatique dans les études urbaines, au milieu des années 1950 [Voorhees, 1959, Batty, 2004]. Dans ce contexte, ce chapitre se concentre sur les modèles dynamiques : en particulier, la catégorie de modèles à base de règles, car l'objectif général de cette recherche est de mettre à profit les techniques informatiques, en particulier le *datamining*, pour construire un système de caractérisation de modèles explicatifs/prédictifs.

3.2 Classification des modèles

En raison de l'augmentation exponentielle du nombre de modèles présentés, une pluralité de recherches, visant à profondément les étudier, ont été proposées. Elles ont, principalement, procédé par identifier des catégories puis classer les modèles selon celles-ci dans l'objectif de les reconnaître, les différencier et mieux les comprendre ainsi que leurs caractéristiques. Par conséquent, de nombreux critères de classification ont été proposés. Par exemple, Merlin [Merlin, 1974] les a classifiés, selon l'objectif de modélisation en modèle exploratoires, descriptifs, opérationnels et prédictifs [Echenique, 1972, Verburg et al., 2004], de leurs coté, ont considéré deux critères de classification : le niveau d'analyse donnant lieu aux catégories niveau micro, niveau macro et niveau croisé ; et le niveau d'intégration permettant d'identifier des modèles à faible intégration, à intégration moyenne, et à forte intégration. Dans un travail plus récent, [Silva and Wu, 2012] ont fourni une classification multi-critères. Ainsi, selon le critère utilité ou objectif de modélisation, ils ont proposé cinq catégories de modèles de développement territorial, à savoir les modèles de changement d'occupation/usage du sol, d'étalement urbain, d'usage du sol pour le transport, d'évaluation d'impacts, et de projection analytique (*comprehensive projection*). Selon l'accent spatial (*spatial emphasis*), Silva et Wu [Silva and Wu, 2012] ont ainsi proposé les catégories : orientées spatial, orientées a-spatiale et intégrées. Ils ont, également, selon le niveau d'analyse, identifié les mêmes catégories de modèles définies par [Verburg et al., 2004] – niveau micro, niveau macro et niveau croisé – et selon les critères d'échelle spatiale et d'échelle temporelle, ils ont identifié, respectivement, les catégories des modèles à échelle régionale, des modèles à échelle métropolitaine et des modèles à plusieurs échelles ainsi que celles des modèles à long, à moyen ou à court terme.

3.3 Classification à base d'approches

Dans cette section, nous proposons un tour d'horizons de la littérature pour identifier des différentes approches de modélisation utilisées. Dans le travail de [Mitasova and Mitas, 1998], quatre approches principales ont été mises en évidence : l'approche stochastique, l'approche déterministe, l'approche multi-agents et l'approche basée sur les règles.

L'approche stochastique consiste à modéliser l'évolution des phénomènes

spatio-temporels en traçant l'évolution de certaines variables (les décrivant) pouvant changer, stochastiquement ou aléatoirement, avec certaines probabilités. Dans ces modèles, la projection est basée sur un ensemble de valeurs aléatoires ; les sorties sont enregistrées et la projection est répétée avec un nouvel ensemble de valeurs aléatoires des variables. Ces étapes sont répétées jusqu'à ce qu'une quantité suffisante de données soit recueillie [Molchanov and Woyczynski, 2012]. La principale différence entre l'approche stochastique et l'approche déterministe est que, dans la déterministe, pour les mêmes entrées, le modèle conduit aux mêmes sorties alors que, pour la stochastique différentes sorties peuvent être produites pour les mêmes entrées.

Dans l'approche multi-agents, un ensemble d'agents [1] et leurs interactions sont employés pour simuler des systèmes adaptatifs complexes qui représentent les phénomènes spatio-temporels (généralement urbains) abordés. SWARM [Iba, 2013] est un exemple classique de simulateurs basés sur des agents. Il permet aux utilisateurs de tester des théories scientifiques, de réaliser différents types d'analyses sur des données naturelles ou autres et de réaliser et visualiser des expériences. En ce qui concerne l'approche à base de règles, reporté dans ce même travail [Mitasova and Mitas, 1998], les auteurs soulignent leur utilité lorsqu'il s'agit de modéliser des systèmes complexes dont les processus locaux sont régis par des règles locales simples à l'instar du type de règles que nous proposons, dans le contexte de cette thèse, qui se base sur l'historique des changements d'occupation/usage du sol des entités géographiques étudiées et les configurations spatiales dans lesquelles ils se situent.

Les autres approches qui ont été proposées dans la littérature comprennent la modélisation à base d'automates cellulaires (ACs), à base d'agent et à base de fractales. [Batty, 2007], considère dans sa propre revue de littérature, que ceux-ci représentent les principales approches pour donner un aperçu de la nature dynamique de la structure d'un territoire (e.g. un territoire urbain) à travers les motifs spatiaux [2] (*spatial pattern*) complexes qu'ils fournissent. En fait, il suppose que les systèmes simulés (e.g. les villes) ne doivent pas être perçues comme une structure spatiale holistique mais plutôt comme des couches de changement urbain, qui pourraient être illustrées à

1. Est un processus autonome qui accomplit une tâche en fonction de recommandations spécifiées par son auteur.
2. Une structure perceptuelle, placement ou disposition d'objets sur la Terre.

l'aide d'un automate cellulaire, ou comme des entités présentant des comportements autonomes pouvant être représenté par des agents. Les approches à base d'agents et celles à base d'automates cellulaires sont évoquées presque dans tous les travaux, car ils traitent des comportements complexes et non linéaires des systèmes étudiés, souvent causés par la combinaison de la dynamique temporelle et spatiale.

[Silva and Wu, 2012] présente une étude plus récente et détaillée qui propose d'autres catégories de modèles (selon l'approche utilisée), à savoir, les modèles mathématiques et statistiques, les modèles à base de SIG et les modèles à base de règles et les modèles hybrides. Il convient de noter que de plus en plus d'études [Lambin et al., 2000, Silva and Wu, 2012, Poelmans and Rompaey, 2010] mettent en évidence l'importance des approches hybrides puisqu'elles sont le résultat de combinaisons d'éléments issus de différentes techniques et donc elles bénéficient de leurs avantages.

Dans le présent chapitre, nous avons choisi d'identifier les approches de modélisation à travers la classification proposée par [Silva and Wu, 2012]. Leur travail a commencé par présenter un aperçu des principales classifications, puis a tenté d'étendre celles-ci en englobant des modèles et des approches de modélisation plus récents. En tout, soixante-quatre modèles sont étudiés et ils ont proposé six catégories de modèles impliquant presque toutes les différentes catégories proposées dans la littérature (approches basées sur les agents, approches basées sur les automates cellulaires, approches mathématiques/statistiques, approches basées sur les règles, approches basées sur les SIG et approches hybrides).

3.4 Approches de modélisation

3.4.1 Modélisation traditionnelle statique : mathématique et statistique

Bien que tous les modèles puissent impliquer des mécanismes mathématiques, les modèles mathématiques sont définis comme ceux se basant essentiellement sur des équations mathématiques pour résoudre un problème de modélisation. Le concept de base de cette catégorie de modèles consiste à assumer son état initial d'équilibre, à la fois dans le temps et dans l'espace, puis, en réponse à un stimulus exogène, évoluer vers un nouvel équilibre [Waddell and Ulfarsson, 2004]. Les propriétés des systèmes modélisés

sont décrites à l'aide de variables et les équations mathématiques décrivent comment une ou plusieurs variables sont liées ensemble.

Représentant la première génération de modèles urbains, cette catégorie de modèles a été introduite lors des années cinquante. Ils étaient principalement statistiques et déterministes et se sont concentrés, essentiellement, sur les problèmes d'allocation des transports et d'occupation du sol, d'où le nom « *land-use transport models* » (LUT). Les modèles urbains mathématiques les plus simples ont consisté en un ensemble d'équations décrivant la croissance démographique et la redistribution pour indiquer le changement d'occupation du sol dans le temps. Depuis les années 1960, de nombreux modèles LUT ont été proposés et certains d'entre eux ont, en effet, assisté à des tâches de plannification urbaine [Klosterman, 1994]. Parmi les exemples les plus connus, citons les modèles économiques urbains consistant, principalement, en des modèles économiques régionaux et des modèles de marché foncier. Ces modèles reposent souvent sur des théories basées sur la micro-économie telles que la théorie de la location [3] [Alonso, 1964]. Selon le modèle de von Thunen [4], où la location et le loyer des terrains sont supposés être fonction de la distance ou des coûts de déplacements à certains centres de marché, la fonction de loyer des offres explique la relation entre l'usage du sol et les valeurs foncières. D'autres exemples de modèles mathématiques sont : les modèles basés sur la physique sociale [5] (les modèles d'interaction spatiale), où les activités sont réparties selon des hypothèses gravitationnelles ; et des modèles basés sur des techniques statistiques telles que la régression. Ces techniques ont été communément utilisées pour traiter la prise de décision et la modélisation des phénomènes sociaux dans le cadre de la modélisation du changement de l'occupation du sol [Mertens and Lambin, 1997]. Les modèles urbains mathématiques ont évolué lentement pour traiter le temps. En outre, ils sont également devenus plus désagrégés et plus liés à l'occupation physique du sol, bien qu'ils restent toujours au niveau de l'allocation d'activité malgré leur nomenclature comme LUT. Quelques-uns des exemples les plus développés sont : UrbanSim, MEPLAN [Hunt and Echenique, 1993], TLUMIP [Weidner and Hunt, 2006] et DELTA [Simmonds, 1999]. Les modèles

3. Appelé en anglais « *bid rent theory* ». C'est une théorie économique géographique qui se réfère à la façon dont le prix et la demande de biens immobiliers changent à mesure que la distance du quartier d'affaires central (CBD) augmente.

4. $https://is.mendelu.cz/eknihovna/opory/zobraz_cast.pl?cast = 62123$

5. est un domaine de la science qui utilise l' analyse des données et les lois mathématiques de la biologie pour comprendre le comportement des foules humaines.

mathématiques urbains ont été critiqués par de nombreux chercheurs comme n'ayant presque rien à voir avec les structures spatiales [Jr and B, 1973]. Ces modèles ont été conçus pour un monde plus simple, statique et plus certain. Cependant, les systèmes du monde réel sont complexes, dynamiques et composés de nombreuses composantes interdépendantes et en traduisant la réalité en équations mathématiques, certaines de ces composantes sont omises. Les principaux avantages de la présente catégorie de modèle sont: leur capacité à faire des déclarations précises sur les processus comportementaux tels que l'évolution des villes par le changement d'occupation/usage du sol ou la croissance urbaine, leur testabilité [Mazur, 1987] et la capacité des techniques, comme la régression logistique, à atteindre une prévisibilité suffisante, même avec une quantité limitée d'informations [Mertens and Lambin, 1997].

3.4.2 Modélisation dynamique

3.4.2.1 Modélisation à base d'automates cellulaires

Dans le domaine de la modélisation urbaine, un modèle à base d'automates cellulaires (AC) repose sur un ensemble de zones physiques d'une grille appelé cellules. Chacune d'elle est caractérisée par un état (e.g. utilisation du sol) qui change, de manière synchrone en pas de temps discret, conformément à un ensemble de règles de transition. Ces fonctions de changement d'état (règles) spécifient, pour chaque état possible du voisinage, un certain état de la cellule considérée.

Cette approche a été largement utilisée dans les modèles urbains principalement grâce à certaines de ses caractéristiques : la simplicité de construction, la spatialité, la décentralisation, la flexibilité (flexible à formuler), la dynamique et la capacité d'être intégrée avec d'autres outils d'analyse spatiale. La capacité à générer des formes spatiales extrêmement complexes, basées sur des règles locales simples, représente un avantage clé des modèles à base d'AC. En outre, ces modèles offrent une vue topologique de l'espace, grâce à leur composante fondamentale – le voisinage –, et sont considérés compatibles avec la plupart des jeux de données spatiales comme il est défini sur un espace considéré comme une grille de cellules (raster). En AC, les changements urbains simulés sont généralement visualisés à travers les transitions des états cellulaires (e.g. de non urbain à urbain), ce qui améliore la lisibilité et la compréhensibilité des résultats de la simulation.

Cependant, parallèlement à ces avantages, plusieurs limitations ont été

rapportées. Par exemple, on a accusé la structure originale d'AC d'être trop simpliste et restreinte pour être appliquée dans des applications urbaines réelles [Sipper, 2001]. En effet, les villes, par exemple, ne peuvent pas être infinies, uniformes ou régulières, c'est pourquoi il est aberrant d'appliquer une représentation par un plan spatial infini (bidimensionnel) et d'un espace régulier uniforme, l'idée centrale du concept de AC, à des systèmes tel que la ville. Une autre critique est que AC se concentre sur leurs cellules de composition et souvent néglige le lien entre elles et les motifs globaux qui peuvent être capturés sur de plus grandes échelles spatiales [Qi et al., 2004]. Un tel problème a été abordé en introduisant des contraintes exogènes sur certains modèles d'expansion urbaine qui sont régis par ces facteurs macro-économiques sociaux et économiques [White and Engelen, 1997, White and Engelen, 2000, Ward et al., 2000].

3.4.2.2 Modélisation à base d'agents

Un modèle à base d'agents (MBA) est défini par un ensemble d'agents interagissant au sein d'un environnement simulé. Dans les applications du monde réel, un agent peut représenter une grande variété d'entités telles que les atomes, les cellules biologiques, les personnes, les organisations, les bâtiments ou les parcelles [Conte et al., 2013, Epstein and Axtell, 1996, Janssen and Jager, 2000, Weiss, 1999]. Un agent est, essentiellement, caractérisé par son autonomie, son adaptabilité, et sa capacité à évaluer sa situation et à agir sur lui-même et sur son environnement. Il prend des décisions liées à son comportement, ses relations et interactions avec ses congénères. Ces règles – pouvant s'agir d'équations mathématiques, de règles si/alors ou d'opérations logiques – sont généralement dérivées de l'observation, des connaissances d'experts et de l'analyse des données. Tous les agents peuvent partager une règle ou chaque agent peut avoir sa propre règle unique. Les règles ne sont pas toujours prédéfinies – elles peuvent être évolutives comme les agents sont dotés de capacités d'apprentissage.

Les MBA présentent plusieurs avantages. Ils sont facilement éditables. Ils permettent un réglage de la complexité des agents en affectant leurs comportements, leur niveau d'agrégation (e.g. groupe et sous-groupes d'agents, des agents individuels, coexistence de différents niveaux d'agrégation), leur capacité d'apprentissage et d'évolution, et leurs règles d'interactions. En effet, la composante règles rend le processus de modélisation beaucoup plus flexible comme l'édition, la suppression ou l'ajout d'une règle se produit, gé-

néralement, indépendamment des autres règles définies. Ces modèles sont, de plus, extensibles (ajouter des agents) et contractiles (enlever les agents). Outre les avantages susmentionnés, les MBA sont capables de saisir des propriétés émergentes résultant de l'interaction et de l'aptitude à évoluer de leurs propres entités. Ce type de propriétés ne sont pas prévisibles par observations à micro-échelle (observation d'entités isolées) mais peuvent, par contre, être capturés à l'échelle macro [Bonabeau, 2002]. En mettant l'accent sur les comportements des individus, les MBAs offrent une modélisation plus réaliste, plus pratique, plus fiable et plus complète. En effet, plus les règles de décision des agents sont définies de sorte à ce qu'elles ressemblent à des décisions humaines, meilleures sont les résultats de la modélisation.

Cependant, cette caractéristique de simulation à niveau micro peut se transformer en une limitation en raison de la consommation élevée de ressources lors de la modélisation de grands systèmes réels et du nombre énorme d'agents et de facteurs d'interaction à prendre en compte. Ainsi, la plupart des modèles MBA actuels ne sont capables de simuler que des paysages hypothétiques très simplifiés [Kanaroglou and Scott, 2002]. Une autre limitation des MBA est que leur dynamique est si complexe que même leurs propres développeurs peuvent ne pas comprendre comment ils sont générés ce qui rend difficile pour eux de détecter les éventuelles erreurs ou artefacts dans la conception ou dans la mise en oeuvre du modèle [Galán et al., 2009].

3.4.2.3 Modélisation à base de systèmes d'information géographique

De nos jours, la gestion des bases de données, l'analyse spatiale, la modélisation spatiale et la visualisation sont les principales utilisations des SIG actuels dans les tâches de modélisation urbaine telles que la planification [Yeh, 1999]. Les données dans un SIG sont structurées sous la forme de cartes numériques organisées l'une sur l'autre selon un référentiel spatial commun utilisé pour localiser des entités géographiques. Ces cartes ou couches sont liées à des données tabulaires supplémentaires avec des informations, essentiellement, sur les objets géographiques inclus. Comme outil de gestion des données, le SIG est utilisé pour le stockage des données spatiales et temporelles (cartes d'utilisation des sols, plans, données socio-économiques et environnementales) liées à l'aide d'un modèle géo-relationnel [6] (georelational

6. Est un modèle de données géographiques représentant les caractéristiques géographiques par un ensemble interdépendant de données spatiales et d'attributs.

model).

Les SIG aident les utilisateurs à interroger ces données et à récupérer des informations utiles afin d'analyser la situation existante de la zone géographique adressée. Les SIG actuels offrent aux planificateurs la possibilité d'exporter des informations géographiques à partir de leurs bases de données vers d'autres programmes de modélisation et d'analyse spatiale et de les combiner à d'autres bases de données tabulaires ou à des enquêtes spécialement conduites pour prendre des décisions de planification efficaces. En tant que boîte à outils, le SIG permet aux planificateurs d'effectuer divers analyses spatiales à l'aide de certaines fonctions de géo-traitement telles que l'interpolation, la mesure de la connectivité, la mise en mémoire tampon (*buffering*) et la superposition de cartes. En effet, cette dernière est probablement la fonction la plus utile car elle est traditionnellement utilisée par les praticiens dans l'analyse de l'aptitude des terres qui représente une composante importante pour certains exercices de modélisation urbaine tels que la planification [Chandio et al., 2011, Şener et al., 2010]. Comme exemples concrets d'utilisation, les outils de géo-traitement, dans les SIG, aident à identifier les zones présentant des conflits de développement urbain et de préservation de l'environnement en superposant les aménagements existants sur les cartes d'aptitude des terres (*land suitability*). Les domaines de préoccupations environnementales peuvent être identifiés à partir de l'analyse des données (informations géographiques et données socio-économiques et environnementales) stockées dans les SIG. Par exemple, les scénarios environnementaux sont étudiés à travers : d'abord, de la projection de la demande future de ressources foncières à partir de la population et des activités économiques, par la suite, la modélisation de la distribution spatiale de cette demande et puis l'utilisation de l'analyse de superposition des cartes SIG pour identifier les zones de conflit. Les travaux de [Yadav et al., 2012] représentent des exemples révélateurs de ce type d'applications. Les auteurs utilisent des outils géo-spatiaux (e.g. télédétection et SIG) pour détecter et représenter les changements d'occupation/couverture des sols, analyser leur conflit avec l'habitat faunique et évaluer, ensuite, son impact sur la biodiversité. En superposant une carte de développement des terres produite à partir de l'analyse des images de télédétection sur le plan d'utilisation des terres, le SIG offre également la possibilité d'examiner si le développement territorial suit le plan d'aménagement du territoire de la région. Outre les analyses d'aptitude des terres, l'évaluation de l'impact et la surveillance de la disponibilité et du développement des terres, l'analyse spatiale et statistique fournie par le SIG peut

être utilisée pour la sélection des sites, l'identification des zones d'action pour la planification, la modélisation du transport terrestre et d'autres pratiques urbaines [Stillwell et al., 2013].

La visualisation des données représente l'un des atouts les plus importants des modèles urbains basés sur le SIG. En effet, le SIG aide les planificateurs à explorer les données (par exemple, la répartition des données socio-économiques et environnementales), calibrer et afficher les résultats des exercices de modélisation tels que la planification urbaine et l'analyse spatiale, [Gu et al., 2009]. Les outils de modélisation basés sur les SIG sont dotés de capacités de cartographie thématique et d'affichage graphique des résultats de la modélisation pour une meilleure interaction homme/machine et également pour un processus de prise de décision amélioré. De plus, la possibilité d'incorporer des matériels multimédias tels que des vidéos, des photographies aériennes, des images, la réalité virtuelle et des sons dans la modélisation basée sur les SIG augmente considérablement la compréhension des praticiens du problème de la modélisation urbaine qu'ils analysent. En effet, les images de time series issues de la télédétection et des satellites représentent une grande source d'informations géographiques car elles aident, par exemple, à détecter les changements d'occupation des sols pour des zones urbaines spécifiques.

Malgré leurs différentes forces, les outils de modélisation basés sur le SIG restent insuffisants lorsqu'il s'agit de traiter l'aspect dynamique du changement urbain. Le mécanisme causal associé aux changements d'occupation/couverture du sol reste mal compris car ils sont limités à une représentation statique de l'état passé et actuel des systèmes modélisés, à travers des cartes et des images satellites, sans fournir aucune compréhension des raisons ou de l'aperçu des perspectives futures [Hopkins, 1999]. Ainsi, les SIG ont été combinés avec d'autres approches capables de saisir la complexité et la dynamique des comportements des systèmes urbains (e.g. la fouille de données). Cela fournit des outils plus puissants capables d'effectuer une analyse spatiale améliorée avec la possibilité de tracer des tendances et des projections.

3.4.2.4 Modélisation à base de règles

Dans le cadre de la modélisation des changements d'usage/occupation du sol, les modèles à base de règles sont utilisés pour déterminer, au moyen d'un ensemble de règles prédéfinies, où une certaine occupation ou usage du sol est susceptible de se produire. La littérature a fait état de diffé-

rentes perspectives et définitions de l'approche de modélisation fondée sur des règles. Selon [Silva and Wu, 2012], les modèles basés sur des règles sont des systèmes experts incorporant des règles de décision explicites, qui permettent aux utilisateurs du modèle (experts) de spécifier comment celui-ci se comportera. Cependant, un aperçu de la littérature montre que même les modèles basés sur les automates cellulaires, les modèles basés sur les agents [Mitasova and Mitas, 1998] et les modèles basés sur certains techniques de fouille de données [Tayyebi, 2013] peuvent être considérés comme des modèles à base de règles car ils s'appuient sur un ensemble de règles locales dans leur processus de modélisation. Les techniques de datamining représentent une méthode révolutionnaire pour extraire des règles plus informatives et significatives que même les experts peuvent omettre en raison de la grande quantité de données et de variables de décision disponibles. Cependant, ces modèles sont a-spatiaux (c'est-à-dire manquent de capacités dans la représentation spatiale et la visualisation spatiale des résultats). Par conséquent, ils sont généralement combinés avec des outils tels que les AC et les SIG. Par exemple, [Liu et al., 2007] proposent d'utiliser un algorithme de réseau de neurones qui, en se basant sur des données historiques urbaines, vise à explorer les motifs d'étalement urbain afin de le prédire. Le SIG, dans ce modèle, est principalement utilisé pour l'importation et la gestion de données qui alimenteront l'algorithme de réseau de neurones – Le modèle a besoin d'au moins deux cartes géographiques correspondant à deux dates différentes à partir desquelles le SIG extrait des caractéristiques spatiales (par exemple l'utilisation des topographies), puis représente chaque cellule de la zone étudiée par un vecteur constitué de ces caractéristiques spatiales et d'un label de classe (*i.e.* urbaine ou non urbaine). Dans un autre exemple, [Gharbi et al., 2014], ont utilisé la recherche de règles d'association pour extraire des règles d'évolution à partir de l'historique des changements du bâti dans la ville de Saint-Denis. Ensuite, ils ont employé ces règles ainsi que certaines fonctions de SIG pour générer et visualiser la future carte de la ville.

La principale force des modèles basés sur des règles est que, une fois que la base de règles est définie, les règles peuvent être gérées par des non-experts en programmation informatique. En effet, l'utilisateur peut facilement les comprendre, les examiner et même les modifier, ce qui assure un développement rapide du modèle et facilite son entretien. Cependant, les règles produites (et non les approches) sont, en général, très détaillées et spécifiques à une application quelconque ce qui rend difficile leur réutilisation même pour des

problèmes similaires avec des données différentes.

3.5 Discussions

Malgré leurs points forts, chacune des approches présentées ci-dessus souffre de certaines faiblesses, lorsqu'il s'agit des différents aspects de la modélisation du changement d'occupation/usage du sol, à savoir le manque de flexibilité dans les modèles mathématiques basés sur les équations, la modélisation statique dans les modèles basés sur les SIG, le manque de représentation des facteurs a-spatiaux (comportement des individus, variables socio-économiques) dans les modèles AC [White and Engelen, 1997, White and Engelen, 2000, Ward et al., 2000], etc. Il y a donc eu de plus en plus besoin de modèles plus sophistiqués et intégrés qui atténuent l'effet de ces limitations et permettent : de mieux répondre aux caractéristiques multiscalaires des systèmes urbains fonciers urbain, de mettre en oeuvre de nouvelles techniques de quantification des effets de voisinage – qui traitent explicitement de la dynamique temporelle et spatiale et qui permettent d'atteindre un niveau d'intégration plus élevé entre les approches disciplinaires et les changements dans les terres urbaines – et de fournir des solutions théoriques qui peuvent être pratiquement appliquées aux problèmes du monde réel.

Même si de nombreux modèles ont été publiés avec succès et ont théoriquement démontré de bonnes performances [Berling-Wolff and Wu, 2004, Briassoulis, 2000] un large écart subsiste entre l'analyse théorique d'une cité et les applications et le développement pratiques d'un autre côté. En effet, quelques problèmes subsistent avec la vérification et la précision des résultats de la modélisation. Les mécanismes de validation et d'évaluation, la transférabilité et la possibilité de réutilisation représentent encore, pour les spécialistes, un fossé à combler dans l'avenir. Pour obtenir une fiabilité suffisante des résultats de la simulation, il convient d'ajouter certaines considérations à la proposition de modèles, à leur conception et aux procédures de leur évaluation. Premièrement, comprendre les modèles urbains actuels, les approches de modélisation populaires et leurs caractéristiques et applications. Deuxièmement, il est nécessaire de comprendre les dynamiques et les processus des changements d'occupation/usage du sol dans différentes échelles temporelles et spatiales et d'étudier les principaux acteurs du modèle, de la collecte des données et des processus décisionnels. Troisièmement, il est important non seulement de construire le cadre du modèle mais aussi de le

traduire en un outil de simulation flexible et portable pouvant être intégré au modèle à différents niveaux / échelles / délais.

Quatrièmement, la performance du modèle, les tests de validation, l'étalonnage et l'évaluation du modèle doivent tous être évalués. Enfin, il existe un besoin continu d'intégrer de nouvelles techniques de multiples domaines disciplinaires et de proposer des approches de modélisation hybrides. Dans ce contexte, on peut citer le modèle proposé par [Silva et al., 2008] qui requiert la contribution des géographes (concernant la croissance urbaine et le changement d'occupation/usage du sol), les planificateurs (afin d'explorer des questions de politique et de pratique d'aménagement – les objectifs de la politique consistant à établir des corridors de voies vertes et les moyens pratiques de le faire en milieu urbain) et les écologistes du paysage (nécessité d'une bonne compréhension de la taille et de la proximité des parcelles forestières / Connectivité). Dans un autre exemple de modèles intégrés, le DG-ABC [Silva and Wu, 2014], les dynamiques spatiales et a-spatiales ont été considérées. De toute évidence, l'analyse et la modélisation de ces dynamiques requièrent une compréhension des questions spatiales telles que l'information géographique, mais aussi de l'économie et des théories et pratiques humaines et sociales. Du point de vue de l'approche de modélisation, le DG-ABC est un modèle hybride qui emploie des agents intelligents pour prendre en compte les influences sociales et économiques à travers les comportements des individus (résidents : capacité financière, structure familiale, plan d'infrastructure, zones exclues, etc.) et l'automate cellulaire pour explorer, analyser et représenter les effets de certains facteurs d'influence spatiale tels que l'infrastructure.

3.6 Conclusion

Une analyse de la littérature a fait ressortir cinq principales approches informatisées pour la modélisation des phénomènes spatio-temporels à savoir : l'approche statistique et mathématique, l'approche à base de GIS, l'approche basé sur l'automate cellulaire, l'approche à base d'agents et l'approche à base règles. C'est dans la dernière catégorie que notre travail de recherche s'inscrit. En effet, nous partons de l'hypothèse que les dynamiques spatiales et les usages des objets géographiques peuvent, en partie, être anticipés par leurs historiques de changements de fonctions et de co-localisations. Ainsi nous proposons d'exploiter la fouille de données, notamment, la recherche

des motifs fréquents et des règles d'associations, pour en extraire des règles d'évolution régissant ces dynamiques. Ces méthodes d'extraction de connaissances seront, plus amplement, abordées dans le chapitre suivant.

Chapitre 4

Recherche de règles d'association

4.1 Introduction

Dans le domaine de la géographie, l'espace est défini comme une étendue ou une surface terrestre pouvant être occupée ou vide [Clarck et clarck 1993]. En effet, les géographes différencient entre l'espace concret constitué d'un ensemble de lieux de faible étendue et l'espace formel [Pierre and Verger, 1970a] qui correspond à des objets ou entités et les relations entre elles ainsi qu'à des mesures, des qualifications et des représentations cartographiques. De ce fait, de nombreux travaux de géomatique considèrent que l'évolution de cet espace, est, en grande partie, le produit d'échanges et de rapports existants entre ses entités constituantes [Mondo, 2011, Perret et al., 2015, Mathian and Sanders, 2015]. Certain géographes, tels que Tobler [Tobler, 1970] ont même allé jusqu'à mettre l'accent sur l'effet de la proximité sur la pertinence de ces rapports quant à la modélisation des dynamiques spatio-temporelles de ces entités.

> « Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés. » (Tobler, A Computer Movie Simulating Urban Growth in the Detroit Region, p.236)

C'est cette notion de voisinage spatial ou de proximité (explicitée dans le chapitre2) qui est au coeur de notre approche de modélisation. En effet, dans cette thèse nous nous focalisons sur l'évolution territoriale en termes de

fonctions des entités, connues dans le jargon des géographes, sous le terme par exemple d'occupation du sol ou d'usage du sol. Nous partons, de l'hypothèse que l'évolution de ces entités peut en partie être explicitée, voire anticipées, par l'historique des mutations précédentes et les configurations spatiales dans lesquelles elles se situent. Notre objectif est, donc, de déterminer un modèle prospectif se basant essentiellement sur deux types de relations entre les entités étudiées : les relations temporelles de changement de fonction et les relations spatiale de voisinage.

Plusieurs méthodes de traitement et d'analyse de données ont rendu possible la découverte de ce genre de modèle. La fouille de données, ou le *datamining* en anglais, est l'une des techniques les plus populaires dont l'objectif est d'extraire des connaissances cachées et non-triviales à partir d'une grande quantité de données. C'est à partir de ces connaissances que des modèles à vocations explicative et/ou prédictive peuvent être définis.

Dans ce chapitre, nous commençons par présenter un aperçu général sur la fouille de données, ses méthodes et ses applications puis nous mettons l'accent sur la méthode de fouille de motif fréquent. C'est une méthode qui consiste à découvrir les corrélations et des règles entre les variables caractérisant les données d'étude ce qui correspond parfaitement à notre objectif de découvrir la corrélation entre relations spatio-temporelles et évolution et d'obtenir des règles explicatives.

4.2 Datamining

Apparu au milieu dans les années 1990 aux États Unis, le datamining ou la fouille de données, constitue une branche se situant à l'interface de plusieurs disciplines telles que la statistique, les bases de données, l'apprentissage automatique et l'intelligence artificielle. Communément définie comme l'exploration de connaissances non triviales utiles et cachées dans une énorme quantité de données [Hand, 1998, Fayyad et al., 1996b], la fouille de données est souvent confondue avec l'extraction de connaissances dans les bases de données (ECD) appelé en anglais *Knowledge Discovery in Database* (KDD) [Fayyad et al., 1996a].

La fouille de données est l'une des cinq étapes que subit une base de données lors du processus de l'ECD, à savoir : la sélection, le pré-traitement, la transformation, la fouille de données et l'évaluation des résultats (cf. figure 4.1). Bien qu'elle constitue l'étape centrale de ce processus, ses résultats

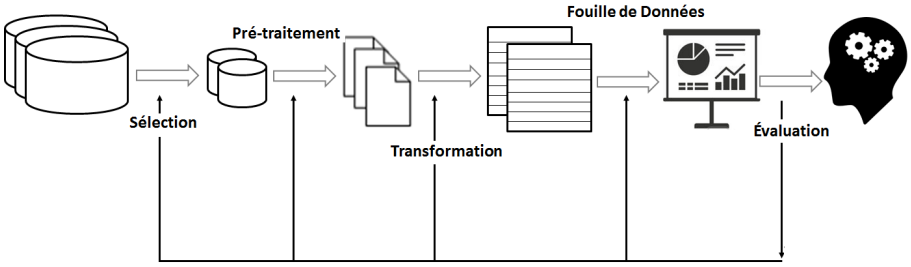


FIGURE 4.1 – Le processus d'extraction de connaissance [Fayyad et al., 1996a]

doivent être interprétables et exploitables par un utilisateur qui est souvent l'expert d'un domaine. Fayyad et al. [Fayyad et al., 1996a] insistent notamment sur l'aspect interactif et itératif du processus. Il considère que c'est à l'utilisateur d'appliquer les différentes étapes de l'ECD et d'ajuster itérativement ses questions en fonction de la réponse renvoyée par le système de fouille de données. En liaison avec notre travail, le processus d'extraction de connaissances s'illustre par l'extraction des règles d'évolution territoriale (chapitre 5) et l'évaluation de leur pertinence une fois extraites (chapitre 6).

Egalement, Fayyad et al. [Fayyad et al., 1996a] mettent l'accent sur l'importance des étapes de pré-fouille et post-fouille pour, respectivement filtrer, compléter et transformer les données, afin de permettre une utilisation optimale de la méthode de fouille de données ; et pour sélectionner visualiser et interpréter les résultats de fouille afin d'en assurer une utilité maximale.

Faisant appel, à la fois, aux statistiques et aux principes de l'apprentissage artificiel, les méthodes de fouille de données visent à définir des modèles qui : à partir d'exemples de solution, assurent une bonne représentativité des données ; sont dotés d'une robustesse aux erreurs et sont également capables de gérer le passage à l'échelle [1]. Nombreuses sont les méthodes de fouille de données qui ont été proposées. Elles varient, principalement, selon le type et la quantité de données étudiées et les problèmes auxquels elles tentent de répondre. Ainsi, nous distinguons les méthodes de régression statistique, les méthodes de groupement ou de *clustering* (*i.e.* les méthodes des plus proches

1. L'aptitude d'un système à résoudre un problème pour lequel il est conçu même avec une plus grande quantité de données.

voisins [Cover and Hart, 2006]), ou de classification supervisée comme les arbres de décision [Breiman et al., 1984, Quinlan, 1986]. Outre les problèmes classiques, traités jusqu'alors par l'apprentissage artificiel, la fouille de données propose des méthodes rattachées à d'autres types de problèmes tels que la recherche de motifs fréquents.

4.3 Recherche de motifs fréquents et l'extraction de règles d'association

Représentant une innovation dans le domaine du datamining, ce problème fut introduit, en 1993, par Agrawal et al. [Agrawal et al., 1993] et résolu par leur algorithme Apriori. Contrairement aux autres problèmes il ne s'agit pas de classifier ou de grouper les données ni de fournir une description globale de celles-ci comme pour les méthodes de régression. En revanche, il s'agit de découvrir des motifs (conjonctions, relations, structures), assez fréquemment présents dans la base de données et d'en définir des règles d'association.

Mis à part sa capacité à capturer les relations et structures incorporées dans les données d'apprentissage, la recherche de motifs fréquents présente plusieurs atouts tels que son exhaustivité en termes de génération de règles ou de motifs fréquents, sa simplicité, son intuitivité et sa capacité à gérer le problème de démarrage à froid posant souvent problème aux autres méthodes de la fouille de données (e.g. classification). En effet, l'utilisation d'un seuil de fréquence (support) garantit la capture des associations tant que l'on dispose d'une quantité suffisante de données même si elle est restreinte. La recherche de motifs fréquents a également évoluée, au fil des années, en terme de performance (temps de réponse et mémoire) et en termes des types de données gérés – spatiales [Sarangi and Sahoo, 2013] , temporelles (e.g. cycliques, séquentielles)[Miao and Shen, 2010] , quantitatives [Gosain and Bhugra, 2013] et floues [Khan et al., 2008]– pour donc produire des règles plus complexes – règles spatio-temporelles[Xuewu et al., 2008], règles multi-niveau et multi-dimensionnelles [Han et al., 2012], floues [Farzanyar and Kangavari, 2012]– et traiter d'autres types de problèmes – classifications, partitionnement, détection d'anomalies, détection d'évènement [Han et al., 2007].

$\mathcal{R}$	i_1	i_2	i_3	i_4
tr_1	×	×	×	×
tr_2	×	×	×	
tr_3			×	×
tr_4	×		×	
tr_5			×	

FIGURE 4.2 – Un exemple de représentation d'une relation binaire sous forme de table

4.3.1 Problème de recherche de motifs fréquents

Dans un formalisme initial, la recherche de motifs fréquents implique un ensemble T composé de n transactions ($T = \{tr_1, tr_2, tr_3, ..., tr_n\}$) et un ensemble I (les attributs) de m items $I = \{i_1, i_2, i_3, ..., i_m\}$ tel que $tr_i \subseteq I$, reliée par la relation binaire $R \subseteq T \times I$. La table de la figure 4.2 constitue un exemple de cette relation binaire. Les colonnes représentent les items ou attributs et les lignes représentent les transactions. Une case est cochée si l'attribut et la transaction qui lui correspond sont en relation, en d'autres termes, l'item figure dans la transaction.

L'objectif de la recherche des motifs fréquents est de déterminer les co-occurrences d'attributs qui apparaissent le plus fréquemment dans les données.

Définitions

Soit les définitions suivantes :

— Un motif M est un ensemble d'items

— Un motif M tels que $M \subseteq I$ couvre une transaction de T si cette transaction est en relation par R avec tous les items de M.

— Le support d'un motif M, noté *supp(M)* et appelé aussi « fréquence absolue », représente le nombre de transaction couvertes par M (*i.e.* nombre de transaction contenant tous les éléments de M) $|tr \in T, \forall i \in M, tr \ni i|$.

— La fréquence relative d'un motif M, notée $Freq_R(M)$, représente la fraction de la fréquence absolue de M sur le nombre n de transactions et est donc une valeur comprise entre 0 et 1.

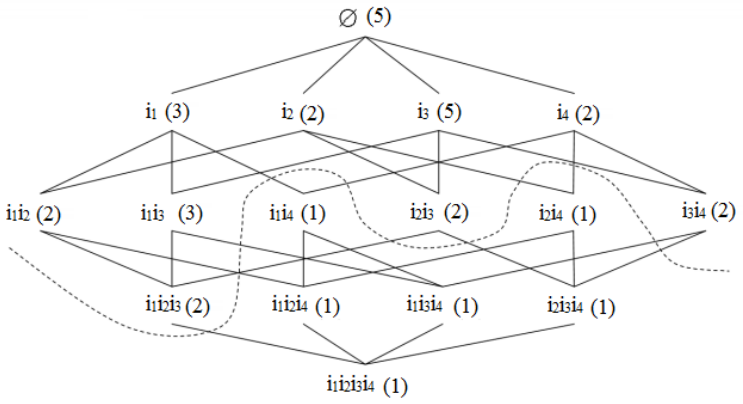


FIGURE 4.3 – Treillis des itemsets correspondant à l'exemple de la figure 4.2.

Tenant compte des définitions ci-dessus, le problème de recherche de motifs fréquents revient à trouver les motifs dont la fréquence est supérieure ou égale à un seuil de fréquence prédéfinie par l'utilisateur. Le seuil peut s'agir d'une fréquence absolue minimale *(minSup)* ou d'une fréquence relative minimale ($minFreq_R$). Dans l'exemple de la figure 4.2 le motif $\{i_2 i_3\}$ couvre les transactions *tr1* et *tr2* et donc a pour support $|\{tr1, tr2\}| = 2$ et pour fréquence relative 2/5. Sur le Treillis (cf. figure 4.3) des motifs correspondant à notre exemple on remarque que la fréquence est une fonction décroissante de l'ensemble de motifs ordonnés par la relation d'inclusion. En effet, une fonction d'un ensemble E ordonné dans un ensemble E' ordonné est dite décroissante si l'application de deux éléments ordonnés dans l'ensemble de départ donne deux éléments inversement ordonnés dans l'ensemble d'arrivée. Dans notre exemple la fréquence est décroissante comme pour $M1 \subseteq M2 \subseteq A$, $Freq_R(M1) \geq Freq_R(M2)$).

Cette propriété de décroissance implique la propriété d'anti-monotonie sur la fréquence des motifs, autrement dit, si un motif M n'est pas fréquent aucun de ses sur-ensembles ne l'est. La contrainte de fréquence notée F est dite anti-monotone si et seulement si pour chaque motif M, si M ne satisfait pas F alors aucun de ses sur-ensembles ne satisfait F.

4.3.2 Extraction de règles d'association fréquentes

Typiquement, une règle d'association représente l'implication de la forme suivante : $X \longrightarrow Y$. X et Y sont deux ensembles d'items, appelés *itemsets* ; $X, Y \subset I$; et $X \cap Y = \emptyset$. Il s'agit d'une règle probabiliste et non déterministe comme elle indique, généralement, la probabilité conditionnelle pour qu'une transaction présentant sa prémisse (le motif X) présente aussi sa conclusion (le motif Y). C'est ce qu'on appelle la confiance *(conf)* qui plus elle est importante plus la règle est intéressante. Représentant intuitivement la relation de causalité ou de corrélation entre sa prémisse et sa conclusion, les experts considèrent que la règle d'association est plus expressive et utile qu'un simple motif fréquent.

4.3.2.1 Définitions

Soit la règle $R : X \longrightarrow Y$.

— La fonction $card(X)$, où X est un itemset, correspond au nombre de transactions de la base qui incluent X.

— Le support d'une règle évalue sa portée dans la base de données. Il est défini par la proportion de transactions de la base qui contiennent à la fois X et Y $(X \cup Y)$:

$$supp(R) = \frac{card(X \cup Y)}{n}$$

n est le nombre de transaction de la base. Il s'agit de la probabilité $Pr(X \times Y)$.

— La confiance d'une règle (e.g. R) représente la proportion de transactions de la base contenant X qui contiennent aussi Y :

$$conf(R) = \frac{card(X \cup Y)}{card(X)}.$$

Il s'agit aussi de la probabilité $Pr(Y|X)$.

Ainsi, considérons toujours l'exemple de la figure 4.2, le support de la règle $R2 : i_1 i_2 i_3$ est 2 et sa confiance est 2/3.

L'objectif de la recherche de règle d'association consiste alors à trouver, à partir des données, un ensemble de règles confiantes (dont les confiances sont élevées) et représentatives (ayant des supports élevés). En effet, étant donnés des seuils minimums de confiance (*minconf*) et de support (*minsup*)

prédéterminés par l'utilisateur ou l'expert, il suffit de générer l'ensemble des itemsets dont la fréquence est supérieure au (*minsup*, et sur ces itemsets fréquents de construire les règles pour sélectionner celles dont la confiance est supérieure ou égale à *minconf*.

4.3.2.2 Algorithmes de recherche de motifs fréquents

4.3.2.2.1 Apriori Lors de l'introduction du problème de recherche de motif fréquent, un algorithme appelé Apriori fut proposé comme une solution. Populaire et très couramment utilisé, cet algorithme adopte une heuristique qui se base sur la connaissance a priori de l'information sur la fréquence des items. La fréquence d'un item ou d'un itemset (*i.e.* un motif constitué d'un ensemble d'items) est évaluée à travers son nombre d'occurrences dans la base (support). Un item ou un itemset est dit fréquent si son support dépasse un seuil, spécifié par l'utilisateur : le *minsup*. Afin d'identifier les itemsets fréquents pour en construire les règles d'association, l'algorithme effectue plusieurs balayages de la base d'apprentissage (cf. algorithme 1, et figure 4.4). En effet, il procède en deux étapes :

— Une étape de génération d'itemsets fréquents utilisant une fonction de jointure pour générer des itemsets candidats et une fonction d'élagage pour ne garder que les candidats fréquents après avoir évalué leurs fréquence dans la base. La jointure se fait en liant la liste d'itemsets fréquents d'ordre directement inférieur à lui-même (e.g. Les candidats de cardinalité 3 sont obtenus à partir d'une jointure sur l'ensemble des itemsets fréquents de cardinalité 2). Pour que deux k-itemsets puissent être liés, ils doivent posséder k-1 items en commun. Par conséquent, la jointure de deux itemsets d'ordre 1 requiert 0 élément en commun tandis que la jointure entre deux itemsets d'ordre 3 requiert deux éléments en commun. (e.g. les 3-itemsets $\{i_1 i_2 i_3\}$ et $\{i_1 i_2 i_4\}$ peuvent être joints pour former le 4-itemset $\{i_1 i_2 i_3 i_4\}$ ce qui n'est pas possible pour les 3-itemsets $\{i_1 i_2 i_3\}$ et $\{i_1 i_4 i_5\}$ car ils n'ont qu'un seul élément en commun : i_1. De plus, à partir de l'ordre 2 de jointure – la jointure des éléments d'ordre supérieur ou égal à deux – un élément généré doit satisfaire une condition supplémentaire dite « principe d'Apriori » pour être généré. Celui-ci se base sur la propriété d'anti-monotonie (qui stipule que tous les sous-ensembles d'un ensemble fréquent doivent être fréquents) afin d'éliminer certains éléments non fréquents sans avoir à parcourir la base pour évaluer leurs supports ce qui allège da-

vantage l'algorithme. En d'autres termes un élément généré d'ordre k est éliminé si au moins un de ses sous-ensembles d'ordre k-1 est non fréquent.

— Une étape de génération de règles d'association à partir des itemsets fréquents trouvés. Pour chaque itemset fréquent f Apriori identifie tous ses sous-ensembles non vides s et génère, pour chacun, une règle de la forme :

$$\mathbf{s} \longrightarrow (\mathbf{f} - \mathbf{s})$$

Algorithme 1 : Le pseudo code le l'algorithme Apriori [Agrawal et al., 1993]

```
1  Données : D;
2  C_k : Liste d'itemsets candidats de taille k;
3  L_k : Liste d'itemsets fréquents de taille k;
4  L_1 = les items fréquents;
5  pour (k = 1; L_k! = ∅; k++) faire
6      C_{k+1} = candidats générés à partir de L_k;
7      pour chaque transaction t dans D faire
8          Incrémenter le décompte de chaque candidat de C_{k+1} contenu
           dans t;
9      L_{k+1} = les candidats dont le support dépasse le seuil minsup;
10 Résultat : ∪_k L_k;
```

Exemple :

— D : la base de transaction de taille 4
— Le seuil de fréquence minimal : $minsup = 2$
— C_k : Liste des candidats de taille k
— L_k : Liste des candidats fréquents de taille k.

4.3.2.2.2 Evolution du concept

Performance Depuis son introduction dans [Agrawal et al., 1993], il y a eu des centaines de publications proposant des améliorations et des extensions différentes allant des algorithmes scalables et des méthodologies plus efficaces pour l'extraction de motifs fréquents, à la gestion

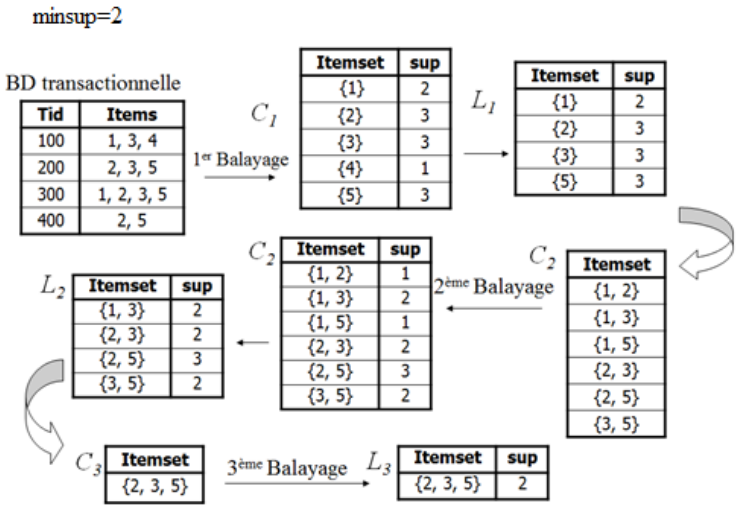


FIGURE 4.4 – Un exemple illustrant l'exécution de l'algorithme Apriori.

d'une diversité de types de données et de tâches d'exploration étendues [Aggarwal and Han, 2014].

La fonction de génération de candidats dans Apriori est un algorithme de recherche complet par niveau (level-wise) basée sur la propriété d'anti-monotonie. Malgré sa capacité à générer des règles en temps linéaire et même à passer à l'échelle de bases de données plus volumineuses [Woo, 2012, Agrawal et al., 1994], Apriori présente trois défis majeurs : les multiples scans de la base de données des transactions, la génération d'un grand nombre de candidats et la charge de travail fastidieuse de compter les supports des candidats. Dans ce contexte, un nombre considérable d'algorithmes essayant de faire face à ces défis et visant à améliorer la capacité de calcul et l'efficacité d'utilisation de la mémoire ont été rapportés. La plupart d'entre eux se sont concentrés sur trois axes principaux :

— La réduction des scans de la base de données transactionnelle, principalement au moyen :
  ◦ Des algorithmes basés sur le partitionnement : dont l'idée est de diviser les données en un ensemble de partitions qui ne se chevauchent pas. Trouver pour chaque partition la liste de tous ses itemsets fréquents, puis fusionner ceux-ci pour générer un ensemble

de tous les itemsets fréquents potentiels [Savasere et al., 1995].
  - Des algorithmes réduisant le nombre de transactions [Vijayalakshmi and Pethalakshmi, 2015].
  - Des algorithmes sans génération de candidats comme FP-growth [Han et al., 2000] and H-mine [Pei et al., 2001], etc..
  - etc.
- Réduction du nombre des candidats à travers:
  - Des algorithmes à base d'échantillonnage : l'idée est de choisir un échantillon aléatoire, pour trouver à l'aide de cet exemple toutes les règles d'association qui se tiennent probablement dans la base de données entière, puis de vérifier les résultats avec le reste de la base de données [Toivonen, 1996].
  - Des algorithmes à base d'hachage [Park et al., 1995] : avec cette extension de l'algorithme Apriori, les itemsets candidats sont divisés en différents blocs [2] et stockés dans un arbre de hachage. Lors du comptage de support, les itemsets contenus dans chaque transaction sont également hachés dans leurs blocs appropriés 4.5. De cette façon, au lieu de comparer chaque itemset dans la transaction avec chaque itemset candidat, il est apparié seulement contre les itemsets candidats qui appartiennent au même bloc.
  - Des algorithmes avec contraintes (algorithme de recherche de motifs fermé [3] A-Close [Pasquier et al., 1999] ; CLOSET [Pei et al., 2000] FPClose [Gösta Grahne and Jianfei Zhu, 2003], AFOPT [Liu et al., 2003]), motif maximal [4] (MAFIA [Burdick et al., 2005] ;) ou d'autres types de contraintes liées à la sémantique, dimension, ou la force (interestingness) des motifs [Boulicaut and Jeudy, 2010].
  - etc.
- Faciliter le comptage des supports des candidats par:
  - Les algorithmes utilisant un format vertical des données tels que ECLAT et CHARM [Zaki, 1998, Zaki and jui Hsiao, 2002]. Ils consistent à transposer la base de données transactionnelles dans un format verticale où chaque item est associé à l'ensemble de tran-

2. Appelé en anglais buckets ou slots
3. Un motif fréquent est dit fermé s'il ne possède aucun sur-motif qui a le même support.
4. Un motif fréquent est dit Maximal si aucun de ses sur-motifs immédiats n'est fréquent.

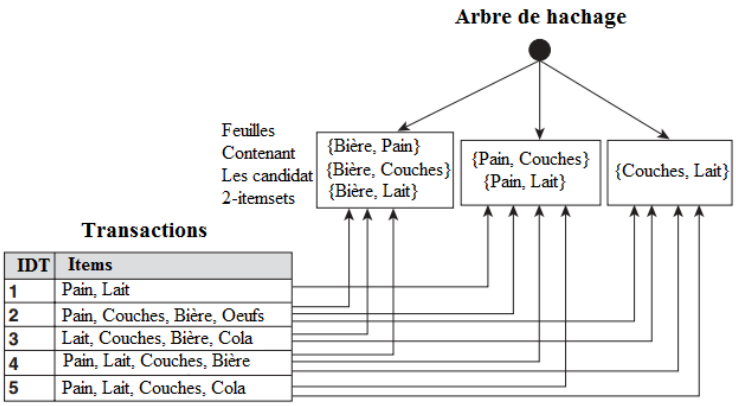


FIGURE 4.5 – Compter le support des itemsets en utilisant la structure de hachage.

sactions qui le contiennent. Le comptage du support d'un itemset est simplement la longueur de l'ensemble de transactions. Le support d'un ensemble peut être compté beaucoup plus facilement en croisant les couvertures des deux de ses sous-ensembles qui donnent ensemble l'ensemble lui-même.

- Les algorithmes à base d'hachage, etc.

— Déploiement sur des environnements Big Data [Woo, 2012]

Des explications plus détaillées sur ces derniers, ainsi que des références d'exemples de leurs implémentations et des exemples d'algorithmes correspondant à d'autres méthodologies, peuvent être trouvés dans les ouvrages suivants [Han et al., 2000, Aggarwal, 2014].

Des règles d'association plus complexes pour des tâches plus variées Outre les améliorations de performance, diverses autres extensions des algorithmes basiques de recherche de règles d'association ont été proposées. Celles-ci se sont, principalement, concentrées sur la manipulation de différents types de données et sur l'intégration d'autres tâches de fouille de données, afin de produire des nouveaux types de règles pouvant offrir une vision plus réaliste et complète des problèmes étudiés.

— **Règles quantitatives et floues :**
Dans leur forme traditionnelle, les règles d'association sont dérivées

à partir de bases de données binaires où un item prend "1" s'il existe dans une transaction et "0" autrement. Le problème avec ce genre de règles est qu'elles impliquent, exclusivement, des attributs catégoriels. Pour pallier à ce problème, des méthodes de recherche de règles d'association quantitatives furent introduites [Gosain and Bhugra, 2013]. L'essence de ses méthodes consiste à diviser les valeurs des variables continues en intervalles ce qui permet de transformer une base de données décrite par des variables continues en une base binaire. En effet, cette dernière est obtenue en vérifiant si un item décrit par la valeur X de la variable continue Y est un membre ou non de l'un de ses intervalles. Cependant, l'adhésion ou la non-adhésion stricte peut mener à un problème de rejet de valeurs près des frontières. Les règles d'association floues [Farzanyar and Kangavari, 2012] ont été proposées pour surmonter ce problème en considérant l'appartenance partielle des items par le biais d'une fonction d'appartenance ainsi que des opérations issues de la théorie des ensembles flous. Ceci permet, ainsi, de générer des règles qui auraient pu être omises avec l'approche quantitative standard.

— **Règles multi-niveaux et multidimensionnelles**
Les données décrivant les problèmes et situations de la vie réelle se localisent, principalement, dans des espaces multidimensionnels et multiniveaux, comme des entrepôts de données et des bases de données relationnelles. C'est-à-dire que les items décrits dans les données sont souvent associés à des informations supplémentaires fournissant plus de détails sur leurs caractéristiques et positions hiérarchique (e.g. l'item « hôpital » est un descendant de l'item « bâtiment administratif »), par exemple l'emplacement, la population et l'usage du sol ou la fonction pour une zone géographique. Dans ce contexte, la recherche de règles d'association a été étendue afin de générer des règles multidimensionnelles et multi-niveaux [Han et al., 2012] dont l'objectif est de tirer parti de ces informations supplémentaires pour être plus riches et plus utiles. Effectivement, impliquer des concepts à différents niveaux d'abstraction apporte de la flexibilité nécessaire pour traiter deux problèmes. Le premier est lié à la difficulté de trouver des règles fortes à des niveaux d'abstraction primitifs en raison de la nature creuse des données, à ces niveaux. Le second consiste à extraire des règles fortes (fréquentes) mais pas très utiles, car elles représentent la connaissance du bon sens. Pratiquement, de nombreuses propositions impliquant les

mesures de support et de confiance, ont été proposées pour une exploitation efficace des règles à plusieurs niveaux. A titre d'exemple, nous pouvons faire varier le support minimum selon les différents niveaux [Gautam and Pardasani, 2011, Lee et al., 2006], ou garder le support uniforme à travers les niveaux et commencer par chercher les itemsets fréquents des niveaux les plus élevés, puis, extraire, seulement, les itemsets des niveaux inférieurs, qui leurs correspondent. Par conséquent, les règles redondantes peuvent être éliminées comme les règles de niveau inférieur peuvent être fondées essentiellement sur des règles de niveaux supérieurs et la distribution des itemsets qui leurs correspondent.

— **Règles spatio-temporelles :** Une autre extension importante de la fouille classique de motifs fréquents est la recherche de règles d'association spatiales, temporelles et spatio-temporelles [Xuewu et al., 2008]. Les règles d'association spatiales traitent des informations tels que l'emplacement et la topologie des items et tentent d'extraire des motifs fréquents montrant, principalement, l'interaction entre deux ou plusieurs attributs dépendant de l'espace ou des objets spatiaux. Alors que les règles d'association temporelles, capturent les attributs dépendant du temps pour révéler des connaissances sur la nature cyclique, périodique ou séquentielle de certains motifs. La véritable valeur ajoutée de l'exploration des règles d'association dans un contexte temporel consiste à gagner des capacités de prédiction. Par exemple, dans l'exploration de règles séquentielles [Rudin et al., 2013a], les occurrences des items ainsi que l'ordre entre ceux-ci comptent pour produire des règles qui tentent de déterminer quel événement sera ensuite révélé, basé sur des séquences d'événements passés.

— **Règles impliquant des éléments rares :**
Dans la recherche de motifs fréquents et également la recherche de règle d'association, l'importance ou la force d'un motif ou d'une règle est évaluée selon la fréquence des items qui y sont impliqués. Or, selon les applications, les données peuvent être déséquilibrées tout en comportant des éléments très fréquents et d'autres très rares qui ne seront jamais pris en compte dans le modèle généré. Ces items, peuvent s'avérer significatifs pour le problème traité si on considère des critères autres que sa fréquence (e.g. critères sémantiques), d'où un problème de perte d'information utile. Dans cette perspective, plusieurs approches ont été proposées, afin de générer des forts motifs ou règles

impliquant, à la fois, des items fréquents ainsi que des items rares pouvant véhiculer des informations importantes [Bhatt and Patel, 2014]. À savoir :

- Une approche basée sur les techniques de recherche opérationnelle telles que l'optimisation par colonie d'abeille [Rousseaux and Ritschard, 2014], l'algorithme génétique [Ghosh and Nath, 2004], l'optimisation par essaims particulaires [Sarath and Ravi, 2013], etc. Ces techniques consistent à générer des règles en formulant le problème de recherche des règles d'association comme un problème d'optimisation qui consistent à minimiser ou maximiser une fonction (appelé la fonction objective) sur un ensemble. Cette fonction sert de critère pour déterminer la meilleure solution à un problème d'optimisation (*i.e.* les meilleures règles d'association). Par exemple dans [Ghosh and Nath, 2004], le problème de recherche de règles d'association est formulé comme un problème d'optimisation multi-objectifs et les mesures de support, de compréhensibilité [5] et d'intérêt [6] sont utilisées pour évaluer les règles. En utilisant ces trois mesures comme les objectifs du problème un algorithme génétique est utilisé pour extraire des règles utiles et intéressantes à partir de n'importe quelle base de données de type panier de la ménagère.
- Une approche basée sur les mesures d'intérêt, en particulier, l'utilisation de plusieurs supports minimum. En effet, l'utilisation d'un seul seuil de support est inappropriée en cas de distribution déséquilibrée des attributs (items) dans les données. Si le minsup est trop élevé, nous pouvons ne pas trouver de règles qui impliquent les attributs minoritaires. On peut donc penser à définir un minsup trop bas, ce qui provoquerait une explosion combinatoire, comme les items fréquents seront associés ensemble de toutes les manières possibles et engendreront, ainsi, trop de règles non pertinentes. Dans ce contexte, de nombreuses recherches ont proposé des algorithmes utilisant plusieurs minsup ce qui qui permettrait aux utilisateurs de trouver des items rares sans générer des règles trop peu significatives (*i.e.* impliquant

5. Est mesurée par le nombre d'attributs impliqués dans la règle et essaie de quantifier la compréhensibilité de la règle.

6. Une mesure pour estimer combien une règle est intéressante.

uniquement des combinaisons d'éléments très fréquents). Comme exemples, nous pouvons citer MSApriori et certaines de ses extensions [Rai et al., 2012, Kouris et al., 2003, Kiran and Re, 2009].

- Dans leurs travail, Fahed et al [Fahed, 2016] proposent de définir une méthode pour la génération d'un modèle prédictif à partir de séquences d'évènements variées et volumineuses.
  Visant à faire apparaitre des types particuliers d'évènements faiblement représentés dans ces séquences longues et bruitées (e.g. évènement distants dont l'horizon d'apparition est éloigné temporellement afin que l'utilisateur puisse avoir le temps de réagir), elles proposent une sémantique particulière de la conclusion permettant de filtrer les conclusions non pertinentes pour, ensuite, ne compléter la recherche de prémisses que pour les conclusions respectants cette sémantique (*i.e.* les règles doivent comporter une distance temporelle entre l'antécédent (le déclencheur) et la conséquence (l'évènement futur), ce qui permet de préciser l'horizon d'apparition de ce dernier évènement). Dans notre travail, nous nous inspirons de cette méthode pour réduire notre espace de recherche en commençant par générer les séquences intéressantes selon une sémantique que nous avons définie puis par trouver leurs antécédents.

— **Règles traitant d'autres tâches de fouille de données :** Les motifs fréquents découverts via des processus d'exploration de donnés ne sont pas seulement intéressants en eux-mêmes, mais aussi utiles à d'autres tâches de fouille de données telles que la classification associative et le regroupement ou le clustering [Romero et al., 2010, De Sá et al., 2011, Dua and Kidambi, 2010].
La classification est une tâche de datamining qui consiste à étiqueter des éléments d'une collection. En d'autres termes, il s'agit de les affecter à des catégories ou à des classes cibles. Dans la classification associative, l'idée générale est que des associations fortes entre les itemsets fréquents et les étiquettes de classe peuvent être découvertes. Ensuite, les règles d'association, ainsi générées, sont utilisées pour la prédiction.
La recherche de règles d'association classifiantes (RAC), aussi appelé en anglais *Class Association Rule mining*, consiste à découvrir des règles sous la forme de $a \longrightarrow b$, où la partie conclusion (b) doit être un item marqué comme une classe. Ces règles peuvent ensuite

être utilisées pour prédire la classe des enregistrements non classifiés. Cette méthode permet aux utilisateurs de diminuer le nombre de règles générées en précisant la forme de règle qui les intéresse.
Les RAC ont été exploitées avec succès dans différentes applications de prédiction. Selon de nombreuses études expérimentales, e.g., [Thabtah et al., 2004], cette approche peut surpasser les méthodes classiques de classification, telles que les arbres de décision (e.g. C4.5 [Kumar and Verma, 2012]), dans la construction de systèmes prédictifs plus précis.
Comme exemples d'algorithmes de classification associative, on peut citer les algorithmes : CBA [Liu et al., 1998], CMAR [Li et al., 2001], (CPAR) [Yin and Han, 2003], RCBT [Cong et al., 2005], HARMONY [Wang and Karypis, 2005] et plus récemment les algorithmes CAR-Miner [Nguyen et al., 2013] et CCAR [Nguyen et al., 2015].
Contrairement à la classification, le clustering consiste à grouper un ensemble d'enregistrements selon des critères de similitude, sans prendre en compte d'étiquettes de classe. Le regroupement dans un espace à haute dimensionnalité représente un défi important [Parsons et al., 2004]. La possibilité d'extraire les motifs fréquents dans les sous-ensembles à haute dimensionnalité, représente une direction prometteuse pour des problèmes tels que le clustering des sous-espaces [7] (*sub-space clustering*), et le clustering des datasets à dimensions élevées. Dans ce contexte, CLIQUE, un algorithme de clustering des sous-ensembles basé sur Apriori a été proposé par [Agrawal et al., 1998]. Cet algorithme intègre des méthodes de clustering basées sur la densité et sur la grille. Il utilise, en fait, la propriété d'Apriori pour trouver des sous-espaces groupables puis identifier les unités qui sont denses pour formant, par la suite, les clusters. L'un des principaux domaines d'application du clustering basés sur les motifs fréquents est la fouille de texte [Zhang et al., 2015]. La fouille de texte basé sur le regroupement des mots clés ou de données issues de micropuces d'ADN [8] (*microarray data*) représente des pro-

7. Le clustering des sous-espaces est une extension du clustering traditionnel qui cherche à trouver des groupes ou clusters dans différents sous-espaces au sein d'un ensemble de données.

8. Appele en anglais microarray data : Une base de données de micropuces d'ADN est un référentiel contenant des données d'expression de gènes d'ADN. micropuces d'ADN représente un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite

blèmes à haute dimensionnalité et cette aproche de clustering à base de recherche de motifs fréquents commence à démontrer son pouvoir et à quel point elle est prometteuse. Dans ce contexte, plusieurs algorithmes et méthodes ont été proposés. Citons, par exemple la méthode de clustering de texte basé sur les termes fréquents [Beil et al., 2002], et l'algorithme pCluster de [Wang et al., 2002], qui est une méthode de regroupement par similarité pour l'analyse des données de micropuces d'ADN.

4.3.2.3 Applications

La recherche de motifs et de règles d'association fréquents représente une technique de data mining employé, principalement, pour analyser et identifier des associations ou des liens dans des données transactionnelles [Han et al., 2011]. Ainsi, elle génére un modèle qui reflète les connaissances incorporées dans les données et qui peut être utilisé pour résoudre des problèmes dans divers domaines d'application. Ce type de modèles a été, d'abord, appliqué pour le problème du panier de la ménagère avec l'objectif de déterminer les habitudes d'achat des clients en trouvant des associations entre différents articles que les clients placent dans leurs paniers [Kaur and Kang, 2016]. Parmi les autres applications indiquées dans la littérature, nous trouvons la recommandation d'achats [Zhao et al., 2014], la détermination du profil client [van Dam and van de Velden, 2015] la gestion de la relation client [Linoff and Berry, 2011], l'analyse des séquences d'ADN [Wang et al., 2016], etc.

Bien qu'elle ait été appliquée pour la première fois à des données transactionnelles simples, essentiellement à des fins analytiques, cette tâche de datamining a été développée et étendue pour traiter d'autres types de données telles que les données temporelles, spatio-temporelles, graphiques et incertaines, etc. Selon [Aggarwal, 2014], de tels types de données ont de nombreuses applications à d'autres problèmes de data mining tels que le regroupement et la classification. Par conséquent, la recherche de motifs fréquents a gagné un pouvoir prédictif en plus de leur capacité d'analyse, comme un modèle de règles d'association classifiantes représente un modèle capable de prédire de valeurs inconnues de certaines variables (classes) en se basant sur les données étudiées.

surface qui peut être du verre.

Dans la prochaine section nous nous focalisons sur la capacité prédictive des règles d'association et nous fournissons des exemples concrets des différents domaines d'application.

4.3.2.4 Règles d'association pour la prédiction

Un modèle prédictif basé sur les règles d'association utilise l'antécédent de la règle pour prédire sa conclusion. En d'autres termes, il indique quel item est susceptible de se produire compte tenu de l'occurrence d'un certain itemset. Ce type de modèles a été employé pour de nombreuses applications dans différents domaines, on peut citer la prédiction des futurs achats des clients probables dans le contexte de l'analyse du panier de la ménagère [Chen et al., 2014] ; en biologie, la prédiction des fonctions des protéines, en se basant sur les réseaux d'interactions protéine-protéine [Park et al., 2015] ; En médecine, la prédiction du niveau de risque pour les patients ayant une maladie cardiaque [Ilayaraja and Meyyappan, 2015] ; dans le domaine de la gestion de l'énergie, les règles d'association ont été employées pour la prédiction de l'emplacement des occupants des bâtiments. Ce dernier problème consiste à déterminer l'emplacement des occupants des bâtiments, en fonction de l'historique de leurs mouvements, afin de maximiser l'efficacité énergétique de chauffage, de ventilation et de climatisation (CVC). En d'autres termes, faire fonctionner le système CVC en fonction des mouvements des occupants pour satisfaire leurs besoins sans dilapider l'énergie [Ryan and Brown, 2013].

Plusieurs autres travaux tels que [Lin and Li, 2015] ont montré l'intérêt de la recherche de règles d'association pour extraire des modèles spatio-temporels explicatifs et prédictifs.

Ce type de modèles découvre, généralement, des motifs et des règles liés à des relations spatio-temporelles qui sont typiquement incorporés dans des données géospatiales au lieu d'être explicitement transcrites dans la base de données. Les règles d'association spatiales traitent des informations telles que l'emplacement et la topologie des items, afin d'extraire des motifs fréquents montrant, principalement, l'interaction de deux ou plusieurs attributs dépendants de l'espace ou des objets spatiaux qui sont en son sein. Alors que les règles temporelles, saisissent les attributs dépendant du temps pour révéler la connaissance de la nature cyclique, périodique ou séquentielle de certains évènements ou motifs. La véritable valeur ajoutée de la recherche de règles d'association dans un contexte temporel, est de gagner

plus de capacités prédictives. Par exemple, dans l'extraction de règles séquentielles [Rudin et al., 2013b], à la fois, les occurrences d'items et l'ordre entre celles-ci comptent dans la production des règles dont l'objectif est de déterminer quel événement sera prochainement révélé en se basant sur des séquences d'événements passés.

Selon le contexte applicatif, les modèles prédictifs à base de telles règles d'association peuvent être employés pour recommander un évènement prédit ou un autre. En guise d'exemples, nous pouvons citer les règles d'épisodes qui, extraites à partir des données séquentielles sur les intersections des internautes avec les sites web et entre eux, peuvent servir pour la prédiction de leurs comportements et, ainsi, pour la recommandation des pages web à visiter [Laxman et al., 2008]. Dans les applications liées aux trafics routiers, des règles générées à partir de séquences d'évènements routiers peuvent être utilisées pour prédire les zones d'embouteillage et donc recommander, aux conducteurs, des itinéraires permettant de les éviter [Cho et al., 2008]. Dans un autre exemple, Laxman et al [Laxman et al., 2009] ont présenté un modèle prédictif de l'état final d'un produit afin de recommander des mesures pour l'améliorer. Les règles d'association, dans celui-ci, ont été extraites à partir des données séquentielles décrivant les étapes de production.

Dans le domaine géographique, bien que rares, ils existent quelques travaux qui se sont basés sur l'analyse des relations spatiales et temporelles afin de produire des règles d'association pouvant être employées pour la prédiction des phénomènes spatio-temporel. Dans notre contexte d'application, l'évolution territoriale, nous pouvons citer le travail de Alouaoui et al [Alouaoui et al., 2015]. Les auteurs, dans ce travail, ont considéré une zone réduite de la ville de Tunis entre 1987 et 2001. Ainsi, ils ont, tout d'abord, appliqué des requêtes spatio-temporelles afin d'extraire les relations spatiales et temporelles reliant les différents objets (deux à deux) au fil du temps ; leur appliquer, par la suite, l'algorithme Apriori afin de générer des règles d'associations ; et enfin généraliser ces règles sur tout le territoire. Indiquant l'association entre les caractéristiques spatio-temporelles des versions prédécesseurs d'un objet et celles de sa version successeure (cf.figure 4.6), ces règles peuvent être utilisées pour prédire l'évolution des objets d'un territoire. Un objet est caractérisé par un attribut non spatio-temporel (e.g. occupation du sol), un attribut décrivant ses relations avec les objets voisins, un attribut décrivant ses caractéristiques géométriques et un attribut indiquant sa date de validation.

La recherche de règles d'association a évolué pour tenir compte des

$X(a_i, gi, pi, ti) \wedge X(a_j, g_j, p_j, t_j) \wedge \ldots \wedge X(a_n, g_n, p_n, t_n) \longrightarrow X(a_m, g_m, p_m, t_m)$

Avec:

- X: l'objet de référence.
- (a): attributs non-spatiotemporels (e.g. occupation du sol …).
- (g): caractéristiques géométriques de X (e.g. Géométrie …).
- (p): relations spatiales de X avec les objets voisins
- $t_i, t_j, \ldots, t_n, t_m$: est une séquence successive de dates avec ($t_i<t_j<t_n<t_m$).

FIGURE 4.6 – Structure des règles spatio-temporelles générées [Alouaoui et al., 2015].

spécificités des données spatio-temporelles ce qui renforce leurs capacités aussi bien prédictives qu'analytiques dans les contextes d'application spatio-temporelles.

En effet, les relations spatiales et temporelles peuvent avoir des aspects hiérarchiques. Par exemple, dans le cas de la couverture terrestre, une zone industrielle peut également être désignée comme zone développée à un niveau hiérarchique supérieur. Par conséquent, les règles impliquant des « zones développées » et d'autres impliquant des « zones industrielles » sont alors générées. Dans ce contexte, la recherche de règle d'association simple à évoluer pour extraire des règles multi-niveaux, c-à-d, des règles qui prennent en considération l'aspect hiérarchique de certains motifs tels que les relations spatio-temporelles (e.g. relation d'inclusion entre deux zones géographiques).

Les règles d'association sont, conventionnellement, conçues pour traiter uniquement des données catégoriques. Cependant, les relations spatiales incluent aussi bien des relations métriques telles que la distance qui sont décrites avec des données numériques. Ce problème a été abordé en présentant l'extraction de règles d'association quantitatives qui a été combinée, par la suite, avec les théories d'ensembles flous pour gérer l'aspect imparfaits des relations spatiales (e.g. déterminer les degrés de voisinage des objets géographiques par l'appartenance partielle à des intervalles de l'attribut de distance) [Farzanyar and Kangavari, 2012].

4.4 Conclusion

Le problème d'extraction de règles d'association est un domaine de recherche très actif. Mis à part sa capacité à capturer les relations et structures incorporées dans les données d'apprentissage, cette méthode présente plusieurs atouts tels que son exhaustivité en termes de génération de règles ou de motifs fréquents, sa simplicité, son intuitivité et sa capacité à gérer le problème de démarrage à froid posant souvent problème aux autres méthodes de la fouille de données (e.g. classification). L'algorithme le plus populaire pour l'extraction de motifs est sans aucun doute Apriori [Agrawal et al., 1998]. Au fils de temps, cet algorithme a connu plusieurs améliorations et extensions allant des algorithmes scalables et des méthodologies plus efficaces pour l'extraction de motifs fréquents, à la gestion d'une diversité de types de données et de tâches d'exploration étendues tels que la classification associative et le clustering. Par conséquent, bien qu'il ait été essentiellement présenté comme un outil analytique, cet outil a acquis des capacités de prédiction d'où son application dans plusieurs problèmes de prévision. Dans le chapitre 5 nous proposons un modèle de prédiction spécialement conçu pour le problème de suivi et prédiction de l'évolution du territoire. Ainsi nous exploitons quelques notions (présenté ci-dessus) pour relever des défis liés essentiellement à : (1) la présentation adéquate des données d'apprentissage ; (2) la gestion du déséquilibre en leur sein ; (3) et la représentation adéquate, compréhensible et facilement implémenté d'un modèle d'apprentissage, dans un système de prédiction. En effet, notre objectif ultime est d'automatiser ces trois tâches au sein de ce système qui prendra en entrée un ensemble de cartes géographiques décrivant un territoire dans le passé, pour donner en sortie sa carte future. Dans le chapitre suivant nous fournissons, également, un état d'avancement par rapport à cet objectif.

Chapitre 5

Vision méthodologique des contributions

5.1 Introduction

Chaque entité physique peut très souvent être associée à une localisation dans l'espace et certains de ses attributs peuvent varier avec le temps. Par conséquent, il est utile de développer des techniques résumant efficacement ces données et permettant, entre autres, la saisie du comportement évolutif de ces entités au fils du temps. Ceci permettrait, ainsi, de fournir un aperçu utile pour le suivi et la prédiction d'éventuelles occurrences d'évènements qui leurs sont liés.

Dans le contexte de l'évolution territoriale, notre travail vise à proposer une approche automatisée qui exploite la fouille de données pour produire un ensemble de règles en régissant les dynamiques. Il convient de noter que dans un travail précédent nous avons extrait des règles se basant uniquement sur les relations exprimant la continuité temporelle d'usage (succession de fonction) des objets étudiés [Gharbi et al., 2014].

Dans ce mémoire, nous partons de l'hypothèse que les prédécesseurs et leurs voisinages influencent les évolutions et donc notre objectif consiste à extraire des règles en conséquence. Notre démarche peut donc se comprendre comme une évolution du précédent travail. Concrètement, elle propose d'explorer les co-occurrences fréquentes entre les valeurs des variables décrivant : l'historique de l'évolution d'un certain objet spatio-temporel (relations temporelles de succession d'occupation), l'historique de ses co-localisations (re-

lations spatiales de voisinage) et sa future évolution d'usage. Ceci pourrait correspondre à une tâche courante de fouille de données : la recherche de règles d'association. Ainsi, une finalité de ce travail consiste à adapter les règles d'association pour souligner l'effet des relations spatio-temporelles sur l'évolution d'un territoire.

Bien qu'elles disposent d'une vocation analytique et explicative, les règles extraites peuvent être utilisées pour prédire un évènement ou un changement si les conséquences des règles sont réservées pour les évènements à prédire. Ceci représente, en fait, un deuxième volet à explorer et une finalité en fonction de laquelle notre approche a été élaborée. En effet, un modèle explicatif et éventuellement prospectif représente plus d'intérêt, comme il permet d'assister à la prise de décision par la recommandation ou non d'une action, selon le contexte d'application (e.g. planification urbaine et aménagement du territoire).

Pour cela, nous estimons que notre approche devrait être capable de répondre aux défis mentionnés et expliqués au début de ce mémoire, à savoir :

— considérer l'autocorrélation spatiale
— prétraiter les données pour un apprentissage pertinent
— traiter l'aspect asymétrie et déséquilibre des données
— donner des pistes pour passer à l'échelle

Dans ce chapitre, nous essayons de fournir une vision méthodologique de nos contributions. Ainsi, dans la section suivante nous présentons un aperçu général de l'approche proposée tout en soulignant les propositions faites pour répondre à ces défis. Celles-ci seront, par la suite, explicités, chacune, dans une section à part.

5.2 Approche globale

Afin de générer un ensemble de règles d'association régissant le phénomène de changement d'occupation du sol, nous sommes partis de l'hypothèse que ce phénomène peut en partie être expliqué à travers l'historique des changements de fonction des entités géographiques, constituant un territoire, et des configurations spatiales dans lesquelles elles se situent. Ainsi, nous avons considéré un jeu de données constitué d'une série de cartes géographiques indépendantes, décrivant un même territoire à des dates consécutives. L'objectif est d'en **extraire un modèle spatio-temporel** capable de fournir la trajectoire d'évolution de chaque entité. En d'autres termes, identifier les

séquences de succession de ses fonctions et de ses relations spatiales (voisinages). Ces trajectoires, que nous appelons « séquences d'évolution », représentent notre jeu de données d'apprentissage. Celui-ci subira, dans une deuxième étape un **prétraitement** afin de le représenter sous un format permettant la génération d'une forme particulière de règles que nous proposons et jugeons adéquate pour notre problème de prédiction. Bien que ce jeu de données paraisse prêt pour l'extraction de ce genre de règles, celui-ci s'est avéré déséquilibré. Une première application de l'algorithme apriori a généré des règles impliquant seulement les relations de voisinage ce qui s'explique par la domination des items correspondant à ces relations par rapport aux items correspondant aux autres types de relations (e.g. relations temporelles de changement de fonctions). Dans ce contexte, nous essayons, dans une troisième étape, de traiter ce problème d'asymétrie de données par **l'application d'un algorithme adapté**.

La figure 5.1 présente les liens entre les différents défis auxquels nous devrons faire face, les étapes de l'approche définie afin de répondre à ceux-ci et les propositions faites pour y répondre. Ainsi, pour :

- **L'autocorrélation spatiale** (*i.e.* la valeur d'une variable est influencée par les valeurs de la même variable à des localisations voisines), ce défi est traité dans la première étape par l'explicitation des relations de voisinage et dans la dernière étape par la génération de règles soulignant l'effet des fonctions des voisins sur le changement de fonction de l'objet considéré. La principale contribution qui traite l'autocorrélation est le passage d'un modèle de type succession d'états (modèle en *snapshot*) à un modèle se basant sur le paradigme identitaire et permettant le suivi de l'évolution sur les niveaux fonction et voisinage.
- **La représentation adéquate des données pour l'apprentissage**, ce défi est traité au sein de la deuxième étape en proposant, d'abord, une forme de règle adéquate pour la génération d'une règle de prédiction (*i.e.* une règle dont la conclusion, réservée à un attribut décrivant la fonction de l'objet successeur, est déterminée par une prémisse qui comporte un attribut décrivant la succession des fonctions de ses prédécesseurs et un ou plusieurs attributs décrivant les fonctions de leurs voisins) et ensuite, en proposant un format d'organisation des données d'apprentissage permettant la génération de ce genre de règles.
- **L'asymétrie des données**, traitée dans la troisième étape à travers la proposition de deux algorithmes. Le premier adapte apriori en proposant d'utiliser plusieurs supports minimums définis selon deux

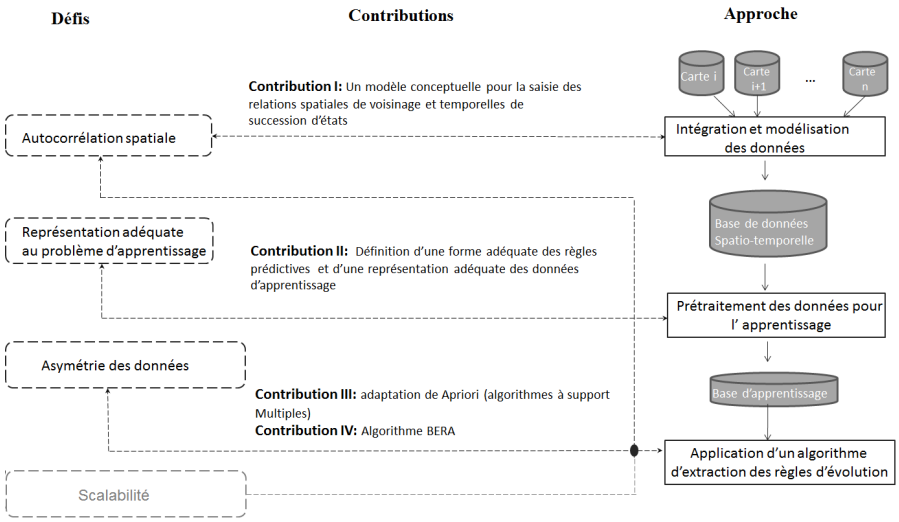


FIGURE 5.1 – Présentation synthétique des contributions méthodologiques.

propositions : une méthode par analyse statistique et une méthode par groupement [Gharbi et al., 2016c] ; Le second appelé *Backtrack Extraction Rule Algorithm* (BERA), se base sur l'idée de partir des items rares pour remonter aux items les plus courants en se basant sur leur sémantique : conclusion puis successions puis voisinages.

— **La scalabilité**, ce défi ne sera pas directement traité dans ce document. Nous proposerons cependant des pistes pour le passage à l'échelle de notre démarche dans les perspectives.

Afin d'exécuter les différentes étapes de notre approche et tester, également, les propositions mentionnées ci-dessus, nous avons développé un dispositif expérimental, appelé SAFFIET (*Spatial And Functional Frequent Itemset Extraction Tool*) [Gharbi et al., 2016a], dont le fonctionnement est illustré par la figure 5.2. Ses différentes étapes seront détaillées à mesure de l'avancement de la description de notre démarche. Il convient de noter, également, que ce même schéma sera repris dans le chapitre 6 tout en soulignant les spécifications informatiques du dispositif développé (*i.e.* les outils et solutions informatiques mis en collaboration, les types et les formats des fichiers impliqués, les langages utilisés, etc.).

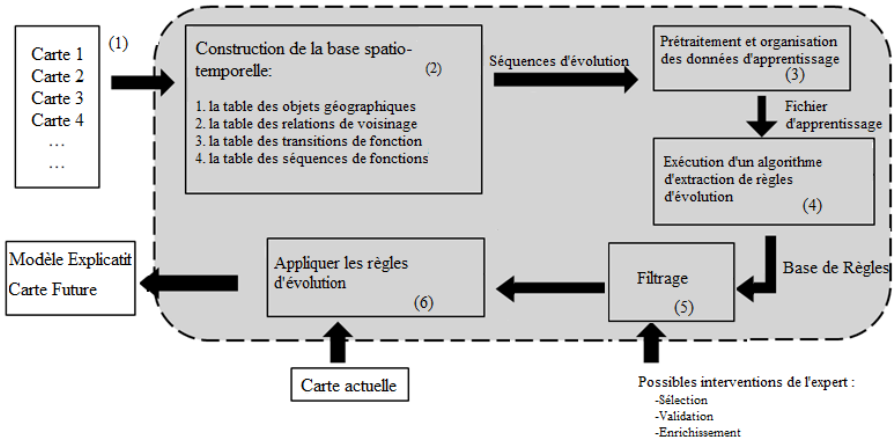


FIGURE 5.2 – Pipeline méthodologique de notre démarche : chaine de traitements pour l'extraction de règles spatiotemporelles en vue de la caractérisation de modèles explicatifs et potentiellement prédictifs des évolutions d'un territoire.

5.3 Modélisation spatio-temporelle des entités évolutives

5.3.1 Données d'entrée

L'objectif principal de cette étape de modélisation consiste au suivi de l'évolution du territoire d'étude afin d'analyser, comprendre et expliquer ses dynamiques. Ainsi, on a, d'abord, besoin de disposer d'une vue globale de ses différents états pendant un certain intervalle d'étude. Ceci est généralement assuré par un modèle dit « par superposition de couches datées » (*snapshot* model en anglais). Ce modèle s'inspire, principalement, de la notion de « couche » utilisée, dans les SIG, pour la représentation de données géographiques telles que les types d'occupation biophysique du sol. Une couche est définie par l'association des données au support spatial qu'elles décrivent et un support spatial est composé d'un ensemble d'objets de même type (e.g. l'objet OB_1 est défini par l'occupation OC_1). À l'encontre de ces couches représentant, séparément, chaque type d'occupation du sol ; le modèle en *snapshot* présente une vue fusionnée de ceux-ci, à une date donnée (cf. figure 5.3). Ainsi la superposition d'un ensemble de couches datées peut infor-

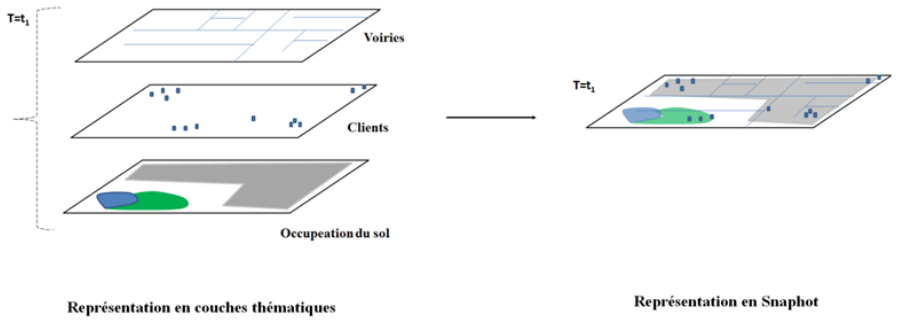


FIGURE 5.3 – Représentations des données : modèle de collecte classique et en modèle spatio-temporel en *Snapshot*.

mer sur l'évolution des objets définissant le territoire d'étude ; ceci permet, en fait, de distinguer les changements de leurs attributs entre deux versions temporelles du territoire (couches).

La base *Corine Land Cover* (CLC), utilisée dans cette thèse, représente un exemple concret d'une telle modélisation [EEA, 2009].

5.3.1.1 Entité spatio-temporelle

Dans ce travail, nous considérons qu'un objet géographique représente le produit de trois composantes qui correspondent, respectivement, à une dimension temporelle, une dimension spatiale et une dimension sémantique [Rodier and Saligny, 2010] (cf. figure 5.4).

Pour modéliser la dimension temporelle de l'évolution, nous adoptons la conceptualisation linéaire du temps comme elle permet de respecter le principe de causalité. Ce principe, qui représente l'axiome de notre travail, stipule que tout évènement est l'effet des actions et évènements qui l'ont précédé. Nous partons, donc, de l'hypothèse que l'occupation du sol est un phénomène qui peut en partie être expliqué par l'historique des changements de fonction des entités géographiques constituant un territoire et des configurations spatiales dans lesquelles elles se situent. Dans ce contexte, nous formalisons le temps à travers l'ensemble des évènements qui se produisent dans l'espace géographique étudié. Ainsi, nous l'appréhendons qualitativement [Thériault and Claramunt, 1999] comme étant une succession d'évènements [De Risi, 2012] de changement de fonction et/ou de forme ou topologie,

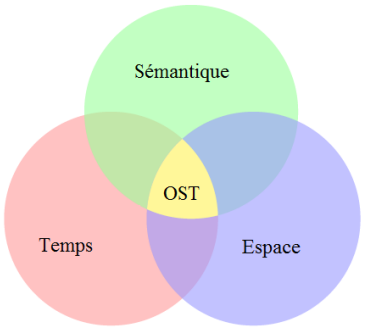


FIGURE 5.4 – Conceptualisation d'un objet Spatio-temporel (OST) selon Rodier et Saligny [Rodier and Saligny, 2010].

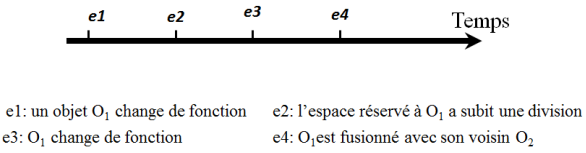


FIGURE 5.5 – Une représentation linéaire, ordonnée et quantitative du temps dans le contexte de changement d'occupation du sol.

ce qui nous permettra d'extraire les séquences de vie des objets, à étudier (cf. figure 5.5).

Pour la dimension spatiale, nous considérons que l'espace est un support immuable où se situent les objets et les relations exprimant leurs configurations spatiales. Selon cette conception newtonienne, l'espace est isotrope et homogène [Peterschmitt, 2012].

5.3.1.2 Suivi de l'évolution d'une entité

Tenant compte des éléments évoqués, l'évolution d'un objet est caractérisée par la modification de la valeur d'au moins un de ses attributs ou composantes (valeur qui correspond à l'une de ses trois dimensions). Ce changement donne, ainsi, lieu à la génération d'une nouvelle version du même objet physique que nous appelons entité. L'entité ou l'atome selon Worboys [Worboys, 1992] est représenté comme une structure en 3D : deux

dimensions pour la portion d'espace couverte par cet atome et une dimension pour l'intervalle temporel de sa validité. Le temps est donc représenté comme étant perpendiculaire et orthogonal au plan spatial. La génération de ces versions est en fait, visible dans le cas des modèles *snapshot* où une couche successeure présente des versions récentes des objets identifiés à la couche de départ.

Dans ces modèles, le suivi de l'évolution d'un objet physique consiste à la projection de ses atomes ou entités sur l'axe de temps. Comme celles-ci sont définies par leurs attributs et que ces attributs sont susceptibles de changer, établir le lien entre elles et l'objet physique auquel elles correspondent est impossible. Ce problème peut être résolu, selon Cheylan [Cheylan and Lardon, 1993], en recourant à l'affectation aux objets comme à ses atomes, d'identifiants persistant tout au long de leurs évolutions. Dans certains jeux de données géo-historiques (*i.e.* CLC), chaque objet est identifié par un identifiant unique pour chaque *snapshot*, ou par cycle de vie. Ces identifiants peuvent changer, d'un *snapshot* à l'autre, lors d'une évolution de forme ou de fonction. Par conséquent, le lien de composition entre chaque objet et ses successeurs est interrompu. La figure 5.6 illustre, sur les données Corinne Land Cover[EEA, 2009], les types d'évolution possibles (changement de fonction, changement de forme distinguée par un changement de la valeur de la superficie, et changement de forme et de fonction) et souligne, également, le changement d'identifiant qui les accompagne.

Une solution fréquente pour remédier à ce problème consiste à placer le traceur de l'identité sur l'un des attributs de l'objet, très souvent son empreintes spatial [Kauppinen and Hyvönen, 2007]. Ainsi, nous identifions un objet par sa zone spatiale allouée puis, aux moyens d'un ensemble de requêtes spatiales, nous tentons de définir son contexte spatio-temporel. Donc, tout d'abord, nous utilisons des requêtes de tangence pour identifier les voisinages de chaque objet, puis nous utilisons des requêtes d'intersection pour chercher leurs successeurs (*i.e.* ces requêtes reçoivent les couples d'objets dont l'intersection est non nulle) (cf. algorithme 2).

Cette solution (identifier les objets à travers leurs empreintes spatiales) exige la disposition d'un support spatial invariable dans le temps et c'est, effectivement, le cas pour nos bases d'études : le géoréférencement, la taille et la forme du support spatial ont été établis par convention dès le commencement de la collecte de la base.

	1990	2000
Evolution fonctionnelle	 < FR-12354; 242; 53.305941673183>	 < FR-54356; 133; 53.305941673183 >
Evolution Spatiale	 <FR-43287; 141; 35.77294786396>	 <FR-22349; 141; 25.77294786396>
Evolution fonctionelle et espatiale	 <FR-45857; 141; 35.77294786396>	 <FR-03455; 142; 25.77294786396>

FIGURE 5.6 – Trois exemples d'évolutions distinguées entre les couches 1990 et 2000 dans les données *Corine Land Cover* [EEA, 2009].

5.3.2 Modèle des données

Dans cette section, nous essayons de donner un aperçu concret sur la base de données à utiliser pour la génération de règles d'évolution. Ainsi, nous présentons les modèles conceptuel et logique décrivant la structure de notre base d'apprentissage (cf. figure 5.7). Il convient de rappeler que l'objectif de ce modèle est de fournir une structuration relationnelle de données qui permet de décrire, au mieux, l'évolution de l'ensemble d'objets constituant un territoire. Par conséquent, la base devrait être capable de décrire les différentes entités impliquées dans les évolutions, leurs voisinages, les transitions qui correspondent aux changements de fonctions et les séquences d'évolution qui en sont composées. Ainsi le modèle conceptuel de notre base considère qu'une entité appartient exactement à une seule couche, peut avoir un ou plusieurs voisins et peut être impliquée dans une ou plusieurs transitions (par exemple, une entité qui subit une division est impliqué dans deux transitions

Algorithme 2 : Pseudo-code de l'algorithme permettant de déterminer le contexte spatio-temporel des objets géographiques

```
1  Données : D;
2  D : Ensemble de couches vectorielles;
3  LS = ∅ ;                                          // Liste des successeurs
4  LV = ∅ ;                                          // Liste des voisins
5  pour chaque objet O dans geoentity faire
6      CoucheC : Couche_Courante(O);
7      CoucheS : Couche_Successeure(CoucheC);
8      LV = RSpatiale.tengence (O);
9      pour chaque v dans LV faire
10         Remplir_Table_Voisinage(O, v);
11     LS = RSpatiale.intersection(O, CoucheS) ∪
         RSpatiale.couverture(O, CoucheS) ∪
         RSpatiale.chevauchement(O, CoucheS);
12     pour chaque s dans LS faire
13         Remplir_Table_Transition(O, s);
```

(cf. figure 5.8). De son côté, une transition peut appartenir à une ou plusieurs évolutions et une évolution contient au moins une transition.

Bien que simple, ce modèle permet de dévoiler les relations temporelles de succession d'états et les relations spatiales de voisinages que ce travail propose d'explorer afin de générer un modèle explicatif et éventuellement prospectif des dynamiques d'un territoire.

Cependant, les séquences telles qu'elles sont stockées dans la base sont difficilement exploitables pour la génération de règles d'évolution utiles et compréhensible et nécessite, donc, un prétraitement additionnel.

5.4 Représentation des données d'apprentissage

L'objectif de ce travail est de générer des règles d'association, dites règles d'évolution. Ces règles doivent être capables de souligner l'effet des relations spatio-temporelles sur l'évolution d'un territoire. En d'autres termes, elles

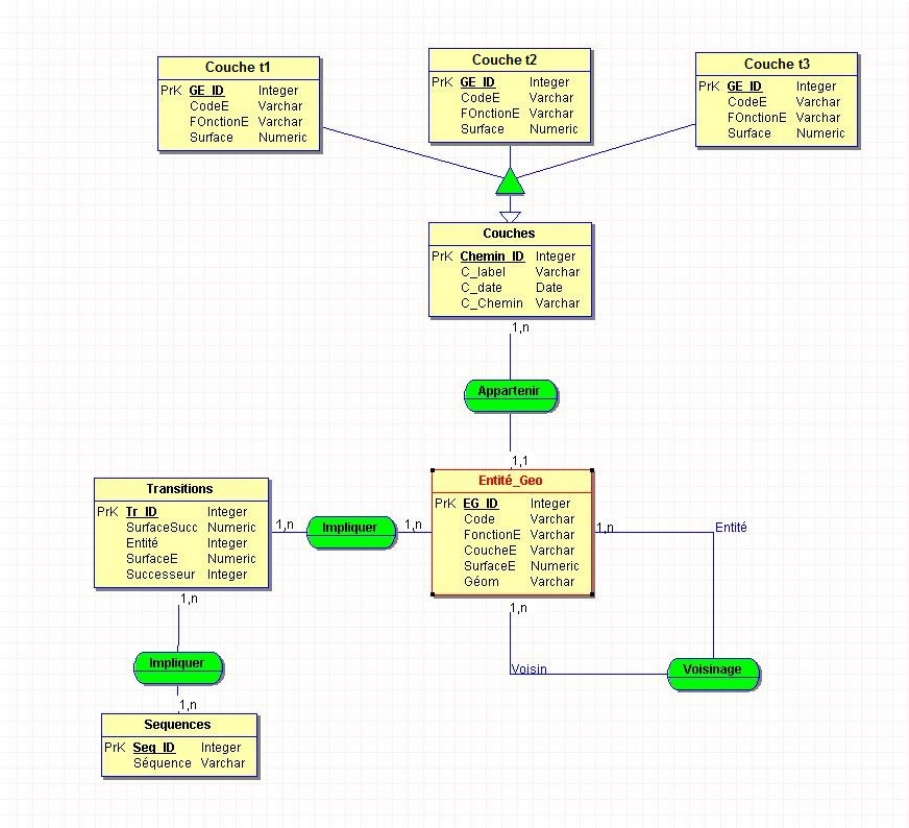


FIGURE 5.7 – MCD correspondant à la base d'apprentissage.

doivent porter dans leurs prémisses sur l'historique de ces relations et dans leurs conclusions sur l'évolution qui lui est, fréquemment, associée. Ainsi, dans notre contexte applicatif, nous définissons une règle cible comme une règle contenant dans son antécédent les éléments qui correspondent, respectivement, à l'historique de changements de fonction d'un objet, et à la configuration spatiale dans laquelle il se situe.

Dans cet objectif une étape de prétraitement est indispensable. Les données d'apprentissage doivent être structurées d'une manière à pouvoir générer ces règles. En effet, nos données d'apprentissage sont modélisées de façon à pouvoir déceler pour chaque cas (objet de référence) ses relations de voisinage et de succession permettant de construire sa trajectoire de vie. Étant nos instances d'apprentissage, ces trajectoires doivent être, explicitement, re-

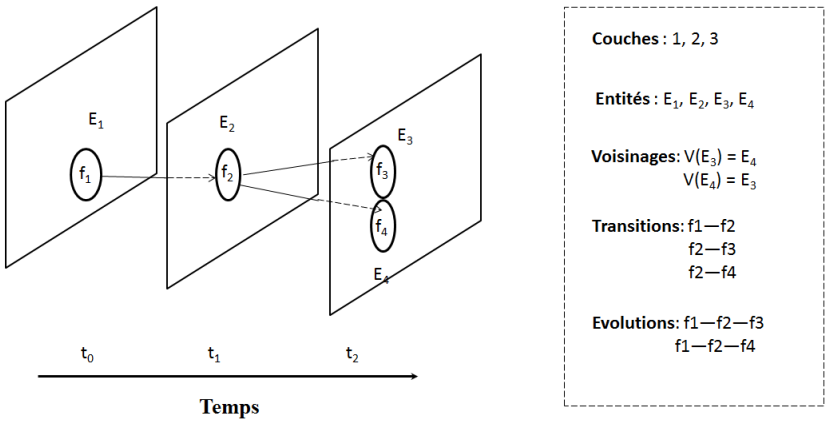


FIGURE 5.8 – Recensement d'une évolution dans la base d'apprentissage.

TABLE 5.1 – Sémantique des items

Symboles	Sémantique
N	La fonction d'un objet Voisin
SPF	Une séquence composée de fonctions des objets impliqués dans l'ensemble d'évolutions précédentes
S	La fonction de l'objet successeur

présentées sous format transactionnel. Les transactions sont formées par un ensemble d'items, les items correspondent aux valeurs des propriétés considérés lors de l'apprentissage. Dans notre cas, ces propriétés, appelés aussi attributs d'étude, correspondent à l'historique des évolutions des prédécesseurs noté *SPF* (*Sequence of Precedent Functions*), l'information sur leurs co-localisations notée N (*Neighbourhood*), et l'évolution future ou la fonction du successeur notée S (*Successor*).

Ainsi, notre transaction d'apprentissage est composée d'un et un seul item de type *SPF*, un et un seul item de type S et un ou plusieurs items de type N comme illustré dans l'exemple de la figure 5.9.

— L'item *SPF* correspond à la séquence de fonctions des entités modélisant son évolution jusqu'à son état actuel (séquence d'évolution). Représenté dans l'exemple, par la séquence composée par les fonctions

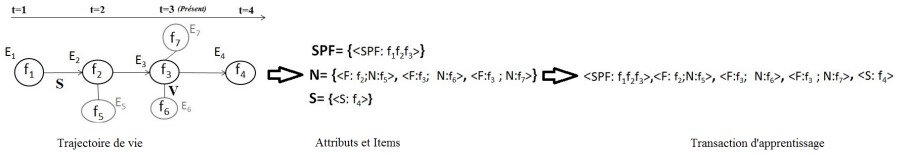


FIGURE 5.9 – La génération d'une instance d'apprentissage transactionnelle [Gharbi et al., 2016b].

f_1, f_2 et f_3. Celles-ci correspondent, respectivement, aux entités E_1, E_2, E_3 avec E_1 et E_2 le prédécesseur, E_3 la version actuelle de l'objet).

— Les items N correspondent aux fonctions des voisins de chaque entité impliquée dans la séquence d'évolution. Représentés dans l'exemple par « $< F : f_2; N : f_5 >$ », « $< F : f_3; N : f_6 >$ « et « $< F : f_3; N : f_7 >$ » avec : $< F : f_2; N : f_5 >$ représentant le voisin E_5 de l'entité E_2 à $t = 2$; $< F : f_3; N : f_6 >$ et $< F : f_3; N : f_7 >$ représentant les voisins E_6 et E_5 de l'entité E_3 à $t = 3$.

— L'item représentant l'attribut (S). Il correspondant à la probable fonction de l'entité successeure. Il est représenté dans l'exemple par « $< S : f_4 >$ ».

Formellement, une transaction peut être définie comme suit: Soit :

$$\left.\begin{array}{l} -I = \{i_1, i_2, ..., i_x\} \mid x \leq n \\ -S = \{s_1, s_2, ..., s_p\} \mid p \leq u \\ -N = \{ne_1, ne_2, ..., ne_q\} \mid q \leq v \\ -SPF = \{spf_1, spf_2, ..., spf_r\} \mid r \leq w \\ -Notons\ que\ I = \{S \cup SPF \cup N \mid S, SPF, N \subset I\} \\ -T = \{tr_1, tr_2, ..., tr_z\} \mid tr_z \subset I; z \leq y \end{array}\right\} tr_z = \{s_p, spf_r, \{neq_1, neq_2, ..., neq_y\}\} \mid s_p \in tr_z, spf_r, ne_q \in tr_z, y \geq 1 \quad (5.1)$$

La figure 5.10 illustre la structure tabulaire du fichier d'apprentissage. Les lignes correspondent aux instances d'apprentissage (trajectoires d'évolution) et les colonnes représentent les différents items qui correspondent aux variables SPF, N, F (cf. tableau 5.1).

Il convient de mentionner qu'une division ou une fusion à n'importe quel niveau temporel de l'évolution, implique la génération d'une nouvelle transaction, comme montré dans la figure 5.11.

Disposant à présent d'une structuration de données adaptée à notre problématique d'étude des dynamiques spatiales, le processus d'apprentissage

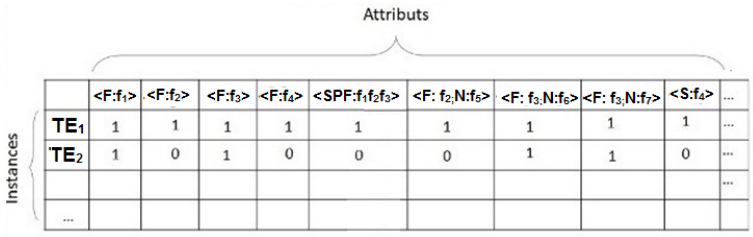


FIGURE 5.10 – Structure de la base d'apprentissage [Gharbi et al., 2016b].

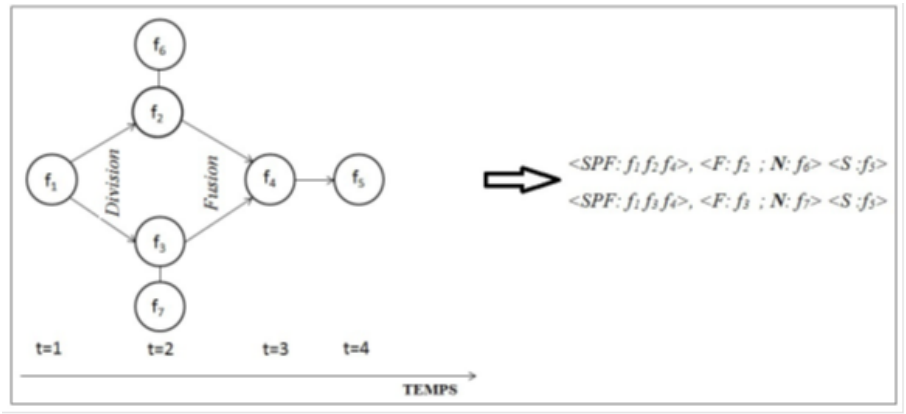


FIGURE 5.11 – Représentation transactionnelle des relations de divisions et de fusions [Gharbi et al., 2016c].

est supposé être capable de produire des règles d'association respectant la forme adéquate à la prédiction.

Dans l'objectif d'explorer dans quelle mesure l'étape précédente de prétraitement rempli cet objectif, nous avons procédé à une première application d'un algorithme classique de génération de règles d'association : Apriori. Lors de l'analyse des résultats préliminaires (voir figure 5.12), nous constatons qu'ils présentent une domination de l'attribut de voisinage par rapport aux autres attributs. Ce constat est valable tant pour les itemsets fréquents générés que pour les règles qui en découlent et qui sont, conjointement et exclusivement, formés, par des items de voisinage. Ceci révèle, en fait, un déséquilibre des données exprimé par un important écart de fréquence, dans la base, entre les items de voisinage (N) et les autres items (S, SPF) (cf.

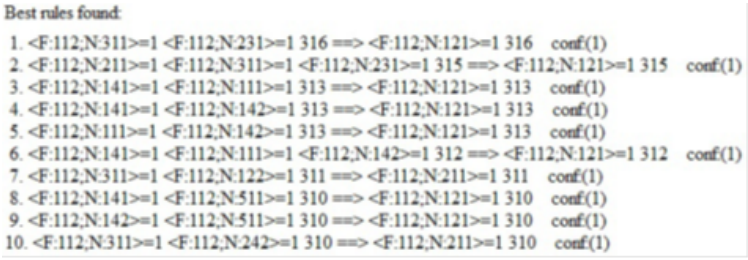

Best rules found:
1. <F:112;N:311>=1 <F:112;N:231>=1 316 ==> <F:112;N:121>=1 316 conf:(1)
2. <F:112;N:211>=1 <F:112;N:311>=1 <F:112;N:231>=1 315 ==> <F:112;N:121>=1 315 conf:(1)
3. <F:112;N:141>=1 <F:112;N:111>=1 313 ==> <F:112;N:121>=1 313 conf:(1)
4. <F:112;N:141>=1 <F:112;N:142>=1 313 ==> <F:112;N:121>=1 313 conf:(1)
5. <F:112;N:111>=1 <F:112;N:142>=1 313 ==> <F:112;N:121>=1 313 conf:(1)
6. <F:112;N:141>=1 <F:112;N:111>=1 <F:112;N:142>=1 312 ==> <F:112;N:121>=1 312 conf:(1)
7. <F:112;N:311>=1 <F:112;N:122>=1 311 ==> <F:112;N:211>=1 311 conf:(1)
8. <F:112;N:141>=1 <F:112;N:511>=1 310 ==> <F:112;N:121>=1 310 conf:(1)
9. <F:112;N:142>=1 <F:112;N:511>=1 310 ==> <F:112;N:121>=1 310 conf:(1)
10. <F:112;N:311>=1 <F:112;N:242>=1 310 ==> <F:112;N:211>=1 310 conf:(1)

FIGURE 5.12 – Un aperçu des résultats préliminaires.

tableau 5.1.

Cependant, une règle, ne comportant que des voisinages, ne prend pas en considération l'historique des évolutions et des voisinages des objets pour en déduire leurs évolutions. Ainsi, les règles générées ne sont ni pertinentes en matière de connaissances requises (informer sur l'évolution) ni conformes à notre hypothèse d'apprentissage qui portent sur le fait que la combinaison des successions des fonctions des objets et de leurs voisinages dans la prémisse devrait pouvoir nous guider sur la fonction de la conclusion.

En conclusion, bien que l'étape de prétraitement permette de considérer les attributs qui nous intéressent dans la tâche d'apprentissage, elle est insuffisante, sans autres développements sur les approches de fouille, pour faire face à nos différents défis, à l'instar de l'asymétrie présente dans les données qui en sont issues. Par conséquent, l'adaptation du processus même d'apprentissage s'avère nécessaire.

5.5 Adaptation du processus de fouille au problème de l'asymétrie

Dans la phase de fouille de règles d'évolution nous sommes partis de l'une des approches les plus populaires et fondamentales dans la génération de règles d'association à savoir l'approche Apriori et ses différentes améliorations (cf. section 4.3.2.2.2). Comme indiqué dans la section 4.3.2.2.1, Apriori procède en deux étapes. La première consiste à générer des itemsets candidats, les évaluer selon leurs fréquences dans la base, et puis les filtrer en supprimant ceux qui ne sont pas fréquents et la deuxième consiste à générer les règles d'associations confiantes à partir des itemsets restants (fréquents).

Étant utilisé pour l'évaluation de la fréquence des itemsets et donc pour la réduction de l'espace de recherche et ainsi pour la limitation du nombre de règles générées, le support minimum (**minsup**) représente l'élément clé de la première étape de la recherche de règles d'association. Cependant, employer un **minsup** unique assume, implicitement, que, dans la base, les fréquences des itemsets pertinents soient proches. Or, souvent, cette situation idéal ne se retrouve pas dans les applications réelles [Bhatt and Patel, 2014], et notamment la nôtre. En effet, au niveau de notre jeu de données, les items qui correspondent aux voisinages spatiaux des objets (*i.e.* les items étiquetés N) sont très fréquents par rapport aux autres types d'items (les items étiquetés S et les items étiquetés SPF). Ceci peut, ainsi, poser deux problèmes lors de la définition du **minsup** :

— si le **minsup** est élevé, seuls les items très fréquents (dans notre cas les items de type N) seront générés. En d'autres termes, les items rares mais pouvant être intéressants (S, SPF) seront omis et ne figureront pas dans les règles générées. Ceci représente, donc, une perte de connaissance et dégrade, ainsi, la qualité de l'apprentissage. Dans le contexte de notre application, les règles générées ne présentant pas des items de type S et SPF, ne sont donc pas conformes à la forme prédéfinie des règles d'évolution recherchées (cf. figure 5.12).

De ce fait, utiliser **un seul minsup** pour **tout** le jeu de données est inadéquat pour faire apparaitre les items rares qui nous intéressent. Partitionner les items en blocs intérieurement homogènes en matière de fréquence puis appliquer séparément l'algorithme de fouille peut paraître comme solution. Cependant ceci omet les corrélations pouvant exister entre les membres de différents blocs et donc entrave la génération de règles les impliquant.

Afin de traiter le problème d'asymétrie de données, nous proposons deux approches :

1. Une première approche qui consiste à un traitement « holistique » où toute la base est considérée. Dans ce cas, Apriori est modifié afin de tenir compte des différents niveaux de fréquence des items par l'utilisation de multiples **minsup**.
2. Une deuxième approche qui consiste en un traitement « en entonnoir » de la base d'apprentissage où à partir d'une base initiale, des sous-bases sont construites au fur et à mesure du traitement, selon les sémantiques des items, des différents niveaux de fréquence. Cette approche est représentée par l'algorithme BERA dont le fonctionnement

est explicité dans la section 5.5.2.

5.5.1 Supports multiples pour la génération des candidats complets

Dans cette approche, nous considérons la totalité du jeu de données d'apprentissage tout en spécifiant plusieurs supports minimums permettant de tenir compte des différents niveaux de fréquence des items. Ainsi, nous proposons d'appliquer une extension d'Apriori appelée MSA-priori [Liu et al., 1999]. Cette dernière permet d'utiliser différents **minsup** pour la génération d'itemsets fréquents (*i.e.* spécifier un **minsup** faible pour les items qui sont rares et un **minsup** élevé pour les items qui sont très fréquents).

Pour la définition et l'affectation de ces seuils, nous proposons et testons quatre méthodes décrites dans la section 5.1.2.

Tenant compte de la structure des règles d'évolution que nous cherchons à extraire, les itemsets pertinents, dits aussi complets, doivent être composés, exactement, d'un item de type *SPF*, un ou plusieurs items de type *N*, et exactement un item de type *S*. Ces itemsets garantissent, ainsi, la génération de règles complètes faisant apparaitre les trois attributs d'apprentissage. Donc, les règles ainsi obtenues intègrent les deux types de relations considérées : temporelles, par les items exprimant les successions de fonctions notamment les items de type *S* et *SPF*, et spatiales par les items de voisinage étiquetés *N*.

Bien qu'utiliser de différents **minsup** permette la génération d'itemsets complets et donc la génération de règles complètes, ces dernières peuvent ne pas être conformes à la forme que nous définissons et jugeons adéquates à l'explication voire la prédiction de l'évolution. En effet, certaines règles pourraient, par exemple, avoir des items de types *S* et *SPF* dans la prémisse et les items de type *N* dans la conclusion contrairement aux règles recherchées ($SPF, N \rightarrow S$). En d'autres termes, les attributs de la prémisse figurent au niveau de la conclusion et vice versa. Ci-après, nous donnons quelques exemples de règles complètes mais inadéquates pouvant être générées à partir d'un itemset complet.

$$S, SPF \rightarrow N$$
$$S, N \rightarrow SPF$$
$$N, N, N \rightarrow SPF, S$$

Afin de répondre à ces exigences, nous apportons deux modifications à MSApriori :

1. Dans la fonction de génération de candidats nous définissons deux contraintes permettant de générer, exclusivement, des itemsets complets. Un itemset est dit complet et est retenu s'il est de taille ≥ 3 et s'il est composé d'un item de type *SPF*, un item de type *S* et au moins un item de type *N*.
2. Dans la fonction de construction de règles, nous définissons une contrainte exigeant que la partie conclusion de la règle contienne exactement un item correspondant à l'attribut *S*.

Dans les sous-sections qui suivent (5.5.1.1 et 5.5.1.2), nous décrivons, respectivement le principe de MSApriori et les méthodes que nous proposons pour la définition des différents **minsup** et pour leur affectation aux items.

5.5.1.1 MSApriori

Dans cette extension d'Apriori, chaque item dispose d'un support minimum (SMI : Support minimum d'item) spécifié au préalable par l'utilisateur. Étant exprimé par les *SMI* des items qui la composent, le **minsup** d'une règle correspond à la valeur la plus petite de ceux-ci.

Soit :

$$R : spf, \{ne\} \rightarrow s$$

Tel que :

— $s \in S = \{s_1, s_2, ..., s_p\} \mid p \leq u$
— $ne \in N = \{ne_1, ne_2, ..., ne_q\} \mid q \leq v$
— $spf \in SPF = \{spf_1, spf_2, ..., spf_r\} \mid r \leq w$
— $I = \{i_1, i_2, ..., i_x\} \mid x \leq n$
— $I = \{S \cup SPF \cup N \mid S, SPF, N \subset I\}$

Dans le contexte de nos données, une règle notée r est dite fréquente si elle satisfait son **minsup**. Autrement dit, si son support dans la base est supérieur ou égal à :

$$min(SMI(s), SMI(spf), \{SMI(ne)\})$$

Ainsi, les règles impliquant des items fréquents auront des **minsup** élevés et les règles impliquant des items rares auront des **minsup** faibles. Par conséquent, les deux types de règles sont considérés comme fréquents et peuvent, ainsi, être retenus.

Bien qu'elle représente la propriété fondamentale de l'approche Apriori, la propriété d'anti-monotonie ne tient plus dans le modèle étendu.

Exemple 1 :

Soit :

- les items : spf_1, s_3, ne_1, ne_5
- $SMI(spf_1) = 10\%, SMI(s_3) = 20\%, SMI(ne_1) = 5\%, SMI(ne_5) = 6\%$
- le 2-itemset spf_1, s_3 formé lors de la première jointure avec $SMI(spf_1, s_3) = min(SMI(spf_1), SMI(s_3))$
- $Sup(spf_1, s_3) = 10\%$

Étant non-fréquent, l'itemset $\{spf_1, s_3\}$ est éliminé de la liste des itemsets fréquents F_k. Par conséquent, les 3-itemsets, $\{spf_1, s_3, ne_1\}$, $\{spf_1, s_3, ne_5\}$ ne peuvent pas être générés lors de la jointure des 2-itemsets fréquents. Or, ces items sont susceptibles d'être fréquents comme minsup(ne_1) est seulement de 5% et minsup(ne_5) est seulement de 6%. Bien que, dans cette logique, éliminer $\{spf_1, s_3\}$ est inappropriée, ne pas l'éliminer représente une violation de la propriété de l'anti-monotonie.

Afin de résoudre ce dilemme, MSApriori propose de trier d'une façon ascendante les *SMI* des items dans une liste T, en construire une liste L à partir de laquelle les 2-itemsets candidats seront générés. Cet ordre persiste dans toutes les étapes ultérieures de l'exécution de l'algorithme.

La construction de la liste L se fait comme suit :

- on identifie le premier élément i de la liste *SM* (ayant le plus petit *SMI*) dont le support satisfait son propre *SMI* (fréquent) et on l'introduit dans L.

— pour chaque élément j successeur à i dans T (*i.e.* $SMI(j) > SMI(i)$), si son support satisfait le *SMI* du premier élément i, j est introduit dans L.

L'exemple 2 nous montre comment cette solution permet de résoudre le problème explicité ci-dessus.

Exemple 2 : Soit :

— D est un dataset de 100 transactions.
— $Sup(ne_1) = 6$, $Sup(ne_5) = 3$, $Sup(spf_1) = 9$, $Sup(s_3) = 25$

Alors :

$T = \{ne_1, ne_5, spf_1, s_3\}$, $L = \{ne_1, spf_1, s_3\}$, $F_1 = \{ne_1, s_3\}$

L'élément ne_5 ne fait pas partie de L car $Sup(ne_5) = 3$ ne satisfait pas $SMI(ne_1) = 5$ (le premier élément de la liste triée T).

Les 2-itemsets candidats générés à partir de L sont : $\{ne_1, spf_1\}$, $\{ne_1, s_3\}$. L'itemset $\{ne_1, spf_1\}$ n'aurait pas fait partie des candidats si on avait utilisé la liste F_1 plutôt que la liste L car, selon la propriété de l'anti-monotonie, son sous-ensemble $\{spf_1\}$ ne fait pas partie de la liste des 1-itemsets fréquents. Or, cet itemset ($\{ne_1, spf_1\}$) peut être fréquent et forme avec $\{ne_1, s_3\}$ le 3-itemset fréquent $\{ne_1, spf_1, s_3\}$ qui peut aussi être fréquent.

Ainsi, utiliser la liste L, créée à partir de la liste triée T, pour la génération des 2-itemsets représente une solution au problème de départ. À l'encontre de la liste F_1, L contient les éléments qui ne satisfont pas leurs *SMI* mais qui peuvent satisfaire le *SMI* d'un élément prédécesseur dans la liste triée T (e.g. l'élément spf_1 ne satisfait pas son *SMI* mais satisfait le *SMI* de ne_1 et peut donc former avec celui-ci un itemset fréquent).

Dans l'algorithme 2, nous donnons le pseudo-code de l'algorithme MSA-priori qui décrit les différentes étapes de son exécution. Il commence par trier les items selon leurs *SMI* (ligne 1), en construit la liste L (ligne2) à partir de laquelle les candidats 2-itemsets seront générés (ligne 6) en utilisant la fonction CandidGen_Niveau2. Pour chaque balayage $k > 1$ l'algorithme effectue trois opérations.

1. Exécuter la fonction de génération de candidats de taille k à partir de la liste des itemsets fréquents F_{k-1}, c-à-d, les fonctions Candid-

Algorithme 3 : Pseudo-code de l'algorithme MSA-priori [Liu et al., 1999]

```
1  Données : D, SM, I ;     // D : une base de données, SM est la liste des minsup
2  T = Trier(I, SM) ;                          // Tri selon les SMI(i) stockés dans SM
3  L = ConstruireL(D, SM); F_1 = {{l} | l ∈ L, Sup(l)/n ≥ SMI(l)} ;   // n
     est le nombre de transactions de la base D
4  pour (k = 2; F_{k-1}! = ∅; k++) faire
5      si k = 2 alors
6          C_k = CandidGen_Niveau2(L);
7      sinon
8          C_k = CandidGenMS(F_{k-1});
9      pour chaque d ∈ D faire
10         pour chaque candidat c ∈ C_k faire
11             si c ∈ d alors
12                 Sup(c)++;
13             si c − {c[1]} ∈ d alors
14                 Supp(c − {c[1]}) + + ; // l'itemset candidat c sans son premier
                     item
15     F_k = {c ∈ Ck | Sup(c)/n ≥ SMI(c[1])};
16 Résultat : F = ∪kF_k ;
```

Gen_Niveau2 pour $k = 2$ et CandidGenMS (ligne8) pour $k > 2$, expliquées, respectivement, ci-après.

2. Balayer le dataset et mettre à jour les supports des différents candidats de C_k (lignes 9-14). Les lignes 13 et 14 représentent la mise à jour des supports des candidats sans leurs premiers items, une opération utilisée pour la génération des règles (cf. section 5.5.1.1.3).
3. Identifier les itemsets candidats fréquents et en construire la liste F_k (ligne 15).

5.5.1.1.1 CandidGen_Niveau2 Cette fonction prend en paramètre la liste L de 1-itemset et renvoie une liste des candidats 2-itemset (de taille 2).

5.5.1.1.2 CandidGenMs Cette fonction procède de la même manière que la fonction classique de génération de candidats de Apriori. Elle com-

mence par une étape de jointure puis par une étape d'élagage qui consiste à éliminer tout candidat c dont, au moins, un de ses sous-ensembles d'ordre $k-1$ n'est pas fréquents (*i.e.* ne figurent pas dans F_{k-1}). Cependant, cette deuxième étape (élagage) est un peu différente en ce qu'elle fait l'exception pour les candidats dont les sous-ensembles ne contiennent pas le premier item du candidat $c\,[1]$.

Exemple 3 :

— **Partie (a)**

Soit :

- $F_2 = \{\{a,b,v\},\{a,b,e\},\{a,v,d\},\{a,v,e\},\{b,v,e\}\}$
- $C_3 = \{\{a,b,v,e\},\{a,v,d,e\}\}$ après élagage $C_3 = \{\{a,b,v,e\},\{a,v,d,e\}\}$

Bien que le sous-ensemble $s = \{v,d,e\}$ de $c = \{a,v,d,e\}$ ne figure pas dans F_2, c n'est pas éliminé lors de l'élagage comme il ne contient pas $c\,[1]$ et donc on ne peut pas être sûr que c est non-fréquent. La 2ème partie de l'exemple apportera plus d'explication à ceci.

— **Partie (b)**

- SMI(v,d,e) = SMI(v)
- SMI(a,v,d,e) = SMI(a)
- SMI(a) $\leq$ SMI(v) (les SMI sont triés en ordre croissant au début de l'algorithme)

Par conséquent, même si v,d,e ne satisfait pas son SMI on n'est pas sûr qu'il ne satisfasse pas le SMI(a) et donc on n'est pas sûr que son sur-ensemble $\{a,v,d,e\}$ ne satisfasse pas SMI(a), *i.e.,* son SMI.

5.5.1.1.3 Génération de règles Dans Apriori, les règles d'association sont générées à partir des itemsets fréquents. Dans le cas d'un **minsup** unique, si f est un itemset fréquent et s est un sous-ensemble de f alors s doit aussi être fréquent. Leurs supports sont, ainsi, enregistrés par l'algo-

rithme et par conséquent la confiance de chaque règle possible à partir de f peut être calculée sans avoir à balayer, de nouveau, la base d'apprentissage. À l'encontre de ce modèle, dans le modèle à base de supports multiples, s peut ne pas être fréquent comme on a vu dans l'exemple 3. Ainsi son support n'est pas enregistré et le calcul de la confiance de toute règle l'impliquant peut poser un problème.

Exemple 4 : Soit : $SMI(a) = 3\%$, $SMI(b) = 0.25\%$, $SMI(c) = 0.15\%$

Si $f = \{c, b, a\}$ avec un support de 0.17%, $s = \{b, a\}$ avec un support de 0.10%, $\{c, b, a\}$ est fréquente et $\{b, a\}$ est non fréquente. Par conséquent, le support de s n'est pas enregistré et, ainsi, calculer la confiance de la règle $r : b, a \rightarrow c$ est difficile (*i.e.* $conf(r) = supp(a, b, c)/supp(c, b)$).

Ce problème peut se poser même avec les règles $b \rightarrow c$, a et $a \rightarrow c$, b comme nous ne pouvons être sûr que a et b sont fréquents.

En effet, ce problème se pose seulement quand la conséquence de la règle contient l'item ayant le SMI le plus faible (le premier item de f). Cette affirmation est prouvée par le raisonnement par l'absurde explicité ci-dessous [Liu, 2011].

Preuve :

— *Énoncé :*

Soit :
- f un itemset fréquent ;
- p le premier item de f ayant le SMI le plus faible dans f et donc $SMI(f) = SMI(p)$;
- la règle $r : X \rightarrow Y$ avec $X, Y \subset f$, $X \cup Y = f$, et $X \cap Y = \emptyset$.

Maintenant, Supposons que le problème exposé ci-dessus se pose aussi si $p \in X$.

— *Raisonnement :*

Puisque $p \in X$ et $X \subset f$ alors p doit avoir le SMI le plus faible dans X et X doit être fréquent (ceci est assuré par MSApriori). Ainsi, son SMI doit être enregistré. Ayant le support de f déjà enregistré dans la base (car

f est fréquent), la confiance de la règle r peut être calculé ce qui contredit la supposition du départ.

Afin de résoudre ce problème, MSApriori propose d'enregistrer les supports des items $f - f\,[1]$ (lignes 13 et 14). Par la suite, la génération des règles dans MSApriori se fait d'une façon similaire à Apriori.

Dans notre travail, notre objectif est de générer des règles visant à déduire les items de type S en se basant sur des items de type SPF et N. Dans ce contexte, deux contraintes sont à définir dans notre fonction de génération de règle GenR, à savoir, la conséquence de la règle doit contenir un seul item et l'item de la conséquence doit être de type S.

5.5.1.2 Méthodes d'affectation des supports minimums

Afin de définir les différents seuils de fréquences (**minsup**), nous adoptons une approche qui consiste à répartir les itemsets en groupes, à définir, pour chaque groupe, un seuil de fréquence, puis à affecter celui-ci à tous les items appartenant à ce groupe.

Dans ce contexte, nous proposons deux méthodes pour le partitionnement des items :

— une méthode exploitant les paramètres d'analyse statistique, en particulier, les indices de centralité des données.
— une méthode se basant sur un algorithme de clustering permettant de cerner les groupes homogènes selon certains critères de similarité.

5.5.1.2.1 Méthode à base des quartiles Une étude statistique, est une opération qui permet de caractériser une collection de valeurs numériques en utilisant des paramètres dits indices statistiques.

La première étape de l'étude statistique consiste à décrire les données tout en définissant la population d'étude, les caractères puis les séries statistiques. La population représente les individus sur laquelle porte l'étude. Les caractères représentent les variables étudiées (par exemple, la température, l'âge, les notes). Ces variables peuvent être quantitatives – ayant des valeurs numériques discrètes ou continues – ou qualitatives plaçant les individus (instance) dans une ou plusieurs catégories nominales si exprimées par des labels,

La Base d'apprentissage transactionnelle :

T_ID	Transition
1	<s_1>; <spf_1>; <n_1, n_2, n_3>
2	<s_1>; <spf_1>; <n_2, n_3>
3	<s_2>; <spf_1>; <n_1, n_3>
4	<s_3>; <spf_2>; <n_1>
5	<s_2>; <spf_2>; <n_1, n_3>

Population : { s_1, s_2, s_3, spf_1, spf_2, n_1, n_2, n_3 }

Caractère : Le nombre d'occurence dans la base

Série Statistique :

s_1	s_2	s_3	spf_1	spf_2	n_1	n_2	n_3
2	2	1	3	2	4	2	4

FIGURE 5.13 – Un exemple illustrant les éléments considérés lors d'une étude statistique.

et ordinale si elles ont un ordre inhérent (exemple : A, B, C .etc.). La série statistique, représente l'association entre les valeurs du caractère et les effectifs qui leurs correspondent. L'effectif d'une valeur d'un caractère étudié représente le nombre d'individus portant cette valeur (e.g. dans une classe d'élèves, 5 personnes ont eu une moyenne de 10. La moyenne est le caractère, 10 est une des valeurs de celui-ci et 5 est l'effectif cette valeur).

Dans le contexte de notre cas d'application, la population est l'ensemble des différents items présents dans la base d'apprentissage, le caractère est le nombre d'occurrence des différents items dans la base (fréquence) qui est une variable quantitative discrète et la série statistique correspond aux différentes valeurs de la variable fréquence associées chacune à son effectif (cf. figure 5.13).

Après avoir décrit les données comme expliqué ci-dessus, plusieurs indicateurs peuvent être calculés afin de saisir les tendances générales de ces données numériques.

Dans ce travail, notre objectif consiste à trouver des valeurs permettant de

séparer nos items selon leurs fréquences. Les indicateurs de centralité peuvent servir de limites entre les valeurs « faibles » et les valeurs « élevées ». La moyenne, et la médiane sont parmi les indicateurs les plus employés pour indiquer la centralité d'un jeu de valeurs. Cependant, la moyenne est particulièrement sensible à la dispersion des données. Dans des séries contenant des quantités extrêmement faibles par rapport aux autres valeurs, la moyenne se trouve loin du centre et donc ne remplit pas sa vocation d'indicateur de centralité. Dans ces cas, recourir à la médiane semble une solution, comme elle est un estimateur robuste de la position centrale dans un échantillon (insensible à la variabilité ou la dispersion des données). En effet, la médiane représente la valeur qui sépare les données en deux telle qu'une moitié de celles-ci a des valeurs qui lui sont inférieures et que l'autre moitié a des valeurs qui lui sont supérieures. Pour améliorer cette répartition, nous pouvons calculer la médiane de la première moitié et donc trouver la valeur qui limite les 25% des valeurs les plus faibles, et la médiane de la deuxième moitié et donc limiter les 25% des valeurs les plus élevé. Ces deux mesures sont appelées respectivement le quartile inférieur ou le premier quartile (Q_1) et le quartile supérieur ou le troisième quartile (Q_3). La médiane représente le deuxième quartile. Dans ce contexte, nous répartissons nos données (les fréquences des items dans la base) en quatre groupes de mêmes tailles selon quatre intervalles limités, respectivement, par la valeur la plus faible et le quartile inférieurs, le quartile inférieur et la médiane, la médiane et le quartile supérieur, et le quartile supérieur et la valeur maximale.

Une fois que le partitionnement des items est fait, une valeur **minsup**, pour chaque groupe, doit être spécifiée. Là aussi nous optons pour les quartiles. Ainsi, nous spécifions la médiane du groupe comme le **minsup** des éléments qui lui appartiennent.

Les Algorithmes 4 et 5 illustrent toutes les étapes de la méthode expliquée, ci-dessus.

Dans cette section, nous avons présenté une méthode basée sur une analyse statistique exploitant exclusivement les indices de centralité, les quartiles, pour partitionner les données selon leurs fréquences dans la base. Afin d'affiner l'opération de partitionnement nous proposons d'exploiter les algorithmes de clustering qui représentent des outils de groupement plus sophistiqués.

Algorithme 4 : Pseudo-code la méthode de partitionnement

Données : D;
Ls: liste des items de taille 1;
$G1, G2, G3, G4 = \emptyset$;
// Partionnement de la liste des items
Incrémenter(D, Ls);
M=Mediane(Ls);
$Q1$=Qinf(Ls);
$Q2$=Qsup(Ls);
pour ***chaque*** $l \in Ls$ **faire**
    si $l.supp \leq Q_1$ **alors**
        $G_1 = G_1 \cup l$;
    si $l.supp > Q_1$ *et* $l.supp \leq M$ **alors**
        $G_2 = G_2 \cup l$;
    si $l.sup > M$ *et* $l.sup \leq Q_3$ **alors**
        $G_3 = G_3 \cup l$;
    si $l.sup > Q_3$ **alors**
        $G_4 = G_4 \cup l$;
Résultat : $G = G_1 \cup G_2 \cup G_3 \cup G_4$;

5.5.1.2.2 Méthode par algorithme de clustering Afin de réaliser un partitionnement plus fin, nous pouvons explorer l'effet d'autres caractéristiques des items et d'autres notions, à part la centralité des données, sur le partitionnement. Nous pouvons également, considérer individuellement les items et les comparer mutuellement selon leurs caractéristiques afin de les regrouper par similarité.

Ceci, en fait, correspond au clustering représentant une méthode d'exploration de données qui émane aussi de l'analyse statistique. Le clustering consiste à grouper les objets d'un jeu de données, uniquement, en fonction des informations trouvées dans les données qui décrivant ces objets et leurs relations. L'objectif est que les objets d'un même groupe soient semblables (ou liés) entre eux et diffèrent (ou non liés) des objets appartenant aux autres groupes. Plus la similitude (similarité ou l'homogénéité) est importante au sein d'un groupe et plus la différence (dissimilarité ou hétérogénéité) entre les groupes est grande, meilleur et plus distinct est le regroupement. Tracées géométriquement, les objets dans un même cluster seront fermés ensemble

Algorithme 5 : Pseudo-code d'affectation de minsup

```
1  Données : G;
2  pour chaque G_i ⊂ {G | i ∈ {1, ..., 4}} faire
3      M_i = Mediane(G_i);
4      si i == 4 alors
5          Q_i3 = Qsup(G_i);
6          ;  // Calculer le quartile supérieur du groupe des items les plus fréquents
               G_4
7          pour chaque l ∈ G_i faire
8              SMI(l) = Q_i3;
9              LSMI = l
10     sinon
11         pour chaque l ∈ G_i faire
12             SMI(l) = M_i;
13             LSMI = l
14 Résultat : LSMI ;
15 ;                                   // La liste des items et leurs minsup
```

tandis que la distance entre les groupes sera plus éloignée. Le clustering peut être réalisé par divers algorithmes qui diffèrent considérablement selon les notions employés pour la définition des clusters (e.g. des petites distances entre les éléments d'un même cluster, une densité importante d'une zone de de données qui définit un cluster) et les façons de les trouver efficacement. L'algorithme de clustering approprié et le réglage des paramètres (e.g. la fonction de distance à utiliser, un seuil de densité ou le nombre de clusters attendus) dépendent de l'ensemble des données individuelles et de l'utilisation prévue des résultats.

Dans notre cas d'application, nous visons à grouper les items similaires du point de vue fréquence dans la base d'apprentissage. Ainsi, la fréquence des items constitue un attribut indispensable pour le clustering qui peut être enrichie par d'autres attributs tel que la sémantique de l'item (item de type S, SPF ou N). Le nombre de groupe peut être spécifié à 2 comme nous visons à traiter séparément les items fréquents et les items rares. Cependant, un écart de fréquence peut exister au sein de ces deux groupes mêmes. Ainsi, utiliser des algorithmes capables de définir par eux-mêmes le nombre de groupe à identifier selon les données, ou ceux dont le fonctionnement

ne nécessite pas la spécification de ce paramètre, nous semble plus approprié. Parmi les plus populaires de ces algorithmes nous trouvons l'algorithme EM, les algorithmes de clustering à base de la densité tels que DBSCAN, OPTICS, et les algorithmes de groupement hiérarchique tels qu'AGNES et DIANA [Patel and Thakral, 2016].

Par analogie à la méthode par quartiles, une fois le groupement fait, nous affectons les médianes de l'ensemble des fréquences des éléments au sein d'un même groupe comme leurs SMI.

5.5.1.2.3 Vers la prise en compte de la sémantique des prédicats Visant à générer des règles informant sur l'évolution d'un objet géographique (en termes de leurs fonctions), nous nous sommes basés sur l'historique des relations spatiales et temporelles reliant ses différentes versions au cours de sa trajectoire de vie. Ainsi, les prédicats de nos règles cibles sont définis selon une sémantique où : un ou plusieurs items, notés N, représentent les voisins d'un objet ; un item, noté SPF, représente la séquence décrivant ses fonctions au cours du temps ; et un item, noté S, représente sa fonction successeure. C'est cette sémantique que nous proposons de considérer afin d'explorer son effet sur la performance des deux méthodes proposées pour l'affectation des différents **minsup**, soit la méthode par quartiles et la méthode par clustering.

1. Au lieu de partitionner les items selon des intervalles de fréquences définis par les quartiles, nous proposons de les partitionner selon leurs sémantiques (S, SPF, N, cf. tableau 5.1) puis utiliser les quartiles pour définir le **minsup** correspondant à chaque groupe. En d'autres termes, définir trois groupes : les éléments étiquetés S, les éléments étiquetés SPF, les éléments étiquetés N, puis calculer, pour chaque groupe, l'indice de centralité adéquat (médiane, quartile supérieur, ou quartile inférieur) et l'affecter comme **minsup** à tous ses membres. Afin de générer plus d'itemsets impliquant les items correspondant aux attributs faiblement représentés dans la base (S, SPF), nous devons maximiser le nombre d'item générés de type S et SPF au détriment des items de voisinage (N). Ainsi, nous définissons, pour les groupes correspondant à ces deux types, la médiane comme leurs **minsup** respectifs et le quartile supérieur comme le **minsup** du groupe des items de type N. Par conséquent, 50% des items de types S et 50% des items de type SPF et seulement 25% des items de type N seront considérés comme fréquents et donc seront générés lors de la première itération de

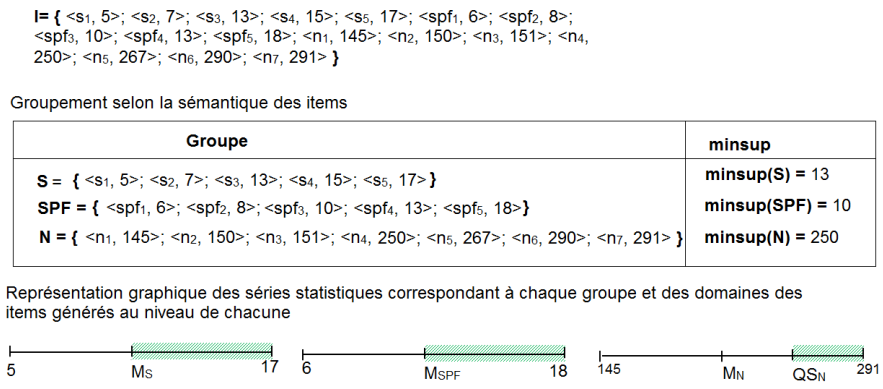


FIGURE 5.14 – Un exemple illustrant les étapes de la génération d'items fréquents selon la méthode par quartile considérant la sémantique de ceux-ci.

MSApriori. La figure 5.14 représente un exemple illustrant les étapes décrites ci-dessus.

2. Pour cette méthode, la considération de l'aspect sémantique se fait simplement par l'ajout d'un attribut de clustering décrivant la sémantique de chaque item (*i.e.* spécifier si c'est un item de type N, de type SPF, ou de type S).

5.5.2 Algorithme BERA

Afin de répondre au défi de l'asymétrie de données, nous avons, dans une première proposition, opté pour une approche holistique. En d'autres termes, nous avons considéré la totalité du jeu de données pour évaluer la fréquence des items tout en spécifiant plusieurs supports minimums. Ceci permet, ainsi, de faire apparaitre les items considérés initialement comme rares conjointement avec les items fréquents, et donc permet de générer des itemsets plus pertinents. Dans une deuxième proposition, nous explorons une autre approche qui consiste à extraire, à partir d'un jeu de données initial (base des transactions sur les transitions observées d'un territoire), des sous-ensembles de données (ou transactions) correspondant chacun à un niveau de fréquence différent. Nous postulons que chaque attribut d'apprentissage représente un niveau de fréquence et peut donc être traité localement dans le sous-ensemble qui lui correspond.

En d'autres termes, pour chaque attribut $(SPF, NetS)$, les items lui correspondant seront évalués selon un **minsup**, propre à l'attribut et au sous-ensemble en cours de traitement, pour en extraire les items fréquents.

Ce sont ces items qui permettront ainsi de filtrer leur base pour en définir une nouvelle utilisée pour traiter un autre attribut présentant un autre niveau de fréquence (*i.e.* supprimer les transactions qui n'implique pas un des items fréquents trouvé). Ainsi, le sous-ensemble final, obtenu après le traitement du dernier attribut, ne sera constitué que par des transactions composées exclusivement par des items fréquents. C'est à partir de ces transactions que les règles sont construites.

Afin d'implémenter cette approche, nous avons proposé l'algorithme BERA (*Backward Extraction Rule Algorithm*) dont le nom émane de l'approche qu'il adopte lors de son fonctionnement (cf. algorithme 6). En fait, cet algorithme génère les règles en remontant en arrière à partir de leurs objectifs ou conclusion pour déterminer leurs prémisses. En d'autres termes, il commence par chercher, dans la base initiale, les items fréquents correspondant à l'attribut de la conclusion puis identifie les transactions contenant un de ces items et en construit une nouvelle base transactionnelle. Par la suite, il répète cette opération pour chaque attribut destiné à figurer dans la partie prémisse jusqu'à obtenir, à la fin, une base de transactions constituées exclusivement par des items fréquents. La mise à jour des supports des items se fait, à chaque itération, selon la nouvelle base traitée. Le **minsup**, est défini par la médiane des fréquences (support) des singletons de cette base.

Dans notre cadre d'application (cf. figure 5.16), nous commençons par trouver les items fréquents qui correspondent à l'attribut de la conclusion, l'attribut S (cf. bloc 1 dans algorithme 6). Nous récupérons les transactions qui contiennent un de ces items et nous en construisons une deuxième base d'apprentissage (DataS dans figure 5.16). Par la suite, nous enchainons avec les attributs de la partie prémisse : SPF et N (cf. bloc 2 dans algorithme 6). Ainsi, nous identifions la liste des items fréquents de type SPF parmi les transactions de la nouvelle base ($DataS$) (dont les transactions contiennent exclusivement des S fréquents), nous récupérons les transactions qui contiennent un de ces items puis nous en construisons une autre base transactionnelle ($DataSPF$). Correspondant à plusieurs items dans la règle à générer, le traitement de l'attribut N est un peu différent (cf. figure 5.15) : considérons la dernière base générée, nous récupérons l'ensemble d'items correspondant à l'attribut N, nous identifions parmi ceux-ci l'ensemble des combinaisons de N qui sont fréquentes puis pour chacune d'elle nous définissons une nouvelle

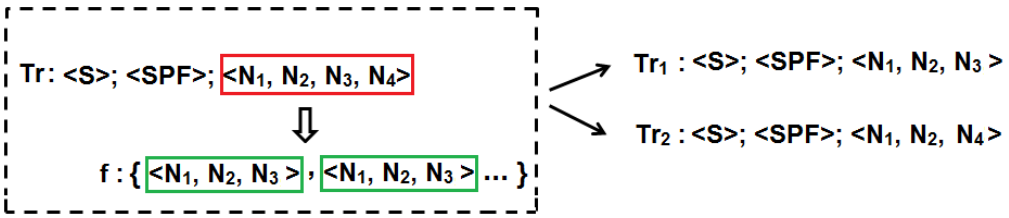


FIGURE 5.15 – Un exemple illustrant la construction des transactions lors du traitement d'un attribut, représenté dans la règle par plusieurs items. Tr est la transaction de départ, f est la liste des itemsets fréquents correspondant à l'ensemble d'item N et tr_1, tr_2 représentent les transactions construites.

transaction que nous ajoutons dans la nouvelle base.

Représentant la base construite lors du traitement du dernier attribut (N), $DataN$ correspond à la base finale, à partir de laquelle nos règles seront construites.

Disposant d'un ensemble de transactions qui chacune contient un item SPF, un item S et des items N fréquents, les règles sont construites comme indiqué dans l'exemple suivant.

Exemple :

Considérons la première transaction de la base résultante « dataN » (cf.figure 5.16) :

$$< s_1, spf_1, n_1 >$$

La règle à générer à partir de cette transaction est la suivante :

$$spf_1, n_1 \rightarrow s_1$$

Afin de filtrer les règles non confiantes, nous employons la mesure de confiance. Une règle est éliminée si la valeur de sa confinace ne satisfait pas un seuil prédéterminé. La confiance de toute règle générée est calculée selon la base initiale.

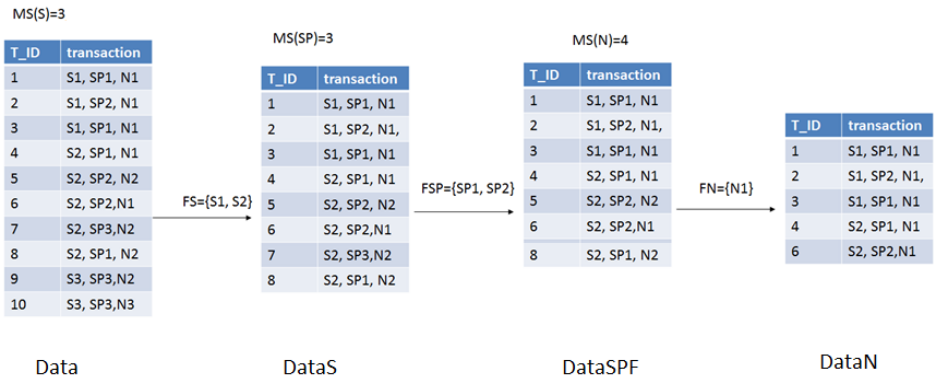


FIGURE 5.16 – illustration de l'exécution de BERA sur un exemple de dataset jeu de données avec $Data$: la base initiale, $DataS$: la base des transactions contenant des s fréquents, $DataSPF$: la base des transactions contenant des spf fréquents, $DataN$: la base des transactions contenant des n fréquents.

5.6 Conclusion

Dans le contexte de l'évolution territoriale, notre travail vise à proposer une approche automatisée qui exploite la fouille de données pour produire un ensemble de règles en régissant les dynamiques. Représentant un aperçu général de l'approche proposée, ce chapitre expose les différents défis rencontrés ainsi que les propositions faites pour répondre à ceux-ci.

Dans ce travail, nous sommes partis de l'hypothèse que l'évolution territoriale peut en partie être expliquée à travers l'historique des changements de fonction de ses entités géographiques et des configurations spatiales dans lesquelles elles se situent. Ainsi, pour répondre à la question de l'autocorrélation spatiale, nous proposons un modèle conceptuel pour la saisie des relations spatiales de voisinage et temporelles de succession de fonctions des objets étudier. Celles-ci nécessitent un prétraitement supplémentaire afin qu'elles soient représentées sous un format permettant la génération d'une forme particulière de règles que nous proposons et jugeons adéquate pour notre mission de départ (l'explication voire la prédiction de l'évolution territoriale). Bien que ce jeu de données paraisse prêt pour l'extraction de ce genre de règles, celui-ci s'est avéré déséquilibré. Une première application de l'algorithme apriori a généré des règles impliquant seulement les relations de voisinage ce qui

s'explique par la domination des items correspondant à ces relations par rapport aux items correspondant aux autres types de relations (e.g. relations temporelles de changement de fonctions). Dans ce contexte, nous essayons, dans une troisième étape, de traiter ce problème d'asymétrie de données en proposant deux approches :

1. Une première approche qui consiste à un traitement « holistique » où toute la base est considérée. Dans ce cas, Apriori est modifié afin de tenir compte des différents niveaux de fréquence des items par l'utilisation de multiples **minsup**.
2. Une deuxième approche qui consiste à un traitement « en entonnoir » de la base d'apprentissage où à partir d'une base initiale, des sous-bases sont construites au fur et à mesure du traitement, selon les sémantiques des items, des différents niveaux de fréquence. Cette approche est représentée par l'algorithme BERA.

Dans le chapitre suivant, ces différentes propositions subiront une étude expérimentale dont l'objectif est de les évaluer en fonction de leurs capacités à répondre aux défis de notre thèse.

Algorithme 6 : Pseudo-code de l'algorithme BERA

```
Données : Data ;                                          // Base transactionnelle initiale
DataC = ∅ ;                                               // Nouvelle base transactionnelle
DataR = ∅ ;                                               // La base transactionnelle résultante
Ac = ∅ ;                                                  // Attribut à figurer dans la conclusion de la règle
IAP = ∅ ;                                                 // Liste des attributs à figurer dans la prémisse
R = ∅ ;                                                   // Liste des règles générées
// Bloc1 : traiter l'attribut de la conclusion
Ic = Singletons (Data, Ac) ; // liste des singletons de Data correspondant à l'attribut Ac
Ic_f = {{ic} | ic ∈ Ic, ic.support ≥ SupAc}; ;            // Filtrer Ic selon SupAc, avec SupAc la Médiane correspondant à l'ensemble Ic
pour chaque t ∈ Data faire
    ic = trouverItemSelonAttribut(t, Ac) ; // trouver l'item de t correspondant à l'attribut Ac
    si ic ∈ Ic_f alors
        DataC = DataC ∪ t;
// Bloc 2 : traiter les attributs de la prémisse
DataOldAtt = DataC;
pour chaqueAp ∈ LAP faire
    DataNewAttr = ∅;
    si condition 1 ; // condition 1 : Ap est représenté par un seul item dans les règles à générer
    alors
        ;                                                 // Sous-boc 2.1
        Iap = Singletons(DataNew, Ap) ; // liste des singletons de DataNew correspondant à l'attribut Ap
        Iap_f = {{iap} | iap ∈ Iap, iap.support ≥ SupAp} ;   // Filtrer Iap selon SupAp, avec SupAp la médiane correspondant à l'ensemble Iap
        pour chaque t ∈ DataOldAtt faire
            iap = trouverItemSelonAttribut(t, iap) ;  // trouver l'item de t correspondant à l'attribut Ap
            si iap ∉ Iap_f alors
                DataNewAttr = DataNewAttr ∪ t;
    sinon si condition 2 ;    // condition 2 : Ap est représenté par plusieurs items dans les règles à générer
    alors
    // Sous-boc 2.1
    DataAp = ∅ ;                                          // une base transactionnelle
    pour chaque t ∈ DataOldAtt faire
        tAp = ConstruireTransaction(t, Ap) ; // une transaction constituée par les items de t correspondants à l'attribut Ap
        DataAp = DataAp ∪ tAp;
    LFAp = TrouverItemSetFréquent(DataAp, SuppAp) ;    // liste des itemsets fréquents de type AP dans DataAp
    pour chaque t ∈ DataOldAtt faire
        Lt = TrouverListeDeTransaction(t, LFAp);
        DataNewAttr = DataNewAttr ∪ Lt;
    DataOldAtt = DataNewAttr;
// Bloc 3 : construire la liste des règles
DataR = DataNewAttr;
pour chaque t ∈ DataR faire
    R = {R ∪ t | R.conf ≥ minConf};
Résultat : R;
```

Chapitre 6

Dispositif Expérimental, résultats et discussions

6.1 Introduction

Le présent chapitre correspond à une étude expérimentale dont l'objectif est d'évaluer les différentes propositions faites en fonction des défis de notre thèse. Ainsi, dans un premier temps, nous présentons le dispositif SAFFIET [Gharbi et al., 2016a] développé afin de mettre en oeuvre nos propositions. Dans un deuxième temps, nous mettons en exergue les paramètres à employer pour l'analyse de celles-ci. Dans un troisième temps, nous exposons les jeux de tests, les résultats générés, les valeurs des paramètres employés ainsi que leurs significations en termes d'évaluation de l'apport de nos propositions pour la résolution des défis de départ.

6.2 Dispositif expérimental

Dans cette section, nous présentons un prototype de notre dispositif expérimental appelé SAFFIET (*Spatial And Functional Frequent Itemset Extraction Tool*). Celui-ci permet, essentiellement, l'extraction de règles d'évolution d'un territoire géographique à partir d'une série temporelle de cartes vectorielles le décrivant.

Il représente une implémentation des différentes propositions faites en vue de répondre aux défis liés à cette tâche. À savoir :

— Une méthode permettant l'identification des objets géographiques

ainsi que leurs caractéristiques – attributaires, spatiales et temporelles – et l'extraction et la modélisation de leurs relations spatiales et temporelles afin de suivre leurs évolutions.

— Une méthode de prétraitement de ces relations permettant leur structuration dans une base d'apprentissage, sous un format prédéfini. Étant exploitable par les algorithmes de fouilles de règles, ce format permet, ainsi, la génération de règles, dites d'évolution. Cette méthode définit également une sémantique, pour les prédicats des règles à générer, jugées adéquates à des fins explicatives et prédictives du phénomène de l'évolution territoriale. Cette sémantique, compréhensibles par l'utilisateur, donne aux règles le potentiel d'être utilisées pour d'autres tâches telles que l'aide à la décision et l'aménagement urbain.

— Un module exploitant la base d'apprentissage, ainsi générée, pour extraire les règles d'évolution. Ce module englobe toutes les propositions faites afin de traiter l'asymétrie inhérente aux attributs d'apprentissage (N, SPF, S) : soit, l'algorithme MSApriori avec affectation de seuils à base de clustering, MSApriori avec affectation de seuils à base de quartiles, leurs variantes considérant la sémantique des items et l'algorithme BERA.

— Un module d'évaluation qui permet d'évaluer les règles formant le modèle d'apprentissage généré à travers des indicateurs fixés (cf. section 6.5).

SAFFIET a été développé en Python (v. 2.7.9). Il exploite les fonctionnalités des différentes solutions existantes à savoir, QuantumGIS (QGIS), PostgreSQL (v. 9.4) et son extension PostGIS, et Weka (3.6.13). Ainsi, il fait collaborer les interpréteurs de ligne de commande Bash (v. 4.3.30), SimpleCLI (v. 3.6) de Weka et psql de PostgreSQL pour exécuter des commandes systèmes, configurer et appeler les algorithmes de fouille de données et gérer le chargement des données géographiques dans la base de données. Le système de gestion de base de données utilisé est PostgreSQL, muni de son extension spatiale PostGIS. La lecture des données se fait par une connexion à la base via le module psycopg qui adapte automatiquement les types Python aux types PostgreSQL. QGIS servira pour la visualisation des résultats d'un éventuel module de prédiction qui, appliquant les règles générées sur la carte actuelle d'un territoire, estime et affiche sa carte future.

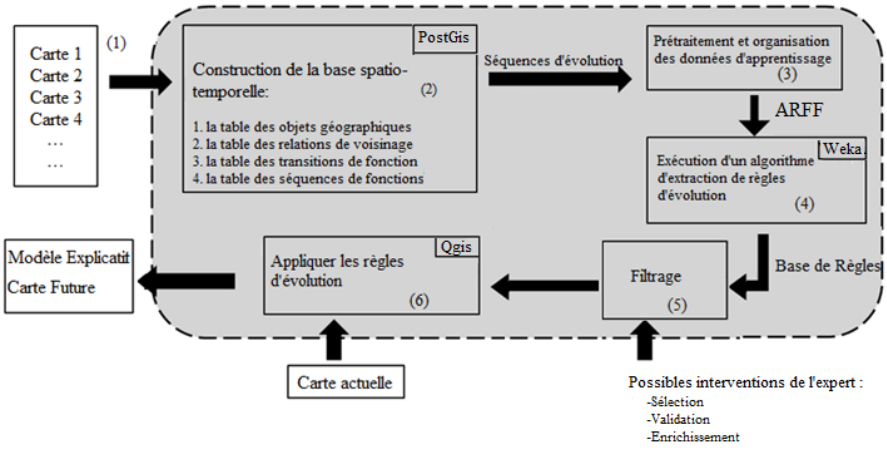


FIGURE 6.1 – Chaine de traitements pour l'extraction de règles spatiotemporelles en vue de la caractérisation de modèles explicatifs et potentiellement prédictifs des évolutions d'un territoire: implémentation.

6.2.1 Suivi de l'évolution

Afin d'accomplir cette tâche nous adoptons une conceptualisation des données proposée par Cheylan et Lardon [Cheylan and Lardon, 1993], se basant sur le paradigme identitaire pour l'identification des objets géographiques. Celui-ci propose l'affectation aux objets comme à ses atomes (des entités représentant des versions spatio-temporelles d'un objet de référence), des identifiants qui persistent tout au long de leurs évolutions. Ces identifiants permettent ainsi d'établir le lien entre un objet de référence et ses versions spatio-temporelles et donc rend possible la construction de sa trajectoire de vie à l'aide de requêtes SQL.

Dans ce contexte, l'outil SAFFIET fournit un menu (cf. figure 6.2) permettant aux utilisateurs de charger des fichiers *shapefile* correspondant aux cartes vectorielles de la série temporelles à étudier (cf. étape (1) dans la figure 6.1). Celles-ci sont, par la suite, converties, importées et stockées temporairement, sous un format compréhensible par l'extension PostGIS de PostgreSQL. Ceci est réalisé à l'aide des commandes shp2pgsql [1] et psql [2] dont les

1. http://suite.opengeo.org/docs/latest/dataadmin/pgGettingStarted/shp2pgsql.html
2. http://postgresguide.com/utilities/psql.html

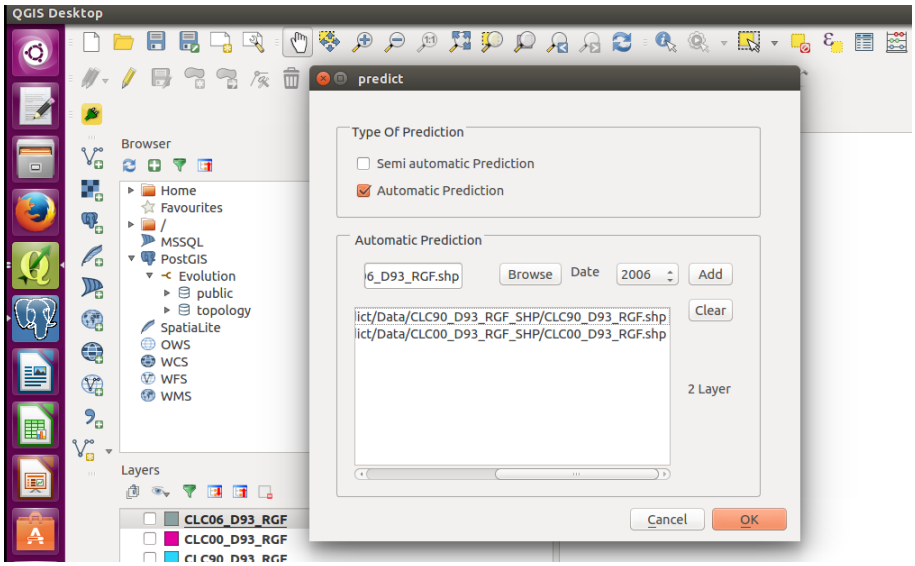


FIGURE 6.2 – Interface graphique de SAFFIET [Gharbi et al., 2016a].

```
shp2pgsql -I -s <SRID> <PATH/TO/SHAPEFILE> <DBTABLE>  >  SHAPEFILE.sql

psql -U postgres -d <DBNAME > -f SHAPEFILE.sql
```

FIGURE 6.3 – Les syntaxes des commandes *shp2pgsql* et *psql*.

syntaxes et des exemples sont, respectivement, illustrés dans les figures 6.3 et 6.4.

Une fois que les cartes sont importées dans la base, une requête SQL extrait et stocke dans une table tous les objets présents dans les cartes en affectant pour chacun un identifiant unique en précisant sa fonction, sa géométrie et la couche à laquelle il appartient. Afin de définir leurs contextes spatiaux, SAFFIET identifie, pour chaque objet, ses différents voisins (par adjacence), à l'aide de la requête *ST_Touches*. Pour identifier les relations temporelles d'évolution, notre outil superpose (combine) deux à deux les cartes de la série temporelle allant de la carte la plus ancienne à la carte la plus récente. Par la suite, il emploie les requêtes *ST_Intersects*, *ST_Overlaps*, *ST_Covers*

```
shp2pgsql -I -s 2154 ./.qgis2/python/plugins/predict/Data/CLC90_D93_RGF_SHP/CLC90_D93_RGF.shp clc90_d93_rgf > clc90_d93_rgf.sql

psql -U postgres -d Evolution -f clc90_d93_rgf
```

FIGURE 6.4 – Un exemple de l'utilisation des commandes *shp2pgsql* et *psql* pour importer la carte du département 93 en France, à la date 1990.

```
SELECT "+layer1+".gid as gidc1, "+layer2+".gid as gidc2, "+layer1+".area_ha as area1, "+layer2+".area_ha as area2 FROM
"+layer1+" INNER JOIN "+layer2+" ON ST_Intersects ("+layer1+".geom, "+layer2+".geom) where
ST_Overlaps("+layer1+".geom, "+layer2+".geom) or ST_CoveredBy("+layer1+".geom, "+layer2+".geom) or
ST_Covers("+layer1+".geom, "+layer2+".geom);
```

FIGURE 6.5 – La requête déterminant la continuité spatiale entre deux cartes temporellement consécutives dans la série temporelle d'étude.

(cf. figure 6.5), couvrant tout type d'intersection, pour déterminer les continuités spatiales (couple d'objets dont l'intersection est non nulle) entre les objets stockés aux différentes dates (cf. algorithme chapitre refalgo:algo1). Par exemple, un objet ID_2 de la carte 2000 dont l'intersection avec l'objet (ID_1) de la carte 1990 est non nulle, représente le successeur de l'objet ID_1 dans un changement possible d'usage. L'ensemble des couples extraits forment la table « Transitions », à partir de laquelle une table « Séquences » (trajectoires) sera construite (cf. étape (2) dans la figure 6.1).

Notons bien que cette approche de suivi suppose que le support spatial est stable dans le temps, ce qui est le cas pour nos données d'étude issues de la base CLC.

6.2.2 Construction du fichier d'apprentissage

Disposant d'une base de données modélisant les objets et leurs contextes spatio-temporels, nous passons au processus d'apprentissage. La première étape de celui-ci consiste à construire la base d'apprentissage qui représente, dans notre cas, un fichier sous format tabulaire dit ARFF (cf. étape 3 dans la figure 6.1). Dans ce contexte, notre système emploie des requêtes SQL pour extraire les trajectoires de vie des objets et les traduire, par la suite, en transactions. C'est à partir de ces dernières que sont déterminées les valeurs des attributs d'apprentissage qui les vérifient. Concrètement, SAFFIET, applique

des requêtes pour trouver les séquences d'évolution. Pour chaque séquence, il extrait la sous-séquence allant du premier élément à l'avant-dernier l'élément, correspondant à la valeur de l'attribut SPF ; il parcourt la séquence courante et exécute pour chacun de ses éléments une requête de même type, sur la table voisinage, servant à identifier les voisins de celui-ci. Ceux-ci représentent les items correspondant à l'attribut N.

Ainsi, le fichier ARFF peut être généré, avec comme lignes les transactions et comme colonnes les items qui les constituent. Un item vérifiant une transaction prend la valeur « 1 » et la valeur « ? » sinon.

6.2.3 Génération de modèle d'apprentissage

La génération des règles d'évolution dans SAFFIET (cf. étape 4 dans la figure 6.1) se fait sous Weka via une ligne de commande (pour le cas de l'approche BERA, se fait directement par l'algorithme BERA). Celle-ci spécifie le fichier ARFF (précédemment généré) représentant la base d'apprentissage et une valeur correspondant au seuil de confiance des règles à générer. Les règles d'évolution produites pourront subir une étape de filtrage par un expert (cf. etape 5 dans la figure 6.1), avant d'être appliquées pour l'explication et/ou la prédiction de l'évolution territoriale (cf. etape 6 dans la figure 6.1).

6.2.4 Critères d'évaluation

L'évaluation du modèle généré se fait à travers une interface permettant à l'utilisateur de charger le jeu d'apprentissage, la méthode de génération de règles – les variantes de MSApriori et l'algorithme BERA – et le jeu de test à employer. Pour chaque combinaison « jeu d'apprentissage - jeu de test », cette interface renvoie des fichiers résumant les valeurs trouvées, objets de la section suivante, pour les paramètres d'évaluation employés.

6.3 Indicateurs d'évaluation des modèles d'apprentissage produits

Les modèles issus des différents algorithmes proposés sont évalués selon trois volets : leurs capacités à gérer le problème de l'asymétrie de données, la richesse des motifs et des règles générées, et la qualité de ces dernières en termes de pertinence de la prédiction et de l'explication.

6.3.1 Indicateurs pour la gestion de l'asymétrie

Afin d'évaluer la capacité de chaque algorithme à gérer le problème de l'asymétrie de données (les items de type N sont très fréquents par rapport aux items de type S, et SPF), nous proposons un ensemble d'indicateurs permettant de détecter et d'analyser l'évolution en termes de génération des itemsets rares – items de type SPF et S –. Ainsi pour un algorithme donné, nous considérons la liste des items (1-itemset) fréquents générés et nous calculons, respectivement, le ratio des items S, SPF par rapport aux items N et le ratio des items S, SPF par rapport à l'ensemble des items fréquents générés. Une vue comparative entre les différents algorithmes peut ainsi être fournie.

6.3.2 Indicateurs sur la richesse des règles générées

Ces indicateurs permettent d'évaluer les algorithmes en termes de volume des motifs pertinents et règles pertinentes et confiantes générés. Un motif (itemset) est dit pertinent ou complet s'il est conforme à la structure que nous considérons porteuse d'information utile pour l'explication et la prédiction de l'évolution territoriale. En d'autres termes, s'il est de taille supérieure ou égale à trois et contient exactement un item de type S, un item de type SPF et au moins un item de type N. Une règle est dite confiante si elle dépasse un seuil de confiance prédéfini et est dite pertinente si elle est conforme à une structure dite adéquate à l'explication et à la prédiction de l'évolution, c.-à-d. si elle contient exactement un seul item de type SPF et au moins un item de type N dans sa prémisse et exactement un seul item de type S dans sa conclusion.

6.3.3 Indicateur de qualité interne des modèles

Dans le contexte de l'explication et de la prédiction des dynamiques spatiales, nous avons proposé un ensemble d'algorithmes permettant chacun, de produire un modèle composé d'un ensemble de règles d'évolution. Une règle d'évolution se base sur l'item correspondant à l'attribut SPF, décrivant les relations temporelles de succession de fonctions d'un objet, et les items correspondant à l'attribut N, représentant ses relations spatiales de voisinage, afin de déterminer l'item qui correspond à l'attribut S représentant sa probable fonction successeure.

Par conséquent, les règles d'association produites représentent, entre autres, des règles de classification permettant l'explication et la prédiction de l'apparition d'une instance de l'attribut S. Le modèle qui en est composé, peut ainsi être évalué à travers une matrice de confusion permettant de représenter ses bonnes et ses fausses prédictions.

Considérons, le modèle à évaluer (M) et un nouveau jeu de test (t), la matrice de confusion (mc) présente dans ses lignes les instances réelles de l'attribut cible (C) (*i.e.* dans notre cas, les instances de S observées dans le jeu de test) et dans ses colonnes les instances prédites de celui-ci. Les cases de la matrice représentent le nombre de cas ou transactions de t (X_{ij}) où une instance réelle c_i a été prédite, par le modèle, comme l'instance c_j.

Dans un cadre binaire, la mc permet d'observer quatre types de réponses pouvant être générées par M :

— vrai positive (VP), si une instance étiquetée réellement c est, à raison, prédite comme telle.
— faux positive (FP), si une instance étiquetée réellement c est à tort prédite comme telle.
— Vrai négative (VN), si une instance étiqueté réellement par une valeur différente de c est à raison prédite par le système comme telle.
— Faux négative (FN), si une instance étiquetée réellement par une valeur différente de c est à tort trouvé par le système comme telle.

Se basant sur celle-ci, la matrice de confusion peut dériver diverses métriques d'évaluation. Les plus communément utilisées sont : le rappel, la précision, la F-mesure, le taux de bonne classification, et le taux de cas sans réponse.

La précision (P) : une métrique qui représente la proportion des instances prédites qui sont pertinentes. Elle permet ainsi, de mesurer la capacité du modèle à rejeter les instances non-pertinentes. Elle se calcule selon l'équation 6.1.

$$P = \frac{VP}{VP + FP} \tag{6.1}$$

Le rappel (R) : une métrique qui représente la proportion des instances pertinentes qui sont prédites. Il permet de mesurer la capacité du modèle à fournir toutes les instances pertinentes (cf. figure 6.2).

$$R = \frac{VP}{VP + FN} \tag{6.2}$$

La F-mesure (F) : un indice qui représente un compromis du rappel et de la précision, cette métrique mesure la performance d'un modèle en mesurant sa capacité à fournir toutes les instances pertinentes et à rejeter les non-pertinentes. Elle correspond à la moyenne harmonique du rappel et de la précision donnée par l'équation 6.3.

$$F = \frac{2RP}{R+P} \tag{6.3}$$

Le taux de bonne classification (TBC) : un indice qui représente une métrique mesurant la proportion de cas (dans notre cas les transactions de t) qui ont été correctement prédits (cf. equation 6.4). Bien qu'il représente une mesure plus intuitive pour l'évaluation d'un modèle de classification, le TBC seul n'est pas fiable. Par exemple, un modèle de détection de fraude qui ne prédit que des « non-fraudes » (i.e. un indice VN important avec des FN et VP qui sont nuls) peut avoir un très bon TBC alors qu'en réalité il n'a aucune utilité pour la détection de fraude. Ainsi, il est primordial de considérer d'autres métriques telles que le rappel et la précision qui, s'ils présentent des valeurs proches (*i.e.* FP et FN proches), indique l'absence de ce genre de défaut.

$$TBC = \frac{VP+VN}{VP+VN+FP+FN} \tag{6.4}$$

Le taux de sans réponse (TSR) : un indice qui représente une métrique mesurant la proportion de cas auxquels aucune règle n'est applicable.

Dans notre cadre expérimental, les résultats ne sont pas sous la forme binaire (positifs/négatifs) mais multi-labels (les différentes fonctions des conclusions de nos règles). Notre matrice de confusion aura la forme présentée dans la figure suivante :

Ainsi soit c_i une conclusion présente dans notre base de tests, alors P_{c_i}, R_{c_i} et F_{c_i} sont définies de la manière suivante :

$$\forall c_i \in C, i \in [1,n] : P_{c_i} = \frac{X_{ii}}{\sum_{j=1}^{n} X_{ji}} \tag{6.5}$$

$$\forall c_i \in C, i \in [1,n] : R_{c_i} = \frac{X_{ii}}{\sum_{j=1}^{m} X_{ij}} \tag{6.6}$$

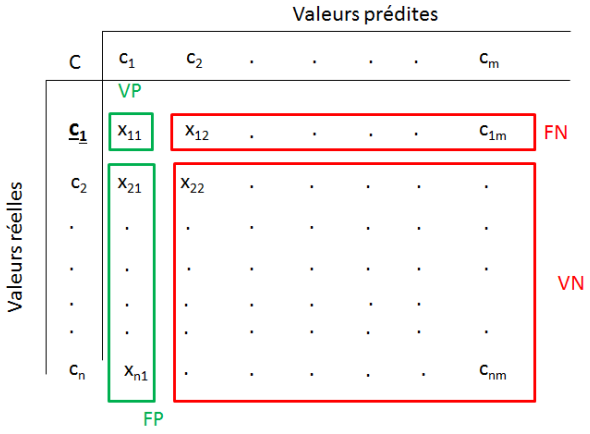


FIGURE 6.6 – Illustration de la matrice de confusion ainsi que les quantités correspondant aux cases représentant les réponses VP, VN, FP, FN pour l'instance c_1.

$$\forall c_i \in C, i \in [1,n] : F_{c_i} = 2\frac{R_{c_i} * P_{c_i}}{R_{c_i} + P_{c_i}} \tag{6.7}$$

Afin d'évaluer la performance globale d'un modèle généré, ces métriques peuvent être calculées pour la totalité des conclusions qu'il génère. En effet, on commence par calculer la valeur de la métrique pour chaque conclusion puis on donne la moyenne des valeurs trouvés. Ainsi, la précision globale (PG), le rappel global (RG), et la F-mesure globale (FG) sont définis comme suit :

$$P_G = \frac{\sum_{i=1}^{n} P_{c_i}}{n} \tag{6.8}$$

$$R_G = \frac{\sum_{i=1}^{n} R_{c_i}}{n} \tag{6.9}$$

$$F_G = \frac{\sum_{i=1}^{n} F_{c_i}}{n} \tag{6.10}$$

6.4 Contexte expérimental

6.4.1 Jeux de données

6.4.1.1 Corine Land Cover

Issue d'un projet mené par l'Agence Européenne de l'Environnement (EEA [3]), CLC représente un inventaire biophysique de l'occupation du sol pour trente-huit pays européen dont la France métropolitaine. L'accès facile à cette base, une bonne documentation, un bon contrôle de la qualité des données, sa simplicité, sa compréhensibilité, et son utilisation par des organismes gouvernementaux et de secteur privé fiables (e.g. IGN [4], CGDD [5], INSEE [6], SOeS [7]) furent certains de nos motivations pour l'utiliser.

Structurée selon une hiérarchie en trois niveaux, la nomenclature de CLC (cf. figure 6.7) comprend, au premier niveau, 5 postes qui correspondent aux grandes catégories d'occupation du sol et qui sont représentés à l'échelle de la planète, 15 au deuxième niveau représentés aux échelles 1/1 000 000 et 1/500 000 et 44 postes au dernier niveau représentés à l'échelle 1/100 000.

Chaque occupation du sol est munie d'un code. Celui-ci est défini en juxtaposant des numéros de postes de chaque niveau de la nomenclature. Par exemple, le code 112 : Zones industrielles ou commerciales pour 1 : Territoires artificialisés et 12 : Zones industrielles ou commerciales et réseaux de communication.

6.4.1.2 Jeux de d'apprentissage et de test

Notre objectif est de vérifier la capacité de nos approches à extraire des règles dont la structure fait écho aux hypothèses premières de notre travail, à savoir la fonction d'un objet géographique dépend de la succession de celles de ses antécédents ainsi que de celles de leurs voisins. Pour cela, afin d'expérimenter notre approche et d'en vérifier la généricité et la qualité, nous avons choisi d'appliquer les algorithmes proposés dans ce contexte, sur deux territoires urbains à assez forte densité : Paris et une partie de la Seine Saint-Denis.

3. En anglais European Environmental Agency
4. Institut géographique national.
5. Commissariat général au développement durable.
6. Institut national de la statistique et des études économiques
7. Service de l'observation et des statistiques.

TABLE 6.1 – Volume des jeux d'apprentissage et de test.

	Seine-Saint-Denis	Paris
Jeux d'apprentissage	494	3913
Jeux de test	3080	4344

CLC fournissant des données pour quatre dates différentes (1990, 2000, 2006 et 2012), nous disposons donc de quatre cartes (ou couche) géographiques datées au format vectoriel pour chaque territoire cas d'étude. Pour chacun, les 3 premières cartes sont réservées à la génération des règles d'évolution alors que la dernière est réservée à évaluer celles-ci en termes de prédiction du phénomène qu'elles régissent.

En d'autres termes, les données d'apprentissage et de test qui en sont extraites consistent, pour chaque territoire, en deux bases transactionnelles : une base réservée à la génération de règles d'association (règles d'évolution) et une base réservée à l'évaluation des performances des règles générées.

Le tableau 6.1 décrit les données correspondant, respectivement, aux départements de Seine-Saint-Denis et de Paris en termes de nombre de cas ou transactions (séquences d'évolution). Utiles pour l'analyse et la compréhension de nos résultats, des informations supplémentaires sur les données d'apprentissage – concernant le volume et la nature des items – sont également fournies dans tablau 6.2.a et tableau 6.2.b.

Nbr total d'items	198
Nbr total d'items S, SPF	54
Nbr total d'items N	144

(a)

Nbr total d'items	190
Nbr total d'items S, SPF	93
Nbr total d'items N	97

(b)

TABLE 6.2 – Bilan des items du point de vue sémantique de prédicats. (a) le cas d'étude de Seine-Saint-Denis, (b) le cas d'étude de Paris

6.5 Résultats

6.5.1 Gestion de l'asymétrie des données

Dans le contexte de notre problématique de recherche, nous visons à générer des règles d'association ayant une structure spécifique et dites pertinentes, à savoir représentant une combinaison d'items correspondant aux trois attributs d'apprentissages S, SPF et N.

Disposant de jeux d'apprentissage présentant une asymétrie en termes de représentativité des trois attributs – les items de type N sont très fréquents par rapport aux items de types S et SPF –, nous avons constaté qu'utiliser l'approche classique de génération de règles d'association – utilisant un seul **minsup** –, ne permet pas de générer nos règles cibles. En effet, fixer un **minsup** assez faible pour faire apparaitre les trois types d'items cause un problème d'explosion combinatoire et fixer un **minsup** permettant d'éviter ce problème ne permet pas de faire apparaitre les items S et SPF dans la liste des items fréquents et ainsi dans les itemsets et les règles généréss à partir de cette liste. En d'autres termes, seules des règles composées d'items de type N sont générées. Afin de remédier à cette problématique, des méthodes utilisant plusieurs **minsup** et afin de faire apparaitre les items S et SPF sont proposées.

Les tableaux 6.3 et 6.4 présentent les performances de ces méthodes en termes de génération des items désirés pour les deux territoires d'étude. Ainsi, ils présentent, les ratios de ces items par rapports aux items fréquents et leurs ratios par rapport aux items correspondant à l'attribut dominant – les items de type N– pour chacune des méthodes.

Une première lecture des résultats, permet de constater que toutes les méthodes proposées ont permis de faire apparaitre les items S, SPF parmi les items fréquents ce qui augmente nos chances pour générer des itemsets complets et ainsi des règles pertinentes.

En comparant les deux familles de méthodes – méthodes à base de quartiles et méthodes à base de clustering – en termes de ratio des items S, SPF fréquents par rapport au total des items fréquents générés et de ratio de S,SPF fréquents par rapport aux items N fréquents, nous constatons que les méthodes à base de quartiles présentent des valeurs supérieurs à ceux correpondant aux méthodes à base de clustering. Nous remarquons également que, pour chaque famille de méthode, la considération de l'aspect sémantique des prédicats lors de la définition et l'affectation des **minsup** a amélioré ces

TABLE 6.3 – Performances en termes de génération des items S et SPF. Comparaisons entre les modèles issus des différents algorithmes proposés : le cas de Seine-Saint-Denis. *BERA est donné à titre indicatif.

	Total Items Fréquents	Total N Fréquents	Total S, SPF Fréquents	Ratio S, SPF Frequents / Total Items Fréquents	Ratio S, SPF Frequents / Total N Fréquents
US : Apriori (minsup=45%)	5	5	0	0	0
MS : QuartilesBased	104	74	30	0.288	0.405
MS : QuartilesBasedSem	80	36	44	0.55	1.222
MS : ClusterBased	111	83	28	0.252	0.337
MS : ClusterBasedSem	104	79	25	0.24	0.316
BERA*	100	74	26	0.26	0.35

ratios. Par exemple, la méthodes à base de quartile considérant l'aspect sémantique (QuartileBasedSem) représente un ratio de 0.55 qui est supérieur au ratio 0.288 de la méthode à base de quartile sans considération de la sémantique des prédicats (QuartilesBased).

Dans ce travail, nous postulons que l'augmentation de nombre des items S, SPF fréquents générés affecte positivement le volume, la richesse, et la qualités des règles générées. Ainsi, nous tentons de confirmer ou non ce postulat en explorant, essentiellement, les corrélations pouvant avoir lieu entre le volume et la richesse des motifs pertinents générés et l'augmentation des ratios des items S, SPF par rapport au total des items fréquents et/ou les ratios des items S, SPF fréquents par rapports aux items N férquents.

Dans ce contexte, les deux sections suivantes présentent des résultats résumants les performances des méthodes proposées, respectivement, en termes de volume d'itemsets générés et en termes de nombre des itemsets complets et de règles pertinentes et confiantes générés.

TABLE 6.4 – Performances en termes de génération des items S et SPF. Comparaisons entre les modèles issus des différents algorithmes proposés : le cas de Paris. *BERA est donné à titre indicatif.

	Total Items Fréquents	Total N Fréquents	Total S, SPF Fréquents	Ratio S, SPF Frequents / Total Items Fréquents	Ratio S, SPF Frequents / Total N Fréquents
US : Apriori (minsup=40%)	6	6	0	0	0
MS : QuartilesBased	89	40	49	0.550	1.225
MS : QuartilesBasedSem	97	24	73	0.753	3.042
MS : ClusterBased	106	57	49	0.462	0.86
MS : ClusterBasedSem	98	49	49	0.5	1
BERA*	96	49	47	0.49	0.959

6.5.2 Performances des méthodes MSApriori en termes de volume des itemsets générés

Dans cette section, nous présentons des résultats en exposant, pour chaque méthode, le volume d'itemsets générés à l'aide des différentes méthodes exécutées chacune sur un nœud d'un calculateur disposant de 144Go de RAM et de 16 cœurs cadencés à 2,93GHz. Les calculs n'étant pas terminés au bout de 192h (8 jours) ont été automatiquement arrêtés.

Ces résultats peuvent ainsi nous donner une idée sur le potentiel de chaque méthode en termes de volumes des itemsets et règles pertinentes pouvant être générés et peuvent également indiquer les éventuels cas d'explosion combinatoire.

La figure 6.8 et la figure 6.9 présentent, respectivement pour Paris et pour Seine-Saint-Denis, une comparaison des volumes des itemsets générés par chaque algorithme.

En examinant ces graphes nous constatons que pour quelques méthodes – QuartilesBasedSem pour Paris ainsi que QuartilesBasedSem, Cluster et ClusterSem pour Seine-Saint-Denis – le nombre des itemsets générés augmente d'une façon très importante. Cette très grande quantité d'itemsets générés induit un coût très important tant en termes de calcul que de mémoire. Aussi, les processus dans ce cadre ne permettent pas de générer les règles d'évolution et donc les modèles explicatifs ou prédictifs correspondant à ces méthodes au bout du temps réservé au traitement (8 jours). Aussi, les résultats correspondant à ces méthodes sont indisponibles dans les tableaux présentés dans les

sections suivantes.

Cette explosion combinatoire est la conséquence de la combinaison de plusieurs paramètres : le fonctionnement de MSApriori, la nature des données d'apprentissage et le choix des différents **minsup**.

En effet, bien qu'elle permette de faire apparaitre des itemsets impliquant des items rares et donc de générer des modèles plus riches et pertinents, l'approche MSApriori ne respectant pas pour certains cas la propriété de Apriori – un itemset fréquent n'est composé que de (sous)-itemsets fréquents – ne permet pas de suffisamment élaguer la treillis (*lattice*) des itemsets possibles. En effet, l'approche MSApriori précise que le minsup d'un itemset correspond à au minimum des minsup des items qui le composent (cf. section 5.5.1.1).

Combiner à l'utilisation de MSApriori, la nature des données d'apprentissage peut également constituer un facteur menant à une explosion combinatoire ce qui est le cas, par exemple, pour les données de Seine-Saint-Denis où l'on dispose d'une valeur importante du ratio attributs/instances. Dans ces cas, le nombre d'instances et le nombre d'attributs sont proches, ce qui fait que la médiane des nombres d'occurences des items S et SPF est très petite. Or cette médiane sert de Minsup. Aussi, au vu de la structure même de MSApriori, l'élagage du treillis est très réduit.

Ainsi, en analysant les motifs fréquents générés nous remarquons que même pour les même données d'apprentissage le choix des valeurs des minsup agit sur la pertinence ou non de l'élagage de l'espace de recherche.

6.5.3 Évaluation de la richesse des modèles générés

Dans cette section, nous fournissons une étude comparative entre les modèles, issus des différentes méthodes d'apprentissage proposées, en termes des motifs et règles générés. Ainsi, nous présentons pour chaque modèle le nombre d'itemsets complets générés et le nombre de règles pertinentes et confiantes qui en sont construites.

La toute première remarque qu'on peut faire, en examinant les tabeaux 6.5 et 6.6, est que l'apparition des items S, SPF parmi les items fréquents a rendu possible la génération de règles et motifs pertinents. En fait, les résultats présentés dans ces tableaux sont en concordance avec les résultats des tableaux 6.4 et 6.3 en ce que les méthodes quartiles, présentant un plus grand nombre d'items S, SPF générés permettent, comparées aux méthodes à base de clustering, de générer plus de motifs complets et de règles complètes et confiantes. Ce qui confirme notre postulat de départ énoncé

TABLE 6.5 – Performances en termes de génération de motifs et règles pertinents. Comparaisons entre les modèles issus des différents algorithmes proposés : le cas de Seine-Saint-Denis.

	ItemSets Complets	Règles Complètes et Confiantes
US : Apriori	-	-
MS : QuartilesBased	1716	1226
BERA	-	437815

comme suit : l'augmentation de nombre des items S, SPF fréquents générés affecte positivement le volume, et la richesse des règles et motifs pertinents générés. Quand à la méthode BERA, nous constatons que cette dernière surpasse, largement, toutes les autres méthodes de l'approche Apriori en termes de volume des règles pertinentes générées.

Bien que les valeurs très élevées de nombres de règles confiantes générées soulignent une capacité remarquable des modèles en termes d'interprétation du phénomène étudié, celle-ci indiquent une difficulté quant à l'interprétabilité du modèle lui même, c'est à dire une difficulté à appréhender la logique derrière l'ensemble des règles générées. Ainsi, une étape de filtrage (faisant l'objet d'une perspective à ce travail) par un traitement automatique, par exemple, à travers la recherche des sous-motifs représentatifs, ou par l'intervention directe de l'expert, s'avère primordiale.

Enfin, nous tenons à souligner que chaque algorithme fournit des règles indiquant la stabilité dans les fonction de l'occupation du sol mais aussi des règles concernant leur évolution. Elles peuvent donc être en elle même riche de sens pour l'expert.

6.5.4 Qualité des règles générées

6.5.4.1 Modèles explicatif

En termes de qualité des règles générées, nous présentons, pour chaque cas d'étude, les valeurs des précision, rappel, f-mesure, TBC et TSR correspondant aux modèles explicatifs issus des différentes méthodes de génération de règles d'évolution (cf. tableaux 6.7 et 6.8).

En comparant, les deux familles de méthodes correspondant à l'approche Apriori – clustering based et quartiles based – nous remarquons que même

TABLE 6.6 – Performances en termes de génération de motifs et règles pertinents. Comparaisons entre les modèles issus des différents algorithmes proposés : le cas de Paris.

	ItemSets Complets	Règles Complètes et Confiantes
US : Apriori	-	-
MS : QuartilesBased	24816	24299
MS : ClusterBased	6274	5747
MS : ClusterBasedSem	9363	8846
BERA	-	423541

TABLE 6.7 – Performances en termes de classification. Comparaisons entre les modèles explicatifs issus de différents algorithmes proposés : le cas de Seine-Saint-Denis.

	R_g	P_g	F_g	TBC	TSR
US : Apriori	-	-	-	-	-
MS : QuartilesBased	0.338	0.254	0.285	0.683	0.917
BERA	0.529	0.529	0.529	1	0.377

pour les indices de qualité des règles les méthodes à base de quartiles surpassent les méthodes à base de clustering (des TBC supérieurs et des TSR inférieurs avec des rappels et précisions proches). Ceci, confirme d'avantage notre postulat de départ liant l'augmentation de nombre des items S, SPF à la génération de règles pertinentes et donc à des meilleurs performances des modèles en termes de prédiction et/ou explication.

Pour la méthode BERA, nous remarquons de bons taux de bonnes classifications. Le modèle arrive à indiquer la bonne évolution (S) pour tous les cas observés pour Seine-Saint-Denis et pour environs 95% des cas pour Paris. Associés à des valeurs de précision et de rappel qui sont presque les mêmes et proches de 50%, ces TBC traduisent vraiment une bonne performance des modèles explicatifs issus de BERA. Les taux de cas sans réponse, qui sont réduits (27% pour Paris et 37% pour Seine-Saint-Denis), soutiennent, également, cette affirmation car ils souligent que ces modèles fournissent une explication pour la plupart des évolutions – e.g. environ 70% des évolutions pour le cas de Paris –.

TABLE 6.8 – Performances en termes de classification. Comparaisons entre les modèles explicatifs issus de différents algorithmes proposés : le cas de Paris.

	R_g	P_g	F_g	TBC	TSR
US : Apriori	-	-	-	-	-
MS : QuartilesBased	0.325	0.417	0.353	0.948	0.598
MS : ClusterBased	0.44	0.42	0.427	0.938	0.901
MS : ClusterBasedSem	0.28	0.415	0.304	0.92	0.534
BERA	0.493	0.523	0.506	0.951	0.277

6.5.4.2 Modèles prédictif

TABLE 6.9 – Performances en termes de classification. Comparaisons entre les modèles prédictifs issus de différents algorithmes proposés : le cas de Seine-Saint-Denis.

	R_g	P_g	F_g	TBC	TSR
US : Apriori	-	-	-	-	-
MS : QuartilesBased	0.19	0.09	0.119	0.234	0.958
BERA	0.166	0.374	0.141	0.422	0.348

TABLE 6.10 – Performances en termes de classification. Comparaisons entre les modèles prédictifs issus de différents algorithmes proposés : le cas de Paris.

	R_g	P_g	F_g	TBC	TSR
US : Apriori	-	-	-	-	-
MS : QuartilesBased	0.233	0.261	0.224	0.848	0.63
MS : ClusterBased	0.241	0.263	0.231	0.722	0.912
MS : ClusterBasedSem	0.224	0.259	0.215	0.852	0.457
BERA	0.4	0.475	0.424	0.864	0.32

Bien qu'ils présentent des valeurs moins bonnes, les valeurs des métriques décrivants les performances des modèles prédictifs (cf. tableaux 6.9 et 6.10) traduisent à peu près les mêmes tendances que celles des modèles explicatifs, c'est-à-dire des rappels et précisions proches, des bons TBR et des TSR réduits et le dévancement des méthodes à base de quartiles par rapport aux méthodes à base de clustering.

6.6 Conclusion

Dans ce chapitre, nous avons présenté une étude expérimentale tentant d'évaluer l'approche générale ainsi que les différentes méthodes (adaptations de MSApriori et BERA) proposées dans le cadre de résolution des défis de notre sujet de thèse. Nous avons cherché à évaluer leur capacité à extraire des règles dont la structure fait écho aux hypothèses premières de notre travail – la fonction d'un objet géographique dépend de la succession de celles de ses antécédents ainsi que de celles de leurs voisins – et également à évaluer les règles obtenues.

Dans ce contexte, nous avons présenté le dispositif expérimental SAFFIET mettant en oeuvre nos propositions, et nous procurant les résultats à évaluer. nous avons également exposé les jeux de tests, les résultats générés, les valeurs des paramètres employés ainsi que leurs significations en termes d'évaluation de l'apport de nos propositions pour la résolution des défis de départ. En effet, les modèles issus des différents algorithmes proposés ont été évalués selon trois volets : leurs capacités à gérer le problème de l'asymétrie de données, la richesse des motifs et des règles qu'ils génèrent, et la qualité de ces dernières en termes de pertinence de la prédiction et de l'explication.

FIGURE 6.7 – Nomenclature de la base de données CLC [EEA, 2009].

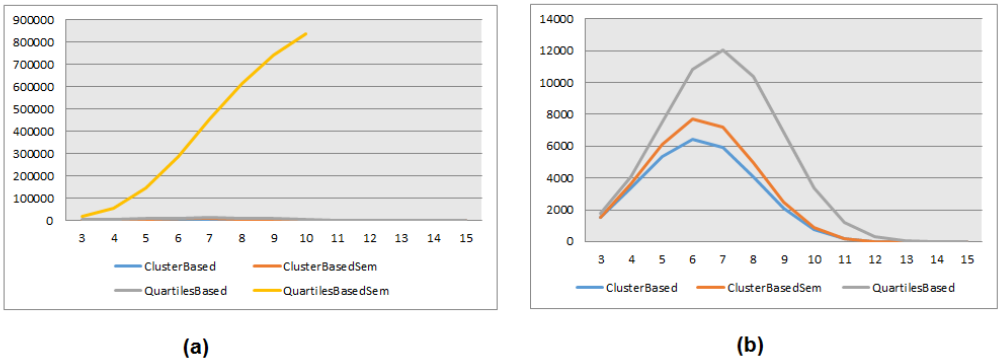


FIGURE 6.8 – Comparaison volumétrique entre les algorithmes correspondants à l'approche MSApriori en termes de générations de k-itemsets : le cas de paris

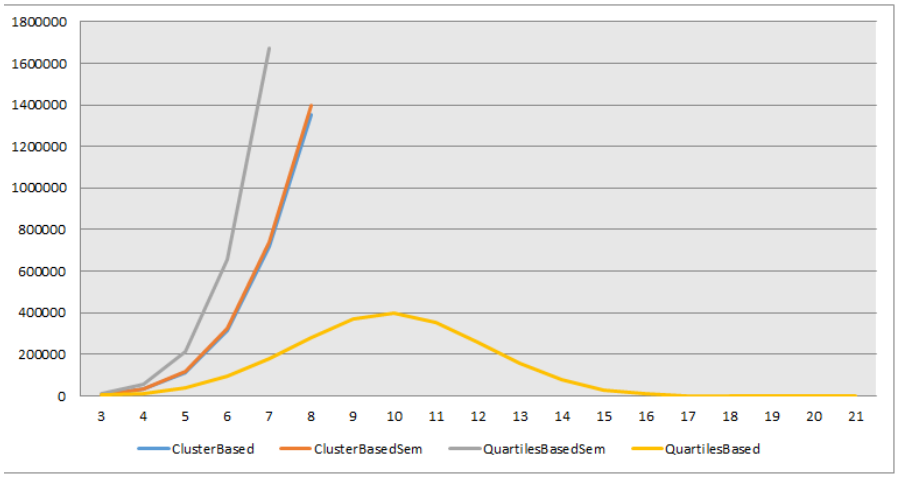


FIGURE 6.9 – Comparaison volumétrique entre les algorithmes correspondants à l'approche MSApriori en termes de génération de k-itemsets : le cas de Seine-Saint-Denis

Chapitre 7

Conclusion et Perspectives

L'information spatiale et temporelle est implicitement présente dans la plupart des bases de données. En effet, quel que soit le domaine d'application, chaque entité physique ou morale peut très souvent être associée à une localisation dans l'espace et certains de ses attributs peuvent varier avec le temps. Par conséquent, il est utile de développer des techniques qui résument efficacement ces données et qui découvrent leurs tendances spatio-temporelles, dans le cadre d'un modèle qui, ainsi, aide à la prise de décision [Cheng et al., 2014]. Ces modèles doivent saisir, entre autres, le comportement évolutif de ces entités au fils du temps et donc fournir un aperçu utile pour le suivi, l'explication et la prédiction d'éventuelles occurrences d'évènements qui lui sont liés. En effet, ces évènements, définis par les changements de caractéristiques des entités, peuvent ainsi être détectés à partir des différentes versions temporelles (ou historiques) de la base de données décrivant les entités étudiées. Ils peuvent également être analysés à travers les techniques de fouille de données afin d'en déduire des modèles spatio-temporelles régissant le phénomène dans lequel ils s'inscrivent. Ces techniques doivent également être capables de répondre aux défis liés à la quantité et à la qualité des données.

Dans un cadre applicatif lié à l'explication et la prédiction de l'évolution territoriale, nous explorons les évènements de changement de fonctions des entités spatio-temporelles et nous tentons de définir une approche permettant le suivi, l'explication et l'anticipation de ceux-ci. Cette approche est définie de façon à traiter l'aspect asymétrique des données spatio-temporelles dont nous disposons.

Afin de répondre à ces problématiques, nous sommes partis d'un ensemble d'hypothèses liées essentiellement à l'appréhension et la conception des no-

tions de temps, d'espace et d'évolution ou d'espace-temps. Dans ce travail, principalement visant à étudier l'évolution d'un espace géographique, nous adoptons la conceptualisation de l'espace comme étant absolue. Conséquemment, le modéliser revient à définir et à étudier les objets ou entités qui sont en son sein. Bien que nous nous basons sur un espace absolu, nous nous positionnons dans l'hypothèse géographique des dynamiques, c'est-à-dire que l'évolution d'une entité dépend des échanges et des rapports qu'elle a avec les autres entités à proximité [Pumain and Saint-Julien, 1997]. Nous nous intéressons, donc, aux relations topologiques entre ces objets afin de construire des relations de voisinage et de succession temporelle dont l'analyse, lors de l'apprentissage, peut apporter des réponses sur la dynamique du phénomène étudié. Nous considérons voisins tous les objets qui sont en adjacence directe.

Dans ce présent mémoire, nous supposons que les données sont acquises et définies à une seule et même échelle spatiale et que le support spatiale reste invariant au fils du temps. Il est, également, important de noter que ce sont la conception linéaire parallèle (time-branching) et l'approche de représentation du temps de Leibniz – le temps est envisagé comme un certain arrangement entre des évènements – qui sont adoptées. En effet, nous percevons le temps comme étant une ligne orientée du passé vers le futur sur laquelle nous plaçons les fonctions des objets selon leurs dates d'observation. À un point donné, cette ligne peut présenter une disjonction en cas de division d'une entité impliquant la présence de deux fonctions sur la même empreinte spatiale, et une conjonction en cas de fusion de deux entités pour former une seule présentant une fonction unique.

Dans le cadre de ces hypothèses, nous définissons une approche qui partant d'un modèle en couches datées indépendantes, exploite le paradigme identitaire pour identifier les objets et leurs relations spatio-temporelles et en construire un modèle traçant leurs évolutions. Ces évolutions ou trajectoires de vie des objets étudiés correspondant à des représentations explicites de leurs changements au cours du temps subissent, par la suite, une opération de prétraitement. Celle-ci vise à les représenter sous un format permettant la génération d'une forme particulière et pertinente de règles dites règles d'évolution.

Bien que ce jeu de données paraisse prêt pour l'extraction de ce genre de règles, celui-ci s'est avéré déséquilibré. En effet, en appliquant un algorithme classique de recherche de règles d'association – Apriori – seules des règles impliquant exclusivement des relations de voisinage sont générées ce qui s'explique par la domination des items correspondant à ces relations par

rapport aux items correspondant aux autres types de relations (e.g. relations temporelles de changement de fonctions).

Dans ce contexte un ensemble de propositions sont faites afin de traiter l'asymétrie inhérente aux attributs d'apprentissage (N, SPF, S) : soit, l'algorithme MSApriori avec affectation de seuils à base de clustering (MS:ClusterBased), MSApriori avec affectation de seuils à base de quartiles (MS:QuartilesBased), leurs variantes considérant la sémantique des items (MS:ClusterBasedSem et MS:QuartilesBasedSem) et l'algorithme BERA.

Parmi ses avantages, l'approche que nous proposons tend vers un traitement complet, automatisé et générique des problèmes liés au suivi, l'explication et la prédiction des phénomènes spatio-temporels tel que l'évolution territoriale. En effet, elle tente de couvrir au mieux toutes les étapes de résolution du problème allant du chargement de données, à leurs modélisation, leurs prétraitement et préparation à l'apprentissage, l'étape d'apprentissage tenant compte des spécificités des données (asymétrie) et leur évaluation. Présenter une conception réalisable d'un outil informatique visant à automatiser les différentes démarches mentionnées ci-dessus, représente, également, l'un des points forts de notre travail. Sur le plan méthodologique et théorique, cette approche définit une sémantique, pour les prédicats des règles à générer, jugée adéquate à des fins explicatives et prédictives du phénomène de l'évolution territoriale. Cette sémantique, compréhensible par l'utilisateur, donne aux règles le potentiel d'être utilisées pour d'autres tâches telles que l'aide à la décision et l'aménagement urbain.

Se focaliser sur les relations spatiales et temporelles incorporées dans les données décrivant un phénomène spatio-temporel offre un certain degré de généricité à l'approche.

Malgré ses avantages, notre approche présente quelque limites tels que la nécessité de finaliser l'ensemble de l'architecture de l'outil SAFFIET afin de permettre la prise en compte de divers formats de données spatio-temporelles et permettre l'éventuel intervention de l'utilisateur tout au long des étapes de l'approche. Sur le plan méthodologique, les contraintes liées aux hypothèses de départ doivent être relachées (*i.e.* échelle spatiale, support spatial, la notion du voisinage, etc.) afin de tendre d'avantage vers une modélisation réaliste du phénomène étudié. De même, les algorithmes d'apprentissage doivent être améliorés afin de réaliser une meilleure gestion de l'aspect asymétrique de données (e.g. agir sur le prétraitement, sur la définition des seuils (**minsup**), etc.) et d'être capable de passer à l'échelle.

Ainsi, les différentes limites mentionnées ci-dessus feront l'objet de nos

perspectives. Celles-ci portent essentiellement sur trois volets: un volet technique concernant la finalisation de l'architecture de l'outil SAFFIET, un volet méthodologique lié à l'enrichissement des hypothèses de notre approche proposée, et un volet qui porte sur l'amélioration des algorithmes d'apprentissages proposés.

7.1 Perspectives

7.1.1 Perspectives techniques

Sur le plan technique, nous envisageons effectuer des développements dont l'objectif est de finaliser l'architecture de l'outil décrite dans la figure 6.1. À savoir:

- modifier le module de chargement de données afin qu'il permette de supporter différents formats et structures des données spatio-temporelles car actuellement seuls des fichiers shapefile dont les attributs sont organisés d'une façon spécifique peuvent être chargés dans la base et traités pour la construction du modèle de données.
- implémenter une interface permettant l'interaction en donnant la main aux experts (par exemple en géographie, en urbanisme, etc.) pour évaluer, enrichir, filtrer ou valider les règles produites. Des retours liés à la performance des algorithmes dans la gestion de l'aspect asymétrique des données sont également envisageables.
- développer un module permettant d'appppliquer les règles d'évolutions les plus fiables pour construire puis visualiser d'une façon cartographiques les prédictions faites par celles-ci : visualiser les fonctions futures de chaque zone géographique d'un territoire étudié.

7.1.2 Perspectives en termes de relachement des contraintes liées aux hypothèses de départ

Parmi ces perspectives, nous pouvons considérer la gestion des données multi-échelles. Également, nous pouvons envisager de générer la variabilité du support spatial entre une couche et une autre dans certains jeux de données.

Sur le niveau conceptualisation de l'espace, nous pouvons considérer d'autres définitions de la relation de voisinage. À savoir, définir le voisinage des objets selon un degré d'adjacence – deux objets de grandes surfaces sont

moins adjacents que deux objets de petites surfaces – ou selon un seuil de distance déterminant d'une façon stricte l'appartenance à la catégorie « voisin » ou la catégorie « non-voisins ». Nous pouvons également intégrer un aspect flou du voisinage, c'est-à-dire, définir un degré de voisinage selon une mesure combinant plusieurs paramètres et déterminant le degré d'appartenance d'un objet à la catégorie voisin.

Sur le niveau modélisation temporelle, nous pouvons adopter une modélisation plutôt quantitative et continue du temps, où un évènement est défini sur un intervalle temporel fixe ce qui permetterait d'estimer la validité temporelle des règles d'évolution produites. Un fenêtrage temporel avec chevauchement peut également être considéré. Ainsi, une fonction peut être valide pour deux intervalles temporels.

Sur le niveau sémantique des données géographiques – *i.e.* la composante attributaire de l'information géographique – qui correspond dans notre cas aux fonctions des objets géographiques étudiés, nous pouvons envisager de l'hierarchiser. En d'autres termes, nous pouvons définir des fonctions à plusieurs niveaux hierarchiques (e.g. la fonction équipement public correspond à un niveau supérieur à école ou hôpital, ainsi les règles impliquant cette fonction peuvent être appliquées pour les objets école et hôpital) et donc générer des règles multi-niveaux.

Sur le niveau représentation des données d'apprentissage, l'attribut SPF, correspondant à la séquence d'évolution précédente, peut être représenté autrement dans les transactions. À savoir, remplacer l'item correspondant à cette séquence par plusieurs items qui chacun correspond à une de ses sous-séquences. Ceci permet, ainsi, d'enrichir, en termes de volume et connaissance les règles générées. Cela permet de faire apparaitre des règles qui auparavant ne pouvaient pas être générées car les itemsets qui sont à leur origine ne sont pas fréquentes si elles impliquent la séquence entière mais fréquentes si elles impliquent une sous-séquence de celle-ci.

7.1.3 Perspectives en termes d'amélioration des algorithmes d'apprentissage

Sur ce volet, nous envisageons des perspectives liées essentiellement au passage à l'échelle des algorithmes d'apprentissage proposés. Bien qu'elles permettent de travailler sur des jeux de données assez large, la performance des algorithmes, en termes de ressources computationnelles et de mémoire,

se dégrade en augmentant la taille des données. Pour certain jeux de données très larges correspondant à des territoires réels et que nous pouvons souhaiter étudier, le traitement poserait quelques problèmes.

C'est essentiellement la recherche des items fréquents qui nécessite un comptage de ceux-ci et le stockage de ces itemsets fréquents qui sont gourmands en ressources. Ainsi, utiliser le paradigme MapReduce [Dean and Ghemawat, 2008] et la parallélisation du traitement peuvent constituer une solution à ce challenge.

En relation avec le traitement de l'aspect asymétrie des données, nous pouvons envisager d'obtimiser la définition des seuils de fréquence (**minsup**) en permettant par exemple à l'expert d'indiquer les attributs dominants ce qui permet une meilleure définition et affectation des minsup.

Bibliographie

[Aggarwal, 2014] Aggarwal, C. C. (2014). *Applications of Frequent Pattern Mining*, pages 443–467. Springer International Publishing, Cham.

[Aggarwal and Han, 2014] Aggarwal, C. C. and Han, J. (2014). *Frequent pattern mining*. Springer.

[Agrawal et al., 1998] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2):94–105.

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.

[Agrawal et al., 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

[Allen, 1983] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

[Allen, 1984] Allen, J. F. (1984). Towards a general theory of action and time. *Artif. Intell.*, 23(2):123–154.

[Allen and Ferguson, 1994] Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579.

[Alonso, 1964] Alonso, W. (1964). *Location and land use: toward a general theory of land rent*. Publication of the Joint Center for Urban Studies. Harvard University Press.

[Alouaoui et al., 2015] Alouaoui, H., Turki, S. Y., and Faiz, S. (2015). Mining spatiotemporal association rules from spatiotemporal databases between two different fixed dates. *Int. J. Knowl. Eng. Data Min.*, 3(2):190–207.

[Armstrong, 1988] Armstrong, M. P. (1988). Temporality in spatial databases. In *GIS/LIS'88*, pages 880–889.

[Augustin d'Hippone, 1864] Augustin d'Hippone, S. (1864). *Oeuvres complètes de Saint Augustin*. L. Guérin et Cie.

[Bailly et al., 2016] Bailly, A., Béguin, H., and Scariati, R. (2016). *Introduction à la géographie humaine - 9e éd.* Géographie. Armand Colin.

[Batty, 2004] Batty, M. (2004). Dissecting the streams of planning history: Technology versus policy through models. *Environment and Planning B: Planning and Design*, 31(3):326–330.

[Batty, 2007] Batty, M. (2007). *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press.

[Beil et al., 2002] Beil, F., Ester, M., and Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 436–442, New York, NY, USA. ACM.

[Beller et al., 1991] Beller, A., Giblin, T., Le, K. V., Litz, S., Kittel, T., and Schimel, D. (1991). Temporal gis prototype for global change research. *GIS/LIS'91, Atlanta, GA, USA, 10/28-11/01/91*, pages 752–765.

[Belussi et al., 1999] Belussi, A., Negri, M., and Pelagatti, G. (1999). Management of data changes in geodatabases: time component in gis. *Geomatics Info Magazine International*, 13(7):41–43.

[Berling-Wolff and Wu, 2004] Berling-Wolff, S. and Wu, J. (2004). Modeling urban landscape dynamics: A review. *Ecological Research*, 19(1):119–129.

[Bharati and Ramageri, 2010] Bharati, M. and Ramageri, M. (2010). Data mining techniques and applications.

[Bhatt and Patel, 2014] Bhatt, U. Y. and Patel, P. A. (2014). A recent overview: Rare association rule mining. *International Journal of Computer Applications*, Volume 107(18):1–4.

[Bonabeau, 2002] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287.

[Boulicaut and Jeudy, 2010] Boulicaut, J.-F. and Jeudy, B. (2010). *Constraint-based Data Mining*, pages 339–354. Springer US, Boston, MA.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

[Briassoulis, 2000] Briassoulis, H. (2000). Analysis of land use change: theoretical and modeling approaches. Technical report, Regional Research Institute, West Virginia University.

[Burdick et al., 2005] Burdick, D., Gehrke, J., Flannick, J., Yiu, T., and Calimlim, M. (2005). Mafia: A maximal frequent itemset algorithm. *IEEE Transactions on Knowledge & Data Engineering*, 17:1490–1504.

[Caloz and Collet, 2011] Caloz, R. and Collet, C. (2011). *Analyse spatiale de l'information géographique*. PPUR Presses polytechniques.

[Chandio et al., 2011] Chandio, I. A., Matori, A.-N., Lawal, D. U., and Sabri, S. (2011). Gis-based land suitability analysis using ahp for public parks planning in larkana city. *Modern applied science*, 5(4):177.

[Charif et al., 2012] Charif, O., Omrani, H., and Basse, R.-M. (2012). Cellular automata based on artificial neural network for simulating land use changes. In *Proceedings of the 45th Annual Simulation Symposium*, ANSS '12, pages 1:1–1:9, San Diego, CA, USA. Society for Computer Simulation International.

[Chen et al., 2014] Chen, J., Miller, C., and Dagher, G. G. (2014). Product recommendation system for small online retailers using association rules mining. In *Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM)*, pages 71–77.

[Cheng et al., 2014] Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., and Wang, J. (2014). *Spatiotemporal Data Mining*, pages 1173–1193. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Cheylan et al., 1999] Cheylan, J.-P., Gautier, D., Lardon, S., Libourel, T., Mathian, H., Motet, S., and Sanders, L. (1999). Les mots du traitement de l'information spatio-temporelle. *Revue internationale de géomatique*, 9(1):11–23.

[Cheylan and Lardon, 1993] Cheylan, J.-P. and Lardon, S. (1993). Towards a conceptual data model for the analysis of spatio-temporal processes: the example of the search for optimal grazing strategies. *Spatial Information Theory A Theoretical Basis for GIS*, pages 158–176.

[Cheylan et al., 1997] Cheylan, J.-P., Lardon, S., Mathian, H., and Sanders, L. (1997). Les problématiques liées au temps dans les sig. *Revue Internationale de Géomatique*, pages 287–305.

[Cho et al., 2008] Cho, C.-W., Zheng, Y., Wu, Y.-H., and Chen, A. L. (2008). A tree-based approach for event prediction using episode rules over event streams. In *International Conference on Database and Expert Systems Applications*, pages 225–240. Springer.

[Claramunt et al., 1997a] Claramunt, C., Parent, C., and Thériault, M. (1997a). Design patterns for spatio-temporal processes. *Data Mining and Reverse Engineering*, pages 455–475.

[Claramunt and Thériault, 1995a] Claramunt, C. and Thériault, M. (1995a). Managing time in gis an event-oriented approach. In *Recent Advances in Temporal Databases*, pages 23–42. Springer.

[Claramunt and Thériault, 1995b] Claramunt, C. and Thériault, M. (1995b). Managing time in gis: an event-oriented approach. *Recent Advances in Temporal Databases*, pages 23–42.

[Claramunt and Theriault, 1996] Claramunt, C. and Theriault, M. (1996). Toward semantics for modelling spatio-temporal processes within gis. *Advances in GIs Research I*, pages 27–43.

[Claramunt et al., 1997b] Claramunt, C., Thériault, M., and Parent, C. (1997b). A qualitative representation of evolving spatial entities in two-dimensional topological spaces.

[Clementini et al., 1993] Clementini, E., Di Felice, P., and van Oosterom, P. (1993). *A small set of formal topological relationships suitable for end-user interaction*, pages 277–295. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Cong et al., 2005] Cong, G., Tan, K.-L., Tung, A. K. H., and Xu, X. (2005). Mining top-k covering rule groups for gene expression data. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 670–681, New York, NY, USA. ACM.

[Conte et al., 2013] Conte, R., Hegselmann, R., and Terna, P. (2013). *Simulating Social Phenomena*. Lecture Notes in Economics and Mathematical Systems. Springer Berlin Heidelberg.

[Cover and Hart, 2006] Cover, T. and Hart, P. (2006). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27.

[Darwen and Date, 2006] Darwen, H. and Date, C. (2006). *A generic model for spatio-bitemporal geographic information*, pages 481–514. Apress, Berkely, CA, USA.

[De Risi, 2012] De Risi, V. (2012). Leibniz on relativity. the debate between hans reichenbach and dietrich mahnke on leibnizś theory of motion and time. *New essays in Leibniz reception: Science and philosophy of science 1800–2000*, pages 143–185.

[De Sá et al., 2011] De Sá, C., Soares, C., Jorge, A., Azevedo, P., and Costa, J. (2011). Mining association rules for label ranking. *Advances in Knowledge Discovery and Data Mining*, pages 432–443.

[Dean and Ghemawat, 2008] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

[Dell'Erba and Libourel, 1997] Dell'Erba, E. and Libourel, T. (1997). Temps et évolution dŠentités géoréférencées. *Treizième journée de Bases de Données Avancées (BDA'97), Grenoble.*

[Di Méo, 1985] Di Méo, G. (1985). Les formations socio-spatiales ou la dimension infra-régionale en géographie. *Annales de Géographie*, 94(526):661–689.

[Dua and Kidambi, 2010] Dua, S. and Kidambi, P. C. (2010). Protein structural classification using orthogonal transformation and class-association rules. *International journal of data mining and bioinformatics*, 4(2):175–190.

[Echenique, 1972] Echenique, M. (1972). Urban space and structures. In Martin, L. and March, L., editors, *Models: A discussion*, pages 164–174. Cambridge University Press, London.

[EEA, 2009] EEA (2009). Corine land cover france guide d'utilisation. Technical report, Commissariat général au développement durable (CGDD), Service de l'observation et des statistiques (SOeS).

[Egenhofer and Franzosa, 1991] Egenhofer, M. J. and Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information System*, 5(2):161–174.

[Epstein and Axtell, 1996] Epstein, J. M. and Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up.* The Brookings Institution, Washington, DC, USA.

[Fahed, 2016] Fahed, L. (2016). *Predicting and influencing the appearance of events in a complex sequence*. Theses, Université de Lorraine.

[Farzanyar and Kangavari, 2012] Farzanyar, Z. and Kangavari, M. R. (2012). Efficient mining of fuzzy association rules from the pre-processed dataset. *Computing and Informatics*, 31(2):331–347.

[Fayyad et al., 1996a] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery: An overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Fayyad et al., 1996b] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996b). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Flewelling et al., 1992] Flewelling, D. M., Egenhofer, M. J., and Frank, A. U. (1992). Constructing geological cross sections with a chronology of geologic events. In *5th International Symposium on Spatal Data Handling, Charleston, South Carolina, USA, IGU Commision on GIS*.

[Forbus, 1984] Forbus, K. D. (1984). Qualitative process theory. *Artificial intelligence*, 24(1):85–168.

[Frank, 2008] Frank, A. (2008). Ontology. In *In K. K. Kemp (Ed.), Encyclopedia of geographic information science*, pages 327–329. Thousand Oaks, CA: SAGE Publications Ltd.

[Frank, 1994] Frank, A. U. (1994). Qualitative temporal reasoning in gis-ordered time scales. In *Sixth International Symposium on Spatial Data Handling, SDH*, volume 94, pages 410–430.

[Freksa, 1992] Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial intelligence*, 54(1-2):199–227.

[Galán et al., 2009] Galán, J. M., Izquierdo, L. R., Izquierdo, S. S., Santos, J. I., Del Olmo, R., López-Paredes, A., and Edmonds, B. (2009). Errors and artefacts in agent-based modelling. *Journal of Artificial Societies and Social Simulation*, 12(1):1.

[Galton, 2000] Galton, A. (2000). *Qualitative spatial change*. Oxford University Press on Demand.

[Galton, 2001] Galton, A. (2001). Space, time, and the representation of geographical reality. *Topoi*, 20(2):173–187.

[Galton, 2004] Galton, A. (2004). Fields and objects in space, time, and space-time. *Spatial cognition and computation*, 4(1):39–68.

[Galton, 2006] Galton, A. (2006). On what goes on: The ontology of processes and events. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, pages 4–11, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[Galton, 2007] Galton, A. (2007). Experience and history: Processes and their relation to events. *Journal of Logic and Computation*, 18(3):323–340.

[Galton and Worboys, 2005] Galton, A. and Worboys, M. (2005). Processes and events in dynamic geo-networks. *GeoS*, 3799:45–59.

[Gautam and Pardasani, 2011] Gautam, P. and Pardasani, K. R. (2011). Efficient method for multiple-level association rules in large databases. *Journal of Emerging Trends in Computing and Information Sciences*, 2(12):722–732.

[Gharbi et al., 2014] Gharbi, A., De Runz, C., Faiz, S., and Akdag, H. (2014). An association rules based approach to predict semantic land use evolution in the french city of saint-denis. *International Journal of Data Warehousing and Mining (IJDWM)*, 10(2):1–17.

[Gharbi et al., 2016a] Gharbi, A., de Runz, C., Faiz, S., and Akdag, H. (2016a). Saffiet : un système d'extraction de règles d'associations spatiales et fonctionnelles dans les séries de données géographiques. In *Extraction et Gestion des Connaissances (EGC)*, Reims, France. Hermann.

[Gharbi et al., 2016b] Gharbi, A., De Runz, C., Faiz, S., and Akdag, H. (2016b). Towards association rules as a predictive tool for geospatial areas evolution. In *2nd International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2016)*, pages 201–206. INSTICC.

[Gharbi et al., 2016c] Gharbi, A., de Runz, C., Faiz, S., and Akdag, H. (2016c). Un modèle à base de règles d'associations spatio-temporelles pour la prédiction de l'évolution territorial. In *EXCES@SAGEO2016*, Nice, France.

[Ghosh and Nath, 2004] Ghosh, A. and Nath, B. (2004). Multi-objective rule mining using genetic algorithms. *Inf. Sci.*, 163(1-3):123–133.

[Gosain and Bhugra, 2013] Gosain, A. and Bhugra, M. (2013). A comprehensive survey of association rules on quantitative data in data mining. In *Information Communication Technologies (ICT), 2013 IEEE Conference on*, pages 1003–1008.

[Gösta Grahne and Jianfei Zhu, 2003] Gösta Grahne and Jianfei Zhu (2003). Efficiently using prefix-trees in mining frequent itemsets. In *In: Proceeding of the ICDM'03 international workshop on frequent itemset mining implementations (FIMI'03)*, pages 123–132, Melbourne, FL.

[Grenon and Smith, 2004] Grenon, P. and Smith, B. (2004). Snap and span: Towards dynamic spatial ontology. *Spatial cognition and computation*, 4(1):69–104.

[Gu et al., 2009] Gu, W., Wang, X., and Geng, L. (2009). Gis-flsolution: A spatial analysis platform for static and transportation facility location allocation problem. In *International Symposium on Methodologies for Intelligent Systems*, pages 453–462. Springer.

[Guida et al., 1999] Guida, G., Lamperti, G., and Zanella, M. (1999). *Software Prototyping in Data and Knowledge Engineering*. Kluwer Academic Publishers, Norwell, MA, USA.

[Hallot and Billen, 2016] Hallot, P. and Billen, R. (2016). Enhancing spatio-temporal identity: States of existence and presence. *ISPRS International Journal of Geo-Information*, 5(5):62.

[Han et al., 2007] Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining: Current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86.

[Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann.

[Han et al., 2012] Han, J., Kamber, M., and Pei, J. (2012). 7 - advanced pattern mining. In Han, J., Kamber, M., , and Pei, J., editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 279 – 325. Morgan Kaufmann, Boston, third edition edition.

[Han et al., 2000] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA. ACM.

[Hand, 1998] Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician*, 52(2):112–118.

[Higginbotham et al., 2000] Higginbotham, J., Pianesi, F., and Varzi, A. (2000). *Speaking of Events*. Oxford University Press.

[Hopkins, 1999] Hopkins, L. D. (1999). Structure of a planning support system for urban development. *Environment and Planning B: Planning and Design*, 26(3):333–343.

[Hornsby and Egenhofer, 1998] Hornsby, K. and Egenhofer, M. J. (1998). Identity-based change operations for composite objects. In *Proceedings of 8th International Symposium on Spatial Data Handling, Edited by POIKER, T. and CHRISMAN, N.*, pages 202–213, Vancouver, Canada. International Geographical Union.

[Hornsby and Egenhofer, 2000] Hornsby, K. and Egenhofer, M. J. (2000). Identity-based change: a foundation for spatio-temporal knowledge representation. *International journal of geographical information science*, 14(3):207–224.

[Hunt and Echenique, 1993] Hunt, J. D. and Echenique, M. (1993). Experience in the application of the meplan framework for land use and transport interaction modeling. In *4th National Conference on Transportation Planning Methods Applications, Volumes I and II. A Compendium of Papers*.

[Iba, 2013] Iba, H. (2013). *Agent-Based Modeling and Simulation with Swarm*. Chapman & Hall/CRC Studies in Informatics Series. CRC Press.

[Ilayaraja and Meyyappan, 2015] Ilayaraja, M. and Meyyappan, T. (2015). Efficient data mining method to predict the risk of heart diseases through frequent itemsets. *Procedia Computer Science*, 70:586 – 592.

[ISO, 2004] ISO, I. O. f. S. (2004). Iso 8601 :2004 Éléments de données et formats d'échange – Échange d'information – représentation de la date et de l'heure. , University of Southern California.

[ISO, 2007] ISO, I. O. f. S. (2007). Iso19136 :2007, technical committee 211, geographic information – geography markup language (gml).

[Janiak, 2016] Janiak, A. (2016). Kant's views on space and time. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

[Janssen and Jager, 2000] Janssen, M. and Jager, W. (2000). The human actor in ecological-economic models: Preface. *Ecological Economics*, 35(3):307–310.

[Jenerette and Wu, 2001] Jenerette, G. D. and Wu, J. (2001). Analysis and simulation of land-use change in the central arizona – phoenix region, usa. *Landscape Ecology*, 16(7):611–626.

[Jiang and Claramunt, 2004] Jiang, B. and Claramunt, C. (2004). A structural approach to the model generalization of an urban street network. *GeoInformatica*, 8(2):157–171.

[Jiang et al., 2000] Jiang, B., Claramunt, C., and Klarqvist, B. (2000). Integration of space syntax into gis for modelling urban spaces. *International Journal of Applied Earth Observation and Geoinformation*, 2(3-4):161–171.

[Jiang and Worboys, 2009] Jiang, J. and Worboys, M. (2009). Event-based topology for dynamic planar areal objects. *International Journal of Geographical Information Science*, 23(1):33–60.

[Jr and B, 1973] Jr, L. and B, D. (1973). Requiem for large-scale models. *Journal of the American Institute of Planners*, 39(3):163–178.

[Kanaroglou and Scott, 2002] Kanaroglou, P. and Scott, D. (2002). Integrated urban transportation and land-use models for policy analysis. *Governing Cities on the Move. Functional and Management Perspectives on Transformations of European Urban Infrastructures, edited by M. Dijst, W. Schenkel, and I. Thomas. Hamshire England: Ashgate.*

[Kantardzic, 2003] Kantardzic, M. (2003). *Data mining: concepts, models, methods, and algorithms*. Wiley-Interscience.

[Kauppinen and Hyvönen, 2007] Kauppinen, T. and Hyvönen, E. (2007). *Modeling and Reasoning About Changes in Ontology Time Series*, pages 319–338. Springer US, Boston, MA.

[Kauppinen et al., 2008] Kauppinen, T., Väätäinen, J., and Hyvönen, E. (2008). Creating and using geospatial ontology time series in a semantic cultural heritage portal. *The Semantic Web: Research and Applications*, pages 110–123.

[Kaur and Kang, 2016] Kaur, M. and Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, 85:78 – 85.

[Khan et al., 2008] Khan, M. S., Muyeba, M., and Coenen, F. (2008). Weighted association rule mining from binary and fuzzy data. In *Industrial Conference on Data Mining*, pages 200–212. Springer.

[Khoshafian and Copeland, 1986] Khoshafian, S. N. and Copeland, G. P. (1986). *Object identity*, volume 21. ACM.

[Kim, 1976] Kim, J. (1976). Events as property exemplifications. In Brand, M. and Walton, D., editors, *Action Theory*, pages 310–326. D. Reidel.

[Kiran and Re, 2009] Kiran, R. U. and Re, P. K. (2009). An improved multiple minimum support based approach to mine rare association rules. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 340–347.

[Klein, 2009] Klein, E. (2009). Quelle est la forme du temps? linéaire ou cyclique. *Conférences du lundi sur l'Inde*, 16.

[Klosterman, 1994] Klosterman, R. E. (1994). Large-scale urban models retrospect and prospect. *Journal of the American Planning Association*, 60(1):3–6.

[Kouris et al., 2003] Kouris, I. N., Makris, C., and Tsakalidis, A. K. (2003). An improved algorithm for mining association rules using multiple support values. In *FLAIRS Conference*, pages 309–313.

[Kulkarni and Michels, 2012] Kulkarni, K. and Michels, J.-E. (2012). Temporal features in sql:2011. *ACM Sigmod Record*, 41(3):34–43.

[Kumar and Verma, 2012] Kumar, R. and Verma, R. (2012). Classification algorithms for data mining: A survey. *International Journal of Innovations in Engineering and Technology (IJIET)*, 1(2):7–14.

[Ladkin, 1987] Ladkin, P. B. (1987). *The logic of time representation.* PhD thesis, Citeseer.

[Lambin et al., 2000] Lambin, E., Rounsevell, M., and Geist, H. (2000). Are agricultural land-use models able to predict changes in land-use intensity? *Agriculture, Ecosystems & Environment*, 82(1–3):321 – 331.

[Langran, 1992] Langran, G. (1992). Time in geographic information systems. *Geocarto International*, 7(2):40–40.

[Langran and Chrisman, 1988] Langran, G. and Chrisman, N. R. (1988). A framework for temporal geographic information. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 25(3):1–14.

[Laxman et al., 2009] Laxman, S., Shadid, B., Sastry, P., and Unnikrishnan, K. (2009). Temporal data mining for root-cause analysis of machine faults in automotive assembly lines. *arXiv preprint arXiv:0904.4608*.

[Laxman et al., 2008] Laxman, S., Tankasali, V., and White, R. W. (2008). Stream prediction using a generative model based on frequent episodes in

event sequences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 453–461, New York, NY, USA. ACM.

[Lee et al., 2006] Lee, Y.-C., Hong, T.-P., and Wang, T.-C. (2006). *Mining Multiple-Level Association Rules Under the Maximum Constraint of Multiple Minimum Supports*, pages 1329–1338. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Li et al., 2001] Li, W., Han, J., and Pei, J. (2001). Cmar: accurate and efficient classification based on multiple class-association rules. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 369–376.

[Lin and Li, 2015] Lin, J. and Li, X. (2015). Knowledge transfer for large-scale urban growth modeling based on formal concept analysis. *Transactions in GIS*, pages n/a–n/a.

[Linoff and Berry, 2011] Linoff, G. S. and Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley technology publication. Wiley.

[Liu, 2011] Liu, B. (2011). *Association Rules and Sequential Patterns*, pages 17–62. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Liu et al., 1998] Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, pages 80–86. AAAI Press.

[Liu et al., 1999] Liu, B., Hsu, W., and Ma, Y. (1999). Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 337–341. ACM.

[Liu et al., 2003] Liu, G., Lu, H., Lou, W., and Yu, J. X. (2003). On computing, storing and querying frequent patterns. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 03, pages 607–612, New York, NY, USA. ACM.

[Liu et al., 2016] Liu, H., He, G., Jiao, W., Wang, G., Peng, Y., and Cheng, B. (2016). Sequential pattern mining of land cover dynamics based on time-series remote sensing images. *Multimedia Tools and Applications*, pages 1–24.

[Liu et al., 2007] Liu, W., Seto, K., Sun, Z., and Tian, Y. (2007). Urban land use prediction model with spatio-temporal data mining and gis. *Urban remote sensing. CRC Press, Taylor and Francis*, pages 165–78.

[Livet, 2008] Livet, P. (2008). La notion d'évènement chez whitehead et davidson. *Noesis*, (13):217–233.

[Longley et al., 2015] Longley, P., Goodchild, M., Maguire, D., and Rhind, D. (2015). *Geographic Information Science and Systems, 4th Edition*. Wiley.

[Malek et al., 2015] Malek, Ž., Boerboom, L., and Glade, T. (2015). Future forest cover change scenarios with implications for landslide risk: An example from buzau subcarpathians, romania. *Environmental Management*, 56(5):1228–1243.

[Maragatham and Lakshmi, 2012] Maragatham, G. and Lakshmi, M. (2012). A recent review on association rule mining. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(6):831–836.

[Mathian and Sanders, 2015] Mathian, H. and Sanders, L. (2015). Temporalités et objets géographiques. *L'Information géographique*, 79(2):55–64.

[Mazur, 1987] Mazur, J. E. (1987). Quantitative analyses of behavior. *The Effect of Delay and of Intervening Events on Reinforcement Value*, 5:55–73.

[McIntosh and Yuan, 2005] McIntosh, J. and Yuan, M. (2005). Assessing similarity of geographic processes and events. *Transactions in GIS*, 9(2):223–245.

[Mercure and Pronovost, 1989] Mercure, D. and Pronovost, G. (1989). *Temps et société*. Québec: Institut québécois de recherche sur la culture.

[Merlin, 1974] Merlin, P. (1974). Methodes quantitatives et espace urbain. *Population*, 29(3):17 – 27.

[Mertens and Lambin, 1997] Mertens, B. and Lambin, E. F. (1997). Spatial modelling of deforestation in southern cameroon: spatial disaggregation of diverse deforestation processes. *Applied Geography*, 17(2):143–162.

[Miao and Shen, 2010] Miao, R. and Shen, X. J. (2010). Construction of periodic temporal association rules in data mining. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 2133–2137.

[Mitasova and Mitas, 1998] Mitasova, H. and Mitas, L. (1998). Process modeling and simulations. *NCGIA Core Curriculum in GIScience*. Exposé le : 1998-12-2.

[Mithal et al., 2011] Mithal, V., Garg, A., Boriah, S., Steinbach, M., Kumar, V., Potter, C., Klooster, S., and Castilla-Rubio, J. C. (2011). Monitoring global forest cover using data mining. *ACM Trans. Intell. Syst. Technol.*, 2(4):36:1–36:24.

[Molchanov and Woyczynski, 2012] Molchanov, S. A. and Woyczynski, W. A. (2012). *Stochastic models in geosystems*, volume 85. Springer Science & Business Media.

[Mondo, 2011] Mondo, G. D. (2011). *Un Modele De Graphe Spatio-Temporel Pour Representer L'evolution D'entites Geographiques. (A Spatio-Temporal Graph-Based Model For The Evolution Of Geographical Entities).* PhD thesis, University of Western Brittany, Brest, France.

[Mourelatos, 1978] Mourelatos, A. P. (1978). Events, processes, and states. *Linguistics and philosophy*, 2(3):415–434.

[Nain and Vardi, 2007] Nain, S. and Vardi, M. Y. (2007). *Automated Technology for Verification and Analysis: 5th International Symposium, ATVA 2007 Tokyo, Japan, October 22–25, 2007 Proceedings*, chapter Branching vs. Linear Time: Semantical Perspective, pages 19–34. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Newton et al., 1759] Newton, I., Châtelet, D., and du Palais des Arts, B. (1759). *Principes mathématiques de la philosophie naturelle*. Principes mathématiques de la philosophie naturelle. Chez Desaint & Saillant.

[Nguyen et al., 2015] Nguyen, D., Vo, B., and Le, B. (2015). Ccar: An efficient method for mining class association rules with itemset constraints. *Engineering Applications of Artificial Intelligence*, 37:115 – 124.

[Nguyen et al., 2013] Nguyen, L. T., Vo, B., Hong, T.-P., and Thanh, H. C. (2013). Car-miner: An efficient algorithm for mining class-association rules. *Expert Systems with Applications*, 40(6):2305 – 2311.

[OGC, 1999] OGC, O. G. C. (1999). Simple feature access-part 2: Sql option.

[Ott and Swiaczny, 2001] Ott, T. and Swiaczny, F. (2001). *Time-Integrative Geographic Information Systems: Management and Analysis of Spatio-Temporal Data*. Number v. 1 in Time-integrative Geographic Information Systems: Management and Analysis of Spatio-temporal Data. Springer Berlin Heidelberg.

[Paque, 2004] Paque, D. (2004). Gestion de l'historicité et méthodes de mise à jour dans les sig. *Cybergeo: European Journal of Geography*.

[Park et al., 2015] Park, H. A., Kim, T., Li, M., Shon, H. S., Park, J. S., and Ryu, K. H. (2015). Application of gap-constraints given sequential frequent pattern mining for protein function prediction. *Osong Public Health and Research Perspectives*, 6(2):112 – 120.

[Park et al., 1995] Park, J. S., Chen, M.-S., and Yu, P. S. (1995). Efficient parallel data mining for association rules. In *Proceedings of the Fourth International Conference on Information and Knowledge Management*, CIKM '95, pages 31–36, New York, NY, USA. ACM.

[Parsons et al., 2004] Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: A review. *SIGKDD Explor. Newsl.*, 6(1):90–105.

[Pasquier et al., 1999] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). *Discovering Frequent Closed Itemsets for Association Rules*, pages 398–416. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Patel and Thakral, 2016] Patel, K. M. A. and Thakral, P. (2016). The best clustering algorithms in data mining. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 2042–2046.

[Pei et al., 2000] Pei, J., Han, J., and Mao, R. (2000). Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 11–20.

[Pei et al., 2001] Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U., and chun Hsu, M. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceeding of the 2001*, pages 215–224.

[Perret et al., 2015] Perret, J., De Runz, C., Rodier, X., Varet-Vitu, A., Dumenieu, B., Saligny, L., Cristofoli, P., Lefebvre, B., and Desjardin, E. (2015). Études des dynamiques de l'occupati n du sol : questionnement, simplification et limites. *Revue Internationale de Géomatique*, 25(3):301–330.

[Peterschmitt, 2012] Peterschmitt, L. (2012). Chapitre 6. l'espace absolu chez newton et les newtoniens: un lieu entre physique et métaphysique. In *Espace et lieu dans la pensée occidentale*, pages 97–112. La Découverte.

[Peuquet, 1994] Peuquet, D. J. (1994). It's about time: A conceptual framework for the representation of temporal dynamics in geographic in-

formation systems. *Annals of the Association of American Geographers*, 84(3):441–461.

[Pierre and Verger, 1970a] Pierre, G. and Verger, F. (1970a). Dictionnaire de la géographie. *Paris, Presses Universitaires*.

[Pierre and Verger, 1970b] Pierre, G. and Verger, F. (1970b). Dictionnaire de la géographie. *Paris, Presses Universitaires*.

[Poelmans and Rompaey, 2010] Poelmans, L. and Rompaey, A. V. (2010). Complexity and performance of urban expansion models. *Computers, Environment and Urban Systems*, 34(1):17 – 27.

[Pumain and Saint-Julien, 1997] Pumain, D. and Saint-Julien, T. (1997). L'analyse spatiale. 1. localisations dans l'espace. paris: Armand colin, coll.«. *Cursus*.

[Qi et al., 2004] Qi, Y., Henderson, M., Xu, M., Chen, J., Shi, P., He, C., and Skinner, G. W. (2004). Evolving core-periphery interactions in a rapidly expanding urban landscape: The case of beijing. *Landscape Ecology*, 19(4):375–388.

[Qiang and Lam, 2015] Qiang, Y. and Lam, N. S. N. (2015). Modeling land use and land cover changes in a vulnerable coastal region using artificial neural networks and cellular automata. *Environmental Monitoring and Assessment*, 187(3):57.

[Quine, 1953] Quine, W. V. O. (1953). Three grades of modal involvement. In *Journal of Symbolic Logic*, pages 168–169. North-Holland Publishing Co.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[Rai et al., 2012] Rai, D., Verma, K., and Thoke, A. S. (2012). Article: Classification algorithm based on ms apriori for rare classes. *International Journal of Computer Applications*, 48(22):52–56. Full text available.

[Ramirez, 1998] Ramirez, J. R. (1998). Revision of geographic data: A framework. In *Proceedings ISPRS Commission IV Symposium GIS-Between Visions and Applications, Stuttgart*, pages 487–493. Citeseer.

[Reitsma and Dubayah, 2007] Reitsma, F. and Dubayah, R. (2007). Simulating watershed runoff with a new data model. *Hydrological processes*, 21(18):2447–2457.

[Reitsma, 2005] Reitsma, F. E. (2005). *A new geographic process data model*. PhD thesis.

[Renolen, 1996] Renolen, A. (1996). History graphs: conceptual modeling of spatio-temporal data. *Proceedings of GIS Frontiers in Business and Science*.

[Rodier and Saligny, 2010] Rodier, X. and Saligny, L. (2010). Modélisation des objets historiques selon la fonction, l'espace et le temps pour l'étude des dynamiques urbaines dans la longue durée. *Cybergeo: European Journal of Geography*.

[Romero et al., 2010] Romero, C., Ventura, S., Vasilyeva, E., and Pechenizkiy, M. (2010). Class association rules mining from students' test data. In *Educational Data Mining 2010*.

[Roshannejad and Kainz, 1986] Roshannejad, A. and Kainz, W. (1986). Handling identities in spatio-temporal databases. In *Proc. of ACSM/ASPRS 1995 Annual Convention and Exposition Tech*.

[Rousseaux and Ritschard, 2014] Rousseaux, E. and Ritschard, G. (2014). An association rule miner for unbalanced data based on artificial bee colony optimization. In *COMPSTAT 2014*.

[Rudin et al., 2013a] Rudin, C., Letham, B., and Madigan, D. (2013a). Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14:3441–3492.

[Rudin et al., 2013b] Rudin, C., Letham, B., and Madigan, D. (2013b). Learning theory analysis for association rules and sequential event prediction. *Journal of Machine Learning Research*, 14:3441–3492.

[Ryan and Brown, 2013] Ryan, C. and Brown, K. (2013). Predicting occupant locations using association rule mining. In Bramer, M. and Petridis, M., editors, *Research and Development in Intelligent Systems XXX*, pages 63–77. Springer International Publishing.

[Sarangi and Sahoo, 2013] Sarangi, B. and Sahoo, L. (2013). A survey on spatial association rule mining technique and algorithms for mining spatial data. *International Journal of Scientific & Engineering Research*, 4:1664–1670.

[Sarath and Ravi, 2013] Sarath, K. and Ravi, V. (2013). Association rule mining using binary particle swarm optimization. *Engineering Applications of Artificial Intelligence*, 26(8):1832 – 1840.

[Savasere et al., 1995] Savasere, A., Omiecinski, E., and Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. In *Proceeding of the 1995 international conference on very large data bases (VLDB'95)*, pages 432–444.

[Şener et al., 2010] Şener, Ş., Şener, E., Nas, B., and Karagüzel, R. (2010). Combining ahp with gis for landfill site selection: a case study in the lake beyşehir catchment area (konya, turkey). *Waste Management*, 30(11):2037–2046.

[Shi and Shibasaki, 2000] Shi, Z. and Shibasaki, R. (2000). Gis database revision–the problems and solutions. *International Archives of Photogrammetry and Remote Sensing*, 33(B2; PART 2):494–501.

[Silva and Wu, 2012] Silva, E. and Wu, N. (2012). Surveying models in urban land studies. *CPL bibliography*, 27(2):139–152.

[Silva et al., 2008] Silva, E. A., Ahern, J., and Wileden, J. (2008). Strategies for landscape ecology: An application using cellular automata models. *Progress in Planning*, 70(4):133–177.

[Silva and Wu, 2014] Silva, E. A. and Wu, N. (2014). Dg-abc: An integrated multi-agent and cellular automata urban growth model. In *Technologies for Urban and Spatial Planning: Virtual Cities and Territories*, pages 57–92. IGI Global.

[Simmonds, 1999] Simmonds, D. C. (1999). The design of the delta land-use modelling package. *Environment and Planning B: Planning and Design*, 26(5):665–684.

[Sipper, 2001] Sipper, M. (2001). *Evolution of Parallel Cellular Machines: The Cellular Programming Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Snodgrass and Ahn, 1985] Snodgrass, R. and Ahn, I. (1985). A taxonomy of time databases. *SIGMOD Rec.*, 14(4):236–246.

[Snodgrass, 1992] Snodgrass, R. T. (1992). *Temporal databases*, pages 22–64. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Spéry et al., 2001a] Spéry, L., Claramunt, C., and Libourel, T. (2001a). A spatio-temporal model for the manipulation of lineage metadata. *GeoInformatica*, 5(1):51–70.

[Spéry et al., 2001b] Spéry, L., Claramunt, C., and Libourel, T. (2001b). A spatio-temporal model for the manipulation of lineage metadata. *Geoinformatica*, 5(1):51–70.

[Sriti et al., 2005] Sriti, M., Thibaud, R., and Claramunt, C. (2005). A fuzzy identity-based temporal gis for the analysis of geomorphometry changes. In *Journal on Data Semantics III*, pages 81–99. Springer.

[Stell, 2003] Stell, J. G. (2003). Granularity in change over time. *Foundations of geographic information science*, pages 95–115.

[Stillwell et al., 2013] Stillwell, J., Geertman, S., and Openshaw, S. (2013). *Geographical Information and Planning: European Perspectives*. Springer Science & Business Media.

[Tayyebi, 2013] Tayyebi, A. (2013). Simulating land use land cover change using data mining and machine learning algorithms.

[Thabtah et al., 2004] Thabtah, F. A., Cowling, P., and Peng, Y. (2004). Mmac: a new multi-class, multi-label associative classification approach. In *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, pages 217–224.

[Thériault et al., 1999] Thériault, M., Claramunt, C., and Villeneuve, P. (1999). A spatio-temporal taxonomy for the representation of spatial set behaviours. In *Spatio-temporal database management*, pages 1–18. Springer.

[Thériault and Claramunt, 1999] Thériault, M. and Claramunt, C. (1999). La représentation du temps et des processus dans les sig: une nécessité pour la recherche interdisciplinaire. *Représentation de l'espace et du temps dans les SIG, Revue internationale de géomatique*, 9:67–99.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

[Toivonen, 1996] Toivonen, H. (1996). Sampling large databases for association rules. In *Proceedings of the 22th International Conference on Very Large Data Bases*, VLDB '96, pages 134–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Turner et al., 2007] Turner, B. L., Lambin, E. F., and Reenberg, A. (2007). The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences*, 104(52):20666–20671.

[UN-Habitat, 2011] UN-Habitat (2011). Annual report 2010. .

[van Dam and van de Velden, 2015] van Dam, J.-W. and van de Velden, M. (2015). Online profiling and clustering of facebook users. *Decision Support Systems*, 70:60 – 72.

[Verburg et al., 2004] Verburg, P. H., Schot, P. P., Dijst, M. J., and Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *GeoJournal*, 61(4):309–324.

[Vijayalakshmi and Pethalakshmi, 2015] Vijayalakshmi, V. and Pethalakshmi, A. (2015). An efficient count based transaction reduction approach for mining frequent patterns. *Procedia Computer Science*, 47:52 – 61. Graph Algorithms, High Performance Implementations and Its Applications ({ICGHIA} 2014).

[Voorhees, 1959] Voorhees, A. M. (1959). Land use and traffic models: a progress report. *J American Institute of Planners*, 25(2):55–57.

[Waddell and Ulfarsson, 2004] Waddell, P. and Ulfarsson, G. F. (2004). Introduction to urban simulation: design and development of operational models. *Handbook in Transport*, 5:203–236.

[Walter and Fritsch, 2000] Walter, V. and Fritsch, D. (2000). Automated revsion of gis databases. In *Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, pages 129–134. ACM.

[Wang et al., 2002] Wang, H., Wang, W., Yang, J., and Yu, P. S. (2002). Clustering by pattern similarity in large data sets. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, SIGMOD '02, pages 394–405, New York, NY, USA. ACM.

[Wang and Karypis, 2005] Wang, J. and Karypis, G. (2005). *HARMONY: Efficiently Mining the Best Rules for Classification*, pages 205–216.

[Wang et al., 2016] Wang, Q., Davis, D. N., and Ren, J. (2016). Mining frequent biological sequences based on bitmap without candidate sequence generation. *Computers in Biology and Medicine*, 69:152 – 157.

[Ward et al., 2000] Ward, D., Murray, A., and Phinn, S. (2000). A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems*, 24(6):539 – 558.

[Weidner and Hunt, 2006] Weidner, T., R. D. J. F. J. E. A. and Hunt, J. D. (2006). Tlumip-transport land use model in portland - current state. In *In Stadt Region Land*, pages 91–102. Aachen: Institut fur Stadtbauwesen und Stadtverkehr, RWTH Aachen.

[Weiss, 1999] Weiss, G. (1999). *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. Intelligent Robotics and Autonomous Agents Series. CogNet.

[White and Engelen, 1997] White, R. and Engelen, G. (1997). Cellular automata as the basis of integrated dynamic regional modelling. *Environment and Planning B: Planning and Design*, 24(2):235–246.

[White and Engelen, 2000] White, R. and Engelen, G. (2000). High-resolution integrated modelling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems*, 24(5):383 – 400.

[Witten and Frank, 2005] Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

[Woo, 2012] Woo, J. (2012). Apriori-map/reduce algorithm. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[Worboys, 2005] Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28.

[Worboys and Hornsby, 2004] Worboys, M. and Hornsby, K. (2004). From objects to events: Gem, the geospatial event model. In *GIScience*, volume 3234, pages 327–344. Springer.

[Worboys, 1992] Worboys, M. F. (1992). Model for spatio-temporal information. In *Proceedings of the 5th International Symposium on Spatial Data Handling*, pages 602–611, Charleston, South Carolina, USA.

[Worboys, 1998] Worboys, M. F. (1998). *A generic model for spatio-bitemporal geographic information*, pages 25–39. Oxford University Press.

[Worboys, 2001] Worboys, M. F. (2001). Modelling changes and events in dynamic spatial systems with reference to socio-economic units.

[Xuewu et al., 2008] Xuewu, Z., Fenzhen, S., Yunyan, D., and Yishao, S. (2008). Association rule mining on spatio-temporal processes. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*, pages 1–4.

[Yadav et al., 2012] Yadav, P., Kapoor, M., and Sarma, K. (2012). Land use land cover mapping, change detection and conflict analysis of nagzira-navegaon corridor, central india using geospatial technology. *International Journal of Remote Sensing and GIS*, 1(2):90–98.

[Yang et al., 2008] Yang, Q., Li, X., and Shi, X. (2008). Cellular automata for simulating land use changes based on support vector machines. *Comput. Geosci.*, 34(6):592–602.

[Yeh, 1999] Yeh, A. G.-O. (1999). Urban planning and gis. *Geographical information systems*, 2:877–888.

[Yin and Han, 2003] Yin, X. and Han, J. (2003). *CPAR: Classification based on Predictive Association Rules*, pages 331–335.

[Yuan, 1999] Yuan, M. (1999). Use of a three-domain repesentation to enhance gis support for complex spatiotemporal queries. *Transactions in GIS*, 3(2):137–159.

[Yuan, 2001] Yuan, M. (2001). Representing complex geographic phenomena in gis. *Cartography and Geographic Information Science*, 28(2):83–96.

[Zaki, 1998] Zaki, M. J. (1998). Efficient enumeration of frequent sequences. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 68–75, New York, NY, USA. ACM.

[Zaki and jui Hsiao, 2002] Zaki, M. J. and jui Hsiao, C. (2002). Charm: An efficient algorithm for closed itemset mining. In *. In: Proceeding ofthe2002SIAMinternationalconferenceondatamining(SDM'02)*, pages 457–473, Arlington,VA.

[Zaki and Wagner Meira, 2014] Zaki, M. J. and Wagner Meira, J. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

[Zhang et al., 2015] Zhang, Y., Chen, M., and Liu, L. (2015). A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 681–685.

[Zhao et al., 2014] Zhao, X. W., Guo, Y., He, Y., Jiang, H., Wu, Y., and Li, X. (2014). We know what you want to buy: A demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1935–1944, New York, NY, USA. ACM.