REPUBLIQUE ALGERIENNE DEMOCATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE SAAD DAHLAB DE BLIDA 1

FACULTE DES SCIENCES

DEPARTEMENT D'INFORMATIQUE

Mémoire de fin d'études proposé pour l'obtention d'un master en Informatique

Option : Génie des Systèmes Informatiques

Framework de nettoyage de données incomplètes basé sur les réseaux bayésiens

Promotrice :
PhD Zhara Fatma Zohra

Présenté par :
M. Diané Moussa
M. Sall Cheick Oumar

Devant le jury composé de :

Mme Arkam	**Présidente**
Mme Mancer	**Examinatrice**
Mme Rezoug Nachida	**Examinatrice**

Blida, Septembre 2014

Remerciement

*Louange à **ALLAH** le clément, le miséricordieux,*
qui nous a donné le courage et la patience de mener
à bien ce travail.

A notre très chère promotrice ZAHRA Fatma Zohra
Nous avons eu le privilège de travailler avec vous
et d'apprécier vos qualités et vos valeurs.
Votre sérieux, votre compétence et votre sens du devoir
nous ont énormément marqués.
Veuillez trouver ici l'expression de notre respectueuse
considération et notre profonde admiration pour toutes vos
qualités scientifiques et humaines.
Ce travail est pour nous l'occasion de vous témoigner
notre profonde gratitude.

A notre présidente du jury Mme Arkam
Vous nous faites l'honneur d'accepter avec une très
grande amabilité de siéger parmi notre jury de soutenance.
Veuillez accepter ce travail maître, en gage de notre
grand respect et notre profonde reconnaissance.

Notre examinatrice Mme Mancer
Vous nous avez honorés d'accepter avec grande
sympathie de siéger parmi notre jury de soutenance.
Veuillez trouvez ici l'expression de notre grand respect et nos
vifs remerciements.

A notre examinatrice Mme Rezoug Nachida
Vous nous faites l'honneur d'accepter avec une très
grande amabilité de siéger parmi notre jury de soutenance.
Veuillez accepter ce travail maître, en gage de notre
grand respect et notre profonde reconnaissance.

Merci très infiniment au peuple algérien de m'avoir logé, nourrit,
former toutes ces années, ses aides ont été cruciales pour moi
sans lesquels je ne pouvais espérer croire en un avenir meilleur.

Nous adressons nos vifs remerciements :
A nos chers parents
Au corps professoral du département d'informatique de l'université
Saad Dahlab Blida 1 qui nous ont accompagnés au cours de ces années
d'étude.
A tous ces gens que nous n'avons pas pu évoqué qu'ils trouvent ici l'expression de nos
Profondes reconnaissances.

Dédicace

Louange à DIEU le Miséricordieux.

A cœur vaillant rien d'impossible
A conscience tranquille tout est accessible
Quand il y a la soif d'apprendre
Tout vient à point à qui sait attendre
Quand il y a le souci de réaliser un dessein
Tout devient facile pour arriver à nos fins
Malgré les obstacles qui s'opposent
En dépit des difficultés qui s'interposent

Les études sont avant tout
Notre unique et seul atout
Ils représentent la lumière de notre existence
L'étoile brillante de notre réjouissance
Comme un vol de gerfauts hors du charnier natal
Nous partons ivres d'un rêve héroïque et brutal
Espérant des lendemains épiques
Un avenir glorieux et magique
Souhaitant que le fruit de nos efforts fournis
Jour et nuit, nous mènera vers le bonheur fleuri

Aujourd'hui, ici rassemblés auprès des jurys,
Nous prions dieu que cette soutenance
Fera signe de persévérance
Et que nous serions enchantés
Par notre travail honoré

Je dédie ce modeste travail à ...

A ma très chère mère Kadia Founé Traoré
Affable, honorable, aimable : Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.
Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études.

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que tu mérites pour tous les sacrifices que tu n'as cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.
Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

Je te dédie ce travail en témoignage de mon profond

amour. Puisse DIEU, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.

A la mémoire de mon Père Feu Moussa
Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous.
Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être.

Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation.

A mon très cher frère Hamidou Diané
Mon cher frère qui m'est le père et la mère, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous.
Mon ange gardien et mon fidèle compagnant dans les moments les plus délicats de cette vie mystérieuse.
Je vous dédie ce travail avec tous mes voeux de bonheur, de santé et de réussite.

A tous les membres de ma famille, petits et grands
Veuillez trouver dans ce modeste travail l'expression de mon Affection

A ma chère et dynamique mère Mme Rezoug Nachida
Un remerciement particulier et sincère pour tous vos efforts fournis. Vous avez toujours été présente.
Que ce travail soit un témoignage de ma gratitude et mon profond respect.

A toute la famille Diané, Traoré, Konaté, Diallo, Touré, Sinaba etc.

A Mes très chèr(e)s ami(e)s

A ma seule et unique famille, la communauté Malienne de Blida

A toute la communauté étrangère

A tous mes ami(e)s de la promotion 2009

A toute ma promotion de Master de l'université de Blida 1.

A tous les internes et résidents des résidences universitaires de Blida.

Résumé

Les données issues du monde réel ne sont pas toujours complètes ce qui constitue un problème majeur puisque l'information à disposition est incomplète et donc moins fiable et ceci semble constituer un phénomène aussi imprévisible qu'inévitable. Ce phénomène se manifeste lorsque les valeurs n'ont pas pu être observées, elles ont été perdues ou elles n'ont pas été enregistrées. Notre projet s'inscrit dans le cadre du traitement de ces données incomplètes. Il existe une abondante littérature sur les méthodes de traitement des données incomplètes. Dans notre travail, nous allons utiliser la méthode d'imputation à travers les réseaux bayésiens en passant par les étapes suivantes : apprentissage de la structure en utilisant l'approche de Friedman et Goldszmidt, l'apprentissage des paramètres et pour finir l'inférence pour pouvoir imputer les données incomplètes.

MOTS-CLES : Données incomplètes, Réseaux bayésiens, Apprentissage structure, Inférence, Imputation.

مجردة

بيانات عن العالم الحقيقي لا يكتمل دائما ما يمثل مشكلة كبيرة لأن المعلومات المتاحة غير مكتملة، وبالتالي أقل موثوقية ويبدو أن هذا ظاهرة غير متوقعة وحتمية. تحدث هذه الظاهرة عندما لم وحظ القيم، وأنها فقدت أو أنها لم تسجل. مشروعنا هو في علاج هذه البيانات غير مكتملة. هناك دراسات واسعة النطاق على طرق معالجة البيانات غير مكتملة. في عملنا، ونحن نستخدم احتساب من خلال شبكات النظرية الافتراضية من خلال خطوات تعلم بنية باستخدام نهج فريدمان وGoldszmidt، المعلمات التعلم والاستدلال أخيرا ل ليحسب في عداد المفقودين البيانات.

الكلمات الرئيسية:.احتساب الاستدلال، هيكل، التعلم الافتراضية النظرية شبكة مكتملة، غير البيانات

Abstract

The data resulting from the real-world are not always complete what constitutes a main issue since information at disposal is incomplete and thus less reliable and this seems to constitute a phenomenon as unforeseeable as inevitable. This phenomenon appears when the values could not be observed, they were lost or they were not recorded. Our

project lies within the scope of the processing of these incomplete data. There exists an abundant literature on the incomplete methods of information processing. In our work, we will use the method of charge through the networks bayesians while passing by the following stages: training of the structure by using the approach of Friedman and Goldszmidt, the training of the parameters and to finish the inference to be able to charge the missing data.

KEYWORDS: Missing data, Networks bayesians, Structure learning, Inference, Imputation.

Sommaire

Chapitre III : Les Réseaux bayésiens

Chapitre IV : Approche proposée et expérimentation

Liste des figures

Chapitre I : Qualité des données

Chapitre II : Traitement des données manquantes

Chapitre III : Les Réseaux bayésiens

Chapitre IV : Approche proposée et expérimentation

Liste des tableaux

Chapitre I : Qualité des données

Chapitre II: Traitement des données manquantes

Introduction générale

1. Contexte du travail

Aujourd'hui, les activités de prise de décisions au sein des différentes organisations sont basées sur les informations obtenues à partir de l'analyse des données. Etant données que ces données sont des ressources très importantes et qu'elles sont collectées par différents Mobiles devices, il peut arriver que ces données soient de sources différentes, hétérogènes, de formats différents, incertaines, incomplètes, avec une perte ou erreur de qualité, la répétition des données sans oublier leurs crédibilités. Il se peut que certains appareils de collecte manquent de fonctions de vérification ou ont une vérification faible de fonction, que le milieu et la température jouent sur le résultat de certains appareils. La qualité des données (QoD, Quality of data) devient alors un élément primordial et critique pour les gestionnaires et les méthodes de traitement de données, et doit être contrôlée et mesurée. En outre, les données de haute qualité peuvent accroître les possibilités de prendre de bonnes décisions, dans des domaines différents. Dans notre travail, l'accent sera mis sur le problème des données incomplètes et leur traitement.

De façon générale, le problème de données incomplètes est présent depuis plusieurs décennies. Ce problème se manifeste lorsque les valeurs n'ont pas pu être observées, elles ont été perdues ou elles n'étaient pas enregistrées. La présence de ces données entraine un dysfonctionnement du processus de traitement de données. Les données manquantes constituent un problème majeur, parce que l'information à disposition est incomplète et donc moins fiable. Le traitement de ces données incomplètes est un problème concret et toujours embarrassant lorsqu'il s'agit de données réelles.

Les techniques classiques de data mining (classification, clustering, association etc.) ne sont pas conçues pour prendre en compte les données incomplètes. En effet nous pouvons citer deux situations où les algorithmes de classification classique présentent des limitations. La première lorsque l'ensemble des données à traiter est de taille trop importante et la seconde lorsque la base d'apprentissage est incomplète.

2. Problématique

En ce qui concerne le nettoyage de ces données, la plupart des solutions se contente de retirer des exemples avec des données manquantes, ce qui ne permet pas de prendre en considération l'ensemble de la base et introduit de nouveaux biais dans l'analyse. En effet pour éviter de supprimer ainsi les données, on peut remplacer une valeur manquante par la

moyenne de la variable correspondante mais cette moyenne est une très mauvaise approximation dans le cas où la variable présente une grande dispersion, cette solution n'est plus satisfaisante car cela exagère les corrélations cet effet nous nous baserons sur procédures et stratégies d'estimation des valeurs des données manquantes basés sur les réseaux Bayésiens.

3. Objectif

Dans le but de traiter des masses de données afin d'avoir une meilleure qualité des données, il est indispensable de traiter la problématique de la validité des données. Le nettoyage des données incomplètes est un parmi les facteurs de la qualité des données.

L'objectif de notre travail est de proposer un framework de nettoyage des masses de données incomplètes en temps réel. Pour ce faire, nous avons proposé une méthode qui se base sur les réseaux Bayésiens pour l'apprentissage séquentiel du modèle d'imputations des valeurs manquantes à partir des données incomplètes. La stratégie permettant d'effectuer cette mise à jour séquentielle consiste à utiliser les techniques d'apprentissage incrémental développées avec des règles de mise à jour récursives et grâce à ces techniques, les informations portées par les nouvelles données sont incorporées séquentiellement dans le model d'apprentissage sans réutiliser ou (très peu) les anciennes données. Et en utilisant l'inférence des réseaux Bayésiens, on estime les valeurs à imputées.

4. Organisation du mémoire

Ce mémoire a été reparti en 4 chapitres, le premier chapitre portera sur la qualité de données en donnant un aperçu sur le problème de la qualité de données et les principales stratégies, techniques, méthodes utilisées.

Le second chapitre est consacré au traitement de données incomplètes qui contiendra les différentes méthodes de traitement de données incomplètes. Cette étude nous permettra de faire le choix de notre méthode de traitement.

Le troisième chapitre présentera les réseaux bayésiens et leurs différentes méthodes d'inférence et d'apprentissage de la structure et des paramètres.

Le dernier chapitre sera dédié à l'approche proposée et expérimentation.

Et enfin nous terminerons par une conclusion générale et la proposition de quelques perspectives et des références bibliographiques.

CHAPITRE I

QUALITÉ DES DONNÉES

Chapitre I : Qualité des données

I.1 Introduction

La croissance explosive de l'E-gouvernement, E-business, et les bases de données scientifiques ont créé le besoin d'une nouvelle génération des techniques et d'outils dans le but de la découverte des connaissances à partir des données de façon intelligente et automatisée [1]. Ces techniques utilisées pour l'amélioration de la qualité de données, sont employées généralement pour identifier les redondances, les valeurs manquantes, l'inconsistance des données, les erreurs et les éliminer [2].

I.2 Données, information

Ces deux notions au cœur de la théorie de l'information se recoupent et les spécialistes du domaine ne les définissent pas de la même manière. Mais dans notre étude, on retiendra les définitions suivantes :

Les données sont des faits ou statistiques pouvant être quantifiés, comptés, mesurés et stockés [3].

Une information est définie comme étant l'ensemble de données fourni avec le contexte nécessaire pour la prise de décision [3].

Les données sont les éléments de base de tous les systèmes et processus d'une entreprise, que ce soit dans les systèmes transitionnels ou les référentiels. Elles sont au centre de l'exploitation de l'entreprise. Les données sont généralement requises dans tous les services de l'entreprise. Le cycle de vie des données est un ensemble d'activités qui permet de représenter et de suivre l'état d'une donnée depuis sa planification jusqu'à sa suppression.

Figure I.1 : Cycle de vie des données [4]

Il existe trois types de données selon [4]: les données structurées, semi structurées, et non structurées.

a) **Données structurées :** Les données sont dites structurées si elles peuvent être stockés dans des champs de base de données et donc entrer dans un modèle de données clairement défini. Il s'agit la plupart du temps des valeurs numériques ou des chaines alphanumériques (ex : l'adresse d'un abonné à un journal, tables relationnelles, données statistiques).

b) **Données semi structurées :** Les données semi structurées sont des données ne répondant pas un schéma fixe : elles ne peuvent pas être stockées dans des bases de données relationnelles (exemple : Corps d'un email).Elles sont structurées mais leur structure est implicite et irrégulière. L'exemple type est un ensemble de pages web (exemple : données XML).

c) **Données Non structurées :** Les données non structures sont définis par *Bill Inmon*, l'un des pères des data warsehouses « tout document, fichier, image, rapport, formulaire, etc. qui n'a pas de structure standard définie qui permettrait de le stocker facilement dans un dispositif de traitement automatisé. Il ne peut pas être défini en termes de lignes et de colonnes ou enregistrements. Les données non structurées sont des e-mails, les feuilles de calcul, les documents etc. Certaines des informations les plus précieuses de l'entreprise résident dans ses données non structurées ».

I.3 Qualité de données

Une fois les données définies, nous pouvons expliciter ce qui fait leur qualité. Le terme « contrôle de qualité » a été popularisé par *Deming [5]* au cours des années 1950. Il a défini quatorze recommandations de gestion. Ces recommandations ont par la suite été le fondement de plusieurs auteurs. Ce sont les entreprises japonaises qui furent les premières à utiliser ces recommandations. En 1954, *Juran [5]* a également contribué à établir des bases pour la gestion de la qualité en proposant 10 étapes pour établir un contrôle de qualité. Au même moment, Ishikawa de son côté, préconise une amélioration constante de la qualité basée sur un diagramme de cause à effet qu'il nomme « fishbone », viennent ensuite les années 1970, durant lesquelles les États-Unis ont mis en place des pratiques qui par la suite, au cours des années 1980, sont devenues le TQM (*Total Quality Management*). Cette méthode est le fruit d'un amalgame crée par les grands auteurs, tels que Deming, Juran et Crosby. Puis, c'est au tournant des années

2000 que la gestion des systèmes d'information a vu le jour avec l'IQM (*Information Quality Management*).

Les données sont dites de qualité si elles satisfont aux exigences de leurs utilisateurs. En d'autres termes, la qualité des données dépend autant de leur utilisation que de leur état *[5]*.

Définir la qualité de données n'est pas simple et il est souvent plus aisé de tenter de définir la non qualité mais on peut retenir la définition suivante : « La qualité de données désigne l'aptitude de l'ensemble de caractéristique intrinsèques des données (fraicheur , disponible, cohérence, fonctionnelle et ou technique, traçabilité, sécurisation, exhaustivité) à satisfaire ces exigences internes (pilotage de donnés, prise de décision …) et des exigences externes (règlementations) à l'organisation. Ces critères sont appliqués dans les démarches normées d'audit également connu sous le nom de « CAVAR » (completness, Accuracy, Validity, Availability, Restricted access) *[3]*.

Le but est d'obtenir des données sans doublon, complètes, sans fautes d'orthographes, sans omission, sans variation superflue et conforme à la structure définie.

I.4 Problèmes de la qualité de données

Le problème de la qualité des données peut être divisé généralement en deux classes qui sont : les données provenant d'une seule source et celles qui proviennent de multiple sources. Selon *[6]*, quatre catégories pour la qualité de données sont identifiés dans la table suivante.

Problème de la qualité des données	Catégorie	Définition
Problème Simple source	Niveau de schéma	Manque de contraintes d'intégrité, créateur de schéma pauvres , Contraintes d'unicité, Intégrité référentielle
	Niveau d'exemple	Erreur de saisie de données, Fautes d'orthographes, redondance de doubles, Valeur contradictoire

Problème multiple source	Niveau de schéma	Modèle et schéma de données hétérogène, conception, conflits de noms
	Niveau d'exemple	Contradiction superposant, inconsistance des données, inconsistance d'agglomération, Inconsistance de Synchronisation

Tableau I.1 : Classification des problèmes de la qualité des données [6]

I.5 Stratégies et techniques de la qualité des données

Il existe deux types de stratégies pour les étapes d'amélioration de la qualité des données, à savoir Data Driven et Process Driven.Chaque stratégie utilise des techniques variées [7] dont le but est l'amélioration de la qualité de données.

Une série de techniques est appliquée à ces deux stratégies : algorithmes, heuristique, et activités basées sur la connaissance.

I.5.1 Data-driven

Cette stratégie améliore la qualité des données en modifiant directement la valeur des données. Quelques techniques appliquée dans ces stratégies sont: acquisition de nouvelles données, la standardisation ou la normalisation, la liaison d'enregistrement, l'intégration de données et de schéma, la source de fidélité, la localisation et correction d'erreur et l'optimisation du coût [7].

I.5.2 Process-Driven

Cette dernière l'améliore en reconcevant les processus qui créent ou modifient des données. Il se compose de deux techniques principales: Process control et process redesign. En effet, dans le contrôle de processus, est inséré les procédures de vérification et de contrôle dans le processus de fabrication de données, tandis que le Process redesign reconçoit des processus afin d'enlever les causes de la mauvaise qualité et introduit les nouvelles activités qui produisent des données de plus haute qualité [7].

En général, le process-driven est mieux à l'exécution que le Data-driven en longue période, car ils éliminent les causes d'origines des problèmes de la qualité complètement. Ce pendant d'un point de vue court terme la conception des processus peut être extrêmement cher. En revanche, les stratégies Data driven sont enregistrées d'être de cout efficace à court terme mais cher en long terme [7]. Une autre classification des données est basée sur la rigueur pour mesurer et d'assurer la qualité des données, qui a deux classes spécifiquement : données élémentaires et données agrégées. Dans une organisation, les données gérées par le processus opérationnel et représentées par des phénomènes atomiques du monde réel sont appelés des données élémentaires (par exemple, le sexe, l'âge), Bien que les données qui sont collectées à partir des données élémentaires pour appliquer la fonction d'agrégation, est appelé des données agrégées (par exemple, le revenu moyen qu'un payeur de l'impôt paye dans une ville précise) [8].

I.6Méthodes, Modèles et Mesures de la qualité de données

Il existe un éventail de méthodes appropriées pour la recherche de la qualité de données, parmi lesquelles on peut citer :recherche-action,Intelligence Artificielle ,raisonnement à base de cas, Data Mining ,Design Science ,économétrie ,empirique ,expérimentale ,Modélisation mathématique ,qualitative ,quantitative ,Analyse statistique ,Le système de conception, mise en œuvre ,théorie et preuves formelles etc.

Différents modèles de la qualité de données ont été proposés pour tenir compte de tous les aspects de données lors de la conception des Frameworks.

Rye et Kyung-seok [9] propose un modèle de maturité de gestion de la qualité des données qui se compose de quatre étapes. La phase initiale travaille au niveau du système de gestion de base de données. La deuxième phase est basée sur le modèle de données logique et physique pour contrôler la qualité de données. La troisième phase est la standardisation de la gestion des données pour résoudre le problème de l'intégration des différentes sources de données et la gestion des métadonnées appartenant également à cette phase. La quatrième phase est l'optimisation de la gestion des données de macro perspectives de l'architecture de données. Dans le traitement des affaires d'une entreprise, le problème de la qualité des données peut apparaître et se propager d'une phase à une autre.

Bagchi S et al. [10] étend le modèle d'entreprise existant avec le contrôle de la propagation d'erreur et fournit l'estimation de la qualité de données. Le modèle est appliqué dans le domaine de la comptabilité et d'audit pour contrôler et corriger les erreurs de transactions et améliore par la suite la qualité des données. La gestion du cycle de vie des données contrôle à partir de la vue de données plutôt qu'application d'entreprise. Il est généralement constitué de trois phases: la phase de conception, phase de fabrication et de la phase d'application. Un Framework distribué basé sur le modèle du cycle de vie basic est conçu pour améliorer la qualité des données [11]. À la phase de conception et de fabrication, des contraintes d'intégrité, comme contraintes de clé primaire, les règles métiers, les règles logiques et fonction dépendante, sont appliqués à prendre la détection de données, analyse et modification au niveau de l'instance. A la phase d'application des données intégrées, les technologies ETL sont appliquées pour intégrer des sources de données distribuées au niveau arrangement.

Lucas. A [12] propose le concept de méta-Framework et de construire un Framework de gestion de données de bas en haut pour l'entreprise entière, y compris les sujets ,les stratégies de communication, les dimensions, les méthodologies ,les techniques et les outils de la qualité des données.

Ils existent deux types de mesures d'amélioration et d'assurance de la qualité des données : mesures préventives et mesures curatives.

Les mesures préventives évitent les problèmes de la qualité des données en prenant certaines précautions pour contrôler la qualité de telle sorte que les conditions nécessaires qui causent des problèmes ne peuvent pas être installées. L'utilisation de quelques stratégies de vérification des données à l'entrée de l'application peut réduire des données pauvres à grande échelle. [13] propose un modèle de probabilité basée sur la théorie des réseaux bayésiens et le gain de l'information gourmand pour vérifier la qualité des données à l'entrée du système d'information. Premièrement, avant la collecte des données, le modèle trie les éléments input-form pour améliorer l'efficacité de la collecte. Pendant l'entrée des données dans le système, le modèle doit connaitre au mieux certaines informations de la qualité des données selon les valeurs de données et le message en temps réel et ajuste alors dynamiquement la forme. Après que les données soient entrées dans le système, le modèle identifiera les erreurs possibles et demandera à l'utilisateur de reconfirmer. Le modèle prend trois phases de vérification pour le

contrôle de la qualité des données. Des mesures correctives sont prises en cas où un problème de données se produit.

La technologie Data Cleaning est une mesure préventive dans le prétraitement des données ou une mesure corrective pour faire face aux problèmes de données. La technologie Data Mining s'applique habituellement dans des outils de réparation.

I.7Dimensions de la qualité de données

Une dimension en qualité des données correspond à une caractéristique d'une donnée qui permet de la classifier et d'en définir les besoins au niveau de sa qualité. Les dimensions sont principalement utilisées pour définir, mesurer et gérer la qualité de celles-ci [14]. Elles sont représentées par des termes, tels que fraicheur, disponibilité, cohérence, exhaustivité etc. Une dimension est définie par un ensemble d'attributs qui représente un certain aspect de la donnée [15]. La définition attribuée à chacune des dimensions peut être de nature théorique ou opérationnelle [16].

Depuis plusieurs années, certaines entreprises ont mis en place des pratiques, des techniques de mesure et de contrôle en qualité des données. Certaines de ces entreprises ont rendu disponible sur le web la méthode qu'elles ont mise en place et qui leur a permis de planifier, d'évaluer, d'analyser, de mesurer et de contrôler la qualité de leurs données. Il existe également des auteurs qui ont publié des articles et des livres portant sur certains aspects liés à cette qualité. Notons entre autres les documents de l'ICIS (Institut canadien d'information sur la santé), l'IMF (*International Monetary Fund*), l'ICES (*Institute for Clinical Evaluative Sciences*) et l'EPA (*United States Environmental Protection Agency*) [15].

Vu l'importance accordée aux dimensions, il est aussi d'une importance primordiale de procéder à une sélection des dimensions pertinentes. D'autant plus que la littérature contient un très grand nombre de dimensions ayant été décrites au fil des années. Ces dimensions demeurent un atout essentiel pour procéder à une analyse qualitative des données. La sélection des dimensions est le point de départ de tout processus lié à la qualité des données. Ce sont les dimensions sélectionnées qui permettront d'établir des concepts plus matures en vue de définir les caractéristiques d'une donnée de qualité, de définir des mesures pertinentes et de mettre en place des processus de contrôle fiables [17].

[3] a détaillé quelques dimensions de la qualité de données qui sont les suivantes :

- Fraicheur : elle est essentielle pour avoir une bonne vision d'une situation à un instant et pour prendre de bonnes décisions.
- Disponibilité : ce concept recouvre deux notions : l'accessibilité, d'une part et d'autre part la trouvabilité.
- Accessibilité : elle dépend de plusieurs facteurs parmi lesquels on peut citer : la robustesse, le classement des données, la présentation des données, l'assistance aux utilisateurs.
- Trouvabilité : L'information peut être disponible, accessible grâce à des outils de requête et pourtant ne pas atteindre son destinataire. Encore faut-il qu'elle puisse être trouvée facilement.
- Cohérence : dans le meilleur des cas, la cohérence des données reçues de l'extérieur peut être garantie par leur dimension institutionnelle ou quasi-institutionnelle.
- Traçabilité : elle permet de suivre l'information de sa collecte à sa restriction en passant par son traitement.
- Sécurisation : elle est l'une des dimensions de la qualité de l'information, même si ce n'est pas forcement celle qui vient en premier à l'esprit.

I.8Méthodes de sélection des dimensions

Puisque la problématique est d'abord liée à l'identification des dimensions qui sont jugées pertinentes à un projet de qualité des données, il a été nécessaire de prendre connaissance des techniques de sélection actuellement disponibles.

Avec les années, différentes approches ont été proposées par des auteurs. Ces approches ont des points communs puisqu'elles sont habituellement basées sur des auteurs d'importance pour le sujet, tels que *Deming, Redman, Wang et Ballou* [15], lesquels ont émis des hypothèses majeures sur le sujet. Voici un bref survol de ces approches.

I.8.1TDQM

Cette méthode, élaborée par Richard Y. Wang du MIT propose une approche qui a fait intervenir près de 400 utilisateurs de données. Elle est composée de deux

importantes par cette étude. Un autre aspect intéressant de cette méthode est dû au fait qu'elle prend racine à la suite d'un questionnaire remis aux utilisateurs. Ce qui semble être un bon départ pour représenter adéquatement le point de vue des utilisateurs. Toutefois, si les utilisateurs ont peu ou pas de connaissances en qualité des données, il pourrait en résulter un nombre très important d'attributs, lesquels nécessiteront beaucoup d'efforts pour en synthétiser la liste. Cette méthode nécessite en soi beaucoup de temps puisque plusieurs utilisateurs interviennent à différents moments dans le processus. Par contre, les résultats peuvent êtres intéressants puisqu'ils permettent de modifier la vision de l'équipe de la qualité des données pour qu'elle soit d'avantage centrée utilisateur. En espérant que cette vision soit celle requise pour le projet.

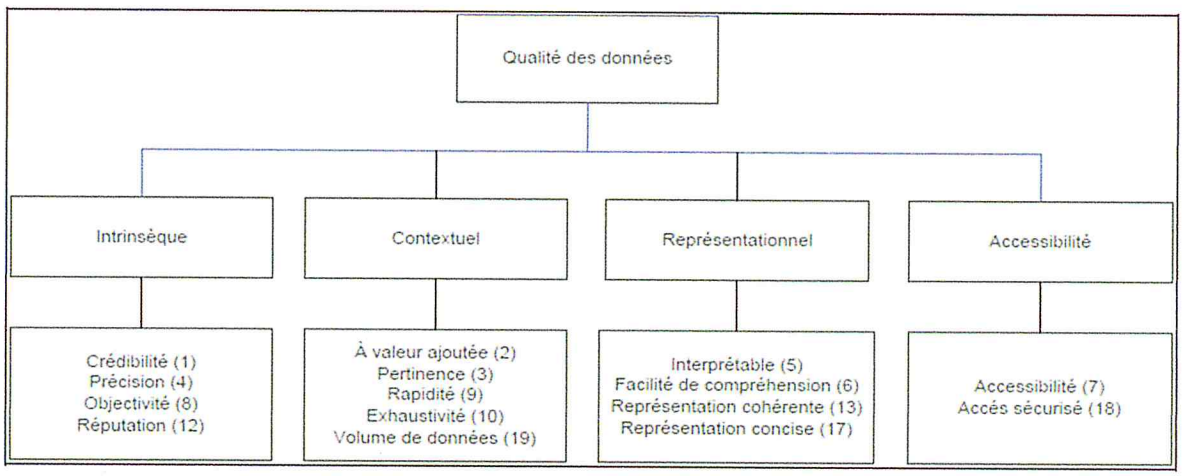

Figure I.2: Regroupement des dimensions selon TDQM [15]

I.8.2Danette McGilvray

Une approche publiée en 2008 utilise un court questionnaire basé sur 12 dimensions qui ont été préalablement sélectionnées par l'auteur. Pour chacune de ces dimensions, deux questions doivent être posées :

1) Dois-je évaluer ces données ?

2) Puis-je évaluer ces données ?

Si la réponse à ces deux questions est oui, la dimension peut-être retenue pour passer à l'étape suivante. La première question sous-entend qu'il est pertinent de mesurer la qualité d'une dimension seulement si celle-ci permet d'atteindre les besoins d'affaires. La deuxième question sous-entend que cette dimension peut être mesurée

avec la structure déjà en place. Elle sous-entend également que si le coût est trop élevé ou si les techniques requises sont non-disponibles, il n'est pas réaliste de penser pouvoir mesurer cette dimension. Une fois la sélection effectuée, il faut procéder à la mesure du niveau de qualité des données pour chacune de ces dimensions. Une fois les résultats connus, une deuxième étape est proposée. Cette étape sert à mesurer l'impact de ces dimensions sur l'entreprise. Les aspects qualitatifs et quantitatifs seront mesurés afin de connaître l'impact qu'a la qualité ou la non-qualité des données sur l'entreprise [14].
Cette méthode est intéressante du fait qu'elle est très simple et ne nécessite pas l'intervention de plusieurs personnes. Toutefois, elle se limite aux 12 dimensions choisies par *McGilvray.* Ce qui ne sera pas nécessairement représentatif pour toutes les entreprises et leur contexte particulier. Il peut constituer un bon point de départ afin de savoir si la mise en place d'un projet de qualité des données est pertinente et peut apporter de la valeur à l'entreprise. Les gestionnaires de données pourraient ajouter des dimensions qu'ils jugeraient appropriées et qui sont absentes de la liste fournie par *McGilvray.* Cette technique de sélection est en somme très rapide et nécessite peu de ressources humaines. La partie qui constitue la mesure du niveau de qualité des données est certainement la partie la plus complexe et la plus délicate de cette méthode.

I.8.3 ICIS

L'Institut canadien d'information sur la santé est considéré comme un pionnier dans les techniques de gestion de la qualité des données dans le milieu de la santé [15]. Il a mis en place, dans les années 2000, un cadre sur la qualité des données. Ce cadre comprend un outil d'évaluation qui permet de mesurer et de documenter les limites et les forces comprises dans les banques de données de l'ICIS. L'Institut canadien d'information sur la santé mentionne que :
« *Pour parvenir à obtenir de l'information de qualité supérieure, il faut déterminer les causes fondamentales des anomalies, prévenir les erreurs dans les processus d'information, établir les exigences en matière de qualité de l'information et contrôler les processus d'information.* » [19]
Cet outil d'évaluation utilise cinq dimensions pour lesquelles 19 caractéristiques et 61 critères ont été définis. Les dimensions utilisées sont l'exactitude, l'actualité, la comparabilité, la facilité d'utilisation et la pertinence. Pendant l'exercice, chacun des critères se voit assigner une cote. Les cotes (respecté, non respecté, inconnu ou non applicable) sont celles généralement utilisées. Ces cotes sont ensuite utilisées afin de

confectionner un plan d'action qui servira à corriger les lacunes. Ce qui est intéressant dans cette démarche, c'est que les cinq dimensions utilisées ont été hiérarchisées afin de donner des niveaux d'importance différents à chacune des dimensions. Ce document attribue à la dimension « pertinence », le plus haut niveau d'importance puisqu'il mentionne que si les autres dimensions sont respectées, mais que la pertinence n'est pas au rendez-vous, alors la donnée est futile et de peu d'importance. La figure I.3 illustre les cinq dimensions au sein du cycle de travail de la qualité des données, tel que représenté par l'ICIS.

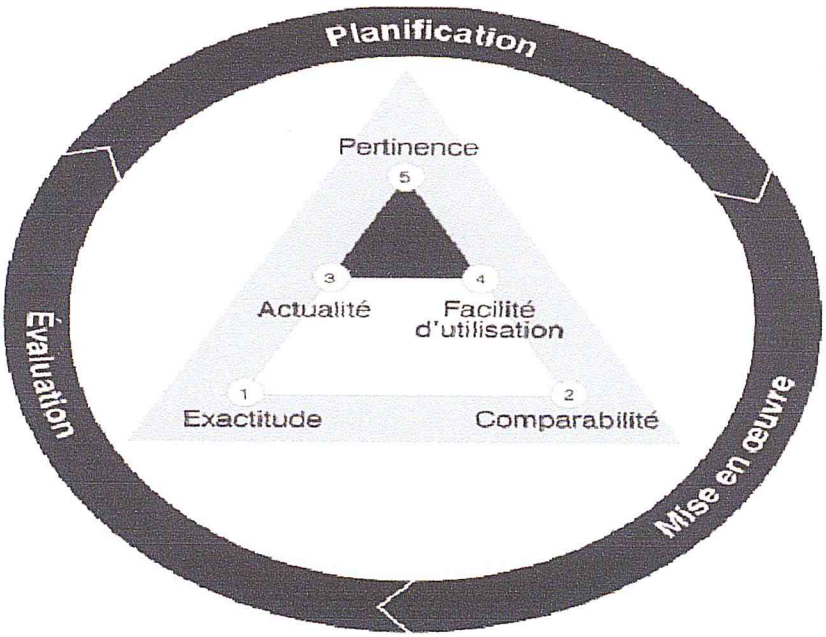

Figure I.3 : Cycle de travail de l'ICIS [19]

L'ensemble de la démarche proposée par l'ICIS a fortement inspiré le département de la santé de la Nouvelle-Zélande qui en a utilisé les fondements en y apportant quelques modifications afin de le conformer à leur contexte.

Cette approche est intéressante puisqu'elle est basée sur un petit nombre de dimensions. De plus, les dimensions font entre elles l'objet d'une certaine hiérarchisation. Ce qui permet de bien orienter les efforts en vue d'atteindre les objectifs. Comme cette technique est conçue pour s'appliquer à un cadre bien précis (la santé), il est difficile de critiquer le nombre de dimensions résultantes avec le choix restreint que ça implique. Ces dimensions étant destinées à un but bien précis. La hiérarchisation permet de mettre en place des pratiques et des techniques évolutives qui

nécessitent de bien maîtriser une dimension avant de passer à la suivante. Ceci peut faire en sorte d'obtenir une grande maîtrise pour chacune des dimensions et ainsi en retirer un apport maximal.

I.9Outils de gestion de la qualité de données

Les outils de gestions de la qualité de données visent à appliquer une solution idéale pour le problème de la qualité. Il y a une centaine d'outils fournis par de nombreuses entreprises et certaines communautés open-source. Des outils de nettoyage de données sont des outils les plus communs de la qualité de données et généralement utilisés dans le domaine de gestion qui visent à réparer et supprimer des données erronées.

De nombreux inventeurs de base de données fournissent non seulement des produits de base de données mais aussi de nettoyage des outils. Certaines sociétés Internet offrent des outils ETL (Extraction Tranform Load) qui a généralement la fonction de gestion de la qualité des données, comme Oracle Data Integrator et warsehouse build, Intégration de Microsoft Dynamics et IBM Data Integrator. Le nettoyage des données est une tâche très importante dans le processus ETL. La solution de rechange pour les questions de la qualité des données est de concevoir ces outils [20].

I.10 Conclusion

Dans ce chapitre, nous avons présenté les principales stratégies et techniques de la qualité des données ainsi que les méthodes, modèles, mesures utilisées.
Dans la suite de ce travail, nous accordons une attention particulière au nettoyage des données incomplètes afin d'assurer leurs qualités.

CHAPITRE II

TRAITEMENT DES DONNÉES INCOMPLÈTES

Chapitre II : Traitement des données incomplètes

II.1 Introduction

Les masses de données issues du monde réel, collectées via différents outils et dispositifs de collecte sont dans une grande partie, incomplètes. Certaines informations ne sont pas disponibles, d'autres non renseignées, ou aberrantes [21]. Ceci semble constituer un phénomène aussi imprévisible qu'inévitable. En effet, les raisons pour lesquelles certaines valeurs peuvent être manquantes sont multiples :

- Un instrument de mesure peut être en panne durant une période.
- Une mesure peut alors ne pas être réalisable lorsque le système entre dans un état particulier (en répondant "non" à une question d'un questionnaire, il arrive souvent que certaines questions ne soient pas pertinentes).
- Il se peut que l'opérateur ait oublié de reporter une donnée, ou que la donnée reportée ne soit pas lisible.

La présence des valeurs manquantes ou nulles dans les bases de données a suscité l'intérêt de plusieurs communautés scientifiques. Les données manquantes constituent un problème majeur, puisque l'information à disposition est incomplète par conséquent, moins fiable. Le traitement des données avec des observations manquantes est un problème concret et toujours embarrassant lorsqu'il s'agit de données réelles. En effet, on est très souvent en présence d'observations pour lesquelles on ne dispose pas de l'ensemble des valeurs des variables descriptives.

Les données manquantes peuvent avoir un impact sur la validité, la fiabilité et la généralisation des estimations effectuées. Le type et la quantité des données manquantes peuvent également avoir un impact différent sur chacun de ces aspects.

La proportion de données manquantes n'est pas suffisante pour déterminer si leur présence est problématique ou non. C'est plutôt la question de recherche des informations contenue dans les données observées et les raisons à l'origine des données manquantes qui déterminent leur impact [22]. Les raisons à l'origine des données manquantes constituent le volet sur lequel le chercheur a le moins de contrôle. Leur présence apporte une incertitude supplémentaire dans l'analyse de données, incertitude bien distincte de la notion d'erreur d'échantillonnage.

II.2 Définition et description des données manquantes

Une donnée manquante peut être définie comme une donnée qui était visée par le processus de collecte, mais qui n'a pas pu être obtenue. Différents éléments peuvent être à l'origine de données manquantes, ils peuvent être associés aux participants, au plan de l'étude ou à une interaction entre les deux. Une donnée peut être manquante au moment de la collecte ou du traitement des données. Toutes ces caractéristiques permettent de décrire les données manquantes pour chacune des analyses menées [23].

II.3 Types des données manquantes (Modèles d'apparition)

Le modèle d'apparition des valeurs manquantes définit la loi selon laquelle les valeurs manquantes apparaissent dans les données. Plus précisément, un modèle indique la probabilité qu'une valeur manquante dépende de sa propre valeur, d'une ou plusieurs autres valeurs ou qu'elle soit complètement indépendante des données. L'immense majorité des travaux sur les valeurs manquantes cite la classification basée sur trois modèles statistiques proposés dans [21]. L'étude menée par *Little et Rubin* propose de répartir les modèles d'apparition des valeurs manquantes en 3 types, "Missing Completely At Random" (MCAR) (complètement aléatoire), "Missing At Random" (MAR) (aléatoire), "Missing Not At Random" (MNAR) (non aléatoire).

II.3.1 MCAR

Une valeur manquante est dite complètement aléatoire lorsque le non réponse est totalement indépendant de toute autre valeur. Une valeur manquante complètement aléatoire affecte donc n'importe quel attribut et n'importe quel objet : la probabilité qu'elle soit manquante est la même pour toutes les valeurs. Par exemple, dans la table suivante où nous représentons deux attributs sexe et poids, la colonne MCAR indique le cas où les valeurs manquantes sur l'attribut poids sont de type MCAR. En effet, sur cette colonne, les valeurs manquantes affectent d'une manière équi-répartie les personnes de sexe féminin ou masculin. De plus, ces valeurs manquantes ne dépendent pas des valeurs réelles de l'attribut poids, qu'on ne possède pas en réalité.

II.3.2 MAR

Une valeur manquante est dite aléatoire lorsque l'absence d'une valeur dépend des valeurs réelles particulières d'autres attributs. Il est à noter que, dans ce cas, la désignation aléatoire prête à confusion, car elle concerne les valeurs réelles de l'attribut

présentant la valeur manquante. Si nous reprenons la table suivante, alors nous remarquons dans la colonne MAR que toute personne de sexe féminin ne présente une valeur manquante sur l'attribut poids. Ces valeurs manquantes dépendent donc du sexe de la personne : on peut, par exemple, déduire que seules les femmes ont tendance à cacher leurs poids, quelle que soit la valeur réelle du poids. On peut très bien remarquer, à partir de cet exemple, que les valeurs réelles du poids de ces femmes n'appartiennent pas à un intervalle de valeurs précises, d'où le terme « aléatoire ».

II.3.3 MNAR

Si une valeur est manquante lorsque la valeur réelle de l'attribut correspondant est particulière, alors le modèle est dit non-aléatoire, i.e., lorsque le phénomène de « non réponse » dépend de la valeur manquante elle-même. Les valeurs manquantes sur l'attribut poids (dernière colonne) illustrent un exemple de valeurs manquantes de type NMAR. En effet, nous remarquons que les personnes concernées par ces valeurs manquantes sont celles dont le poids est supérieur à 100, car les personnes ayant un surpoids ont tendance à le cacher.

sexe	poids			
	Valeur réelle	MCAR	MAR	NMAR
M	80	?	80	80
F	67	67	?	67
F	53	?	?	53
M	132	?	132	?
M	62	62	62	62
F	57	57	?	57
M	70	?	70	70
F	62	62	?	62
M	115	115	115	?
F	55	55	?	55
F	145	?	?	?
F	110	110	?	?

Tableau II.1 : Exemples de trois modèles d'apparition des valeurs manquantes [21]

En présence de valeurs manquantes, il est primordial de considérer leurs modèles d'apparition. Ceci permet de comprendre pourquoi les données sont manquantes, et facilite par la suite leur traitement. En considérant l'exemple des personnes présentant

un surpoids et qui ont tendance à le cacher, il s'avère intéressant de pouvoir caractériser le fait que toutes les personnes présentant un poids manquant sont des personnes concernées par un surpoids. Il est alors facile de classer toute personne ayant un poids manquant dans la catégorie des personnes obèses [21].

La principale conséquence des données manquantes *MCAR* est la perte de puissance statistique. Cette situation engendre néanmoins des estimations non biaisées. Les données manquantes *MAR* engendrent également des estimations non biaisées lorsque les analyses contrôlent pour le mécanisme des données manquantes. La troisième situation, *MNAR*, est la plus problématique parce qu'elle engendre des estimations biaisées [24].

II.4 Méthodes de traitement des données manquantes

Il existe une abondante littérature sur les méthodes de traitement des valeurs manquantes. On peut répartir l'ensemble de ces méthodes en des méthodes simples ou palliatives et des méthodes plus élaborées qui concernent plusieurs domaines tels que la statistique, l'apprentissage supervisé, les bases de données, la théorie des ensembles approximatifs ou la fouille de données. Les méthodes de traitement des données manquantes se distinguent selon deux approches, les méthodes supprimant les données manquantes et les méthodes utilisant toute l'information disponible.

II.4.1 Méthodes de traitement basées sur la suppression des données manquantes

Dans la première catégorie, on trouve les techniques connues sous l'appellation analyse des cas complets (listwise deletion) et analyse des cas complets par paires (pairwise deletion).

II.4.1.1 Étude des cas complets (Listwise deletion)

Cette méthode permet de se ramener à une base de données complète par réduction de la dimension du problème. Pour cela, tous les exemples de la base contenant des valeurs manquantes sont supprimés (On peut également choisir de supprimer toutes les variables dont certaines observations manquent, mais il faut être prudent car certaines peuvent être essentielles pour l'analyse). Par conséquent, cette méthode sacrifie un grand nombre de données [25]. Selon [25], la suppression de 10 % des données de chaque variable dans une matrice de cinq variables peut facilement provoquer l'élimination de 59 % des observations de l'analyse.

ABDERRAZAK et al. [25], rapporte qu'il a vu une baisse de dimension de l'échantillon de 624 à 201 avec l'utilisation de la méthode de suppression *listwise*.

Les techniques statistiques d'analyse des données ayant besoin d'un nombre suffisant d'observations pour que leurs résultats soient valides. Dans des cas qui ne sont pas rares, où la quasi-totalité des exemples possède des valeurs manquantes, elle devient même inutilisable.

D'autre part, les statistiques, telles que la moyenne ou la variance, seront fortement biaisées, à moins que le mécanisme de génération des données ne soit complètement aléatoire (MCAR). Malgré le fait que la grande perte de données réduit la puissance et l'exactitude statistiques, cette technique, du fait de sa simplicité est fréquemment l'option implicite pour l'analyse dans la plupart des progiciels statistiques. Cette méthode est aussi appelée par certains auteurs, la méthode d'analyse des données disponibles (*available-case analysis*) [25].

II.4.1.2 Étude des cas disponibles (Pairwise deletion)

Dans cette méthode, on ne considère que les cas où ces variables sont complètement observées. Par exemple, si la valeur de l'attribut A est absente pour une observation, les autres valeurs pour le reste des attributs de la même observation pourraient encore être employées pour calculer des corrélations, telles que celle entre les attributs B et C. Comparée à la première méthode (Étude des cas complets), cette méthode conserve beaucoup plus de données qui auraient été perdues si on employait la méthode d'étude des cas complets [25].

Il s'agit d'une autre méthode proposée par les logiciels statistiques, mais généralement problématique : le nombre d'observations (n) varie pour le calcul de chaque valeur de la nouvelle base de données, le risque d'obtenir une base de données réduite est grand et, encore une fois, la représentativité sera biaisée si les données manquantes ne sont pas distribuées de façon complètement aléatoire.

Les études de Monte Carlo ont montré que la suppression par la méthode **Listwise deletion** donne des évaluations moins précises des paramètres d'estimation. La méthode **Pairwise deletion** est uniformément plus précise, bien que les différences puissent parfois être minimes [25].

Toutefois, il existe certaines raisons, pour la considérer une bonne méthode CD doit être appliquée uniquement dans les cas où le nombre de valeurs manquantes est relativement faible.

De plus, il paraît que, même si les données sont manquantes selon un mécanisme complètement aléatoire, les méthodes qui utilisent l'ensemble de l'information contenue dans la matrice de données sont plus efficaces que les méthodes basées sur les données complètes [25].

II.4.2 Méthodes de traitement utilisant toute l'information disponible

Dans la deuxième approche qui utilise toute l'information, beaucoup de travaux ont été proposés, consistant à chercher, pour chaque valeur manquante, une valeur de remplacement.

On parle d'imputation des valeurs manquantes [21]. L'imputation des données manquantes n'est pas nouvelle et il existe à ce jour une foule de techniques développées à cet effet.

L'imputation consiste à produire une « valeur artificielle » pour remplacer la valeur manquante, dont l'objectif est de produire des estimations approximativement sans biais. Le grand avantage de l'imputation est qu'elle permet de créer des bases de données complètes. Cependant, cela n'est pas sans conséquence dans le calcul de certains estimateurs, en particulier de la variance.

Chaque technique d'imputation conduit à une formule de variance ainsi qu'à une estimation de variance particulière [26].

Les principales méthodes d'imputation sont les suivantes:

1. **Méthodes de déduction.** Une valeur manquante est déduite à l'aide d'une règle logique construite à partir des réponses aux autres questions.
2. **Méthodes de substitution.** Une valeur manquante est remplacée par la valeur correspondante d'une observation similaire. Ce répondant similaire peut être choisi dans l'enquête courante (méthodes de hot deck qui donnent pour la valeur manquante, la valeur observée d'un individu répondant, le donneur, choisi selon une procédure adéquate, soit aléatoire, soit séquentiel, soit métrique) ou dans d'autres sources (méthodes de cold deck qui utilisent l'information obtenue à partir des répondants d'une autre enquête).
3. **Méthodes de prédiction.** L'ensemble des données est utilisé pour construire un modèle qui prédit la valeur de l'item manquant à partir des valeurs des items

corrélés (parmi les méthodes classiques de prédiction par imputation, on peut citer l'imputation par la moyenne, la régression…) [27].

Différentes procédures de remplacement des données manquantes ont été élaborées au cours des années. Généralement on constate que les différences entre les diverses méthodes diminuent avec:

(a) une plus grande dimension de la base de données,

(b) un plus petit pourcentage des valeurs manquantes et

(c) une diminution au niveau des corrélations entre les attributs [25]

Cependant, [25] a rapporté qu'il y a une différence entre les méthodes de remplacement des données manquantes si les effets des traitements sur les statistiques analytiques sont pris en considération. Avec de plus grandes dimensions de la base de données, en fait, les différences entre les diverses procédures de remplacement augmentent ; ceci fournit d'avantage d'évidence qu'en évaluant l'efficacité des traitements des données manquantes, l'exactitude d'estimer la valeur des données manquantes et l'exactitude d'estimer les effets statistiques doivent être considérées [25]. De façon générale, les stratégies de traitement des données manquantes par imputation sont fort utiles dans la pratique, car une fois les valeurs manquantes imputées, on peut avoir recours à des logiciels de données complètes existants pour analyser les données.

L'imputation offre plusieurs avantages

Elle facilite l'analyse des données car elle mène à la création d'un fichier complet à partir duquel on peut effectuer des analyses de données, des régressions... : elle assure donc la cohérence des résultats issus de différentes analyses. Par exemple, si on impute les yi manquants par $\hat{y}$k, on obtient un ensemble de données complétées :

$$y_k^* = \begin{cases} yk\ k \in r\ (ensembles\ des\ repondants\ ) \\ \hat{y}k\ k\ \in S - r(\ non - repondants\ ) \end{cases} \quad (1)$$

L'imputation évite également la perte d'informations car elle utilise toutes les données observées sur les répondants partiels. Si les méthodes d'imputation ont toutes pour objectif d'améliorer la qualité des données finales, elles ne résolvent cependant pas tous les problèmes liés à la présence de non réponse partielle.

Les limites des méthodes d'imputation sont aussi multiples.

Les méthodes d'imputation offrent de nombreux avantages mais il ne faut pas oublier lors de l'exploitation que la présence de données imputées a des conséquences sur la qualité des inférences, surtout si beaucoup de réponses ont été imputées : aucune méthode d'imputation ne remplacera une vraie réponse.
Les méthodes d'imputation sont soit simples ou multiple [26].

II.4.2.1 Méthode d'Imputation simple

L'imputation simple consiste à remplacer chaque donnée manquante par une estimation (et une seule) de sa valeur et à analyser la base de données ainsi complétée. Les méthodes d'imputation simple sont généralement celles générant les moins bons résultats.
D'un point de vue statistique, cette procédure de remplacement peut être stochastique ou déterministe, selon qu'elle implique ou non le tirage d'un nombre aléatoire.

1. **Méthodes déterministes**

On distingue le groupe de méthodes suivant :

- les méthodes déductives : la donnée manquante est déduite des réponses aux autres questions, Ce type d'imputation par règle déterministe est surtout utilisé dans les enquêtes d'entreprises pour corriger des données intervenant des équations comptables.
- les méthodes de type "cold-deck" : elles utilisent l'information obtenue à partir des répondants d'une autre enquête, par exemple, la donnée manquante peut être remplacée par une valeur obtenue en dehors de l'enquête (enquête précédente dans le cas d'un panel, recensement ou autres sources).
- les méthodes utilisant la prévision par un modèle de régression ;

2. **Méthodes stochastiques**

- les méthodes de type "hot-deck" : elles donnent pour la valeur manquante la valeur observée d'un individu répondant choisi selon une procédure adéquate. Il existe différentes procédures connues, et régulièrement utilisées en pratique, de choix des données ; citons par exemple le hot-deck aléatoire, séquentiel hiérarchisé, métrique, etc.

3. **Imputation par substitution**

L'élément manquant est remplacé par l'utilisation d'une unité initialement non présente dans l'échantillon, mais qui est similaire à celle manquante (par exemple non-répondants dans les enquêtes sur la population).

Il est clair que pour les données rassemblées avec cette méthode, il faut les traiter comme des données imputées [28].

4. Imputation par la moyenne (node) ou médiane (I.Mean)

Les valeurs manquantes de chaque attribut sont remplacées par la moyenne de l'attribut considéré. Il y a deux variantes de l'imputation par la moyenne : Imputation par la moyenne totale, imputation par la moyenne de sous-groupe. Pour l'imputation par la moyenne totale, la valeur absente d'un attribut est remplacée par la moyenne des valeurs de cet attribut de toutes les observations. Pour l'imputation par la moyenne de sous-groupe (classe), la valeur manquante est remplacée par la moyenne du sous-groupe (classe) de l'attribut en question.

L'inconvénient de cette méthode est la sous-estimation de la variance et de biaiser la corrélation entre les attributs, cela veut dire que la distribution des données est loin d'être préservée.

Les études ont été quelque peu concluantes concernant l'efficacité de la substitution par la moyenne. Selon [25], la substitution par la moyenne est moins précise que la méthode **Listwise deletion**, alors que d'autres, ont prouvé que la substitution par la moyenne est plus précise que le **Listwise deletion** et le **Pairwise deletion** [25].

5. Imputation par la méthode du plus proche voisin (I.KNN)

La technique de k-plus proche voisin (kNNI - k-nearest-neighbors imputation), est une technique utilisée pour la substitution des valeurs manquantes, avec la valeur du plus proche voisin dans l'ensemble de données [25].

Pour chaque observation contenant des valeurs manquantes, on recherche ses k plus proches voisines. Dans le cas de variables continues, la valeur de remplacement correspond simplement à une moyenne pondérée des valeurs prises par ces k voisins pour la variable en question.

Chaque valeur manquante est remplacée par la moyenne arithmétique des plus proches voisins de la valeur manquante dans la même classe (S + ou S-).

Les trois cas possibles sont les suivants:

- Si la valeur est située entre deux valeurs existantes, la valeur manquante est remplacée par la moyenne de ces deux valeurs.

- Si la valeur est située au début d'une classe, alors elle sera remplacée par la plus proche valeur inférieure.
- Si la valeur est située à l'extrémité d'une classe, alors elle sera remplacée par la plus proche valeur supérieure.

L'avantage de cette méthode est de ne faire aucune supposition quant à la distribution des données, et de prendre en considération la corrélation entre variables [25].

6. Imputation par l'algorithme EM (Espérance-Maximisation)

Initialement développé par A. P. Dempster, N.M. Laird et D. Rubin en 1977, l'algorithme EM est un algorithme itératif de calcul d'estimateur de vraisemblance par des modèles paramétriques lorsque les données sont observées. Dans le cadre du traitement des données manquantes, Il permet de compléter les valeurs manquantes en se basant sur la vraisemblance maximale (maximum-likelihood estimation) de l'ensemble des données disponibles.

L'algorithme EM est une succession de deux étapes : une étape (E) où on évalue l'espérance de la log-vraisemblance pour la valeur courante du paramètre puis une autre étape (M) où on actualise le paramètre en maximisant cette nouvelle fonction du paramètre. L'estimation ainsi obtenue est celle qui maximise la probabilité d'observer ce qui a été réellement observé. L'algorithme converge vers un point stationnaire sous des hypothèses de régularité [29].

Il est généralement utilisé pour estimer les paramètres d'une densité de probabilité. Il peut être appliqué sur des bases de données incomplètes, et présente l'avantage de procéder à l'estimation des valeurs manquantes en parallèle de l'estimation des paramètres.

Cette technique est très coûteuse en temps de calcul comme beaucoup d'approches itératives [25]. De plus elle demande la spécification d'un modèle de génération des données.

Cette tâche implique de faire un certain nombre d'hypothèses, ce qui est toujours délicat. Pour ces raisons, l'application d'EM pour remplacer les données manquantes n'est pas toujours envisageable.

III.4.2.2 Méthode d'imputation multiple

La méthode d'imputation multiple a été originellement proposée en 1978 par Rubin dans le domaine des sciences sociales [30]. Par la suite, la mise en application de la méthode a été développée par Rubin, dans le contexte de bases de données importantes issues d'enquêtes complexes et destinées à être exploitées par de nombreux utilisateurs et pour des analyses variées [30]. L'imputation multiple est alors appliquée à l'ensemble de la base de données incomplète, sans tenir compte des analyses ultérieures.

A l'époque, les développements de l'imputation multiple sont encore obscurs et son utilisation est limitée à des experts [30]. Les équipes assurant les étapes d'imputation multiple et d'analyse de ces bases de données sont donc des entités distinctes. L'objectif est alors que les équipes d'utilisateurs puissent effectuer toutes les analyses prévues en appliquant les commandes usuelles.

Au cours du temps, la mise en œuvre de l'imputation multiple est devenue plus accessible du fait de l'apparition d'ordinateurs performants et de programmes dédiés [30].

Le principe général de l'imputation multiple consiste à remplacer chaque valeur manquante par un ensemble de M valeurs plausibles, de façon à prendre en compte l'incertitude liée au processus d'estimation des valeurs manquantes [30].

L'imputation multiple permet de prendre en compte l'incertitude de prédiction des valeurs manquantes de l'imputation simple. L'idée de base, élaborée par Rubin [30], est la suivante : (a) imputer les valeurs manquantes en utilisant un modèle approprié qui incorpore la variation aléatoire, (b) répéter cette opération m fois (en général trois à cinq fois) afin d'obtenir m fichiers de données complets. Les analyses statistiques sont alors effectuées sur chaque fichier complété et les résultats sont combinés pour obtenir le modèle final.

Les buts sont : refléter correctement l'incertitude des NA, préserver les aspects importants des distributions, préserver les relations importantes entre les variables, et non pour de prédire les données manquantes avec la plus grande précision et décrire les données de la meilleure façon possible [30].

L'imputation multiple se décompose en trois phases :

- ***Phase d'imputation***

Les données manquantes sont estimées M fois à partir d'un modèle spécifique pour obtenir M bases de données complètes et potentiellement différentes.

- ***Phase d'analyse séparée***

L'analyse retenue est réalisée séparément sur chacune des $m = 1,..., M$ bases de données imputées pour obtenir M estimations (valeur centrale et variance).

- ***Phase d'analyse combinée***

Les résultats obtenus à partir des M analyses sont combinés selon des règles établies par Rubin pour obtenir une seule estimation finale.

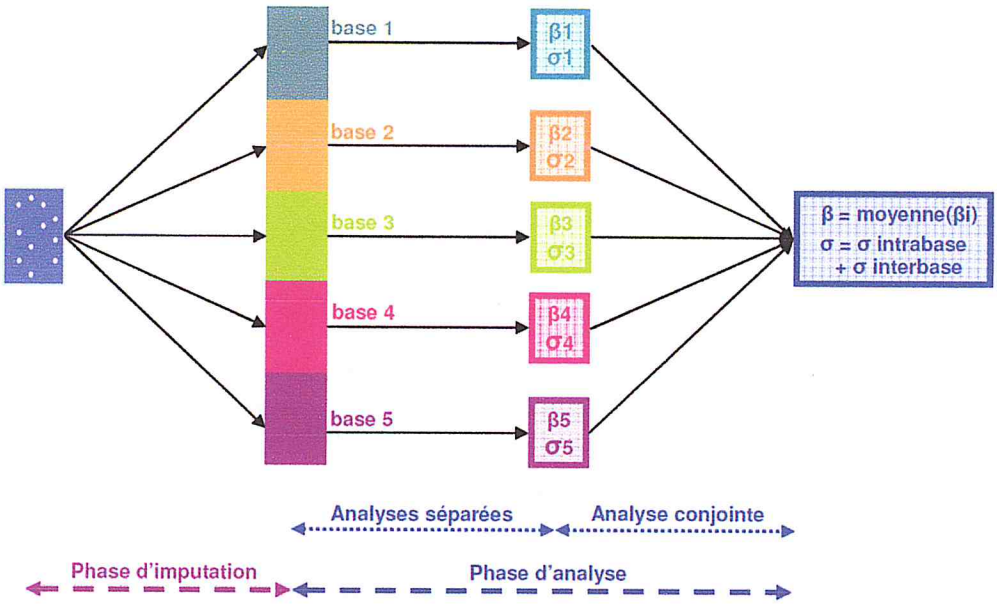

Figure II.1 : Représentation graphique des étapes de l'imputation multiple *[25]*

Alors que l'étape d'analyse de multiples bases de données est simple pour la plupart des estimations, la phase d'imputation constitue une étape complexe, dont dépend la validité des estimations finales.

1. Imputation par la régression (I. Reg)

C'est une approche en deux étapes : d'abord, on estime les rapports entre les attributs, et puis on emploie les coefficients de régression pour estimer la valeur manquante [25]. La condition fondamentale de l'utilisation de l'imputation par régression est l'existence d'une corrélation linéaire entre les attributs. La technique suppose également que les valeurs sont manquantes au hasard.

Dans le contexte des valeurs manquantes, deux modèles de régression sont en général employés : la régression linéaire et la régression logistique. Cette dernière est

plutôt utilisée pour traiter les variables discrètes, alors que la régression linéaire est appliquée sur des variables continues [25].

Pour chacune de ces méthodes, il est possible de tenir compte de l'information de classe en n'utilisant que les observations d'une même classe pour estimer les paramètres de régression.

L'inconvénient de cette méthode, c'est les hypothèses qui sont faites sur la distribution des données. Supposer une relation linéaire entre les variables, revient à faire des hypothèses qui sont rarement vérifiées, dans cette situation, le remplacement des valeurs manquantes par des valeurs prédites basées sur un modèle biaisé ne constitue pas un traitement approprié.

Ces méthodes, seraient beaucoup plus efficaces, exclusivement dans le cas où le modèle de régression est adéquat [25].

Un grand nombre de méthodes de traitement des valeurs manquantes sont citées ci-dessous. Le tableau suivant présente un résumé de ces différentes méthodes.

Technique	**Description**	**Champ d'application**	**Avantage**	**Inconvénient**	**Reference**
Procédures de suppression					
Etude des cas complets (listwise deletion)	Supprime toutes les observations dont certaines valeurs sont manquantes	Il convient d'éviter	Facile à utiliser (par défaut dans la plupart des logiciels statistiques)	Sacrifie une grande quantité de données et un impact négatif sur les paramètres d'estimations (corrélation – régression)	Kim and curry(1977),Raymond(1986), Malbotra(1987) ,Little and Rubin (2002)
Etude des cas disponibles (pairwise deletion	Crée une matrice de corrélation avec les valeurs disponibles (chaque couple de variables est pris deux à deux)	Lorsque les données sont relativement faibles (moins de 10%)	Préserve davantage les données et est plus précise que la	Corrélations ou covariances biaisées	Gleason and Staelin(1975), Kim and curry(1977), Raymond(1986),Roth (1994)

			suppression listwise)		
Procedures de replacement					
Imputation par la moyenne totale (Total mean substitution)	Remplacer par la moyenne des valeurs disponibles de la variable, toutes les valeurs manquantes pour la même variable)	Lorsque les corrélations entre les variables sont faibles(r<|20|) et le taux de leurs manquantes moins que 10%)	Préserve la taille de la base de données et la rend facile à utiliser). La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservés)		Ford(1976), Raymond(1986),Little and Rubin (1987),Kaufma n (1988), Quinten an Raaijmakers (1999)
Imputation par la moyenne de chaque classe (Subgroup mean substitution)	Remplacer par la moyenne des valeurs disponibles de la variable de la même classe, toutes les valeurs manquantes pour la même variable et dans la même classe.	Quand il est facile de définir les classes (classification supervisée)	Donne des meilleurs résultats, par rapport à l'imputation par la moyenne totale)	La sous-estimation de la variance et de biaiser la corrélation entre les variables (la distribution des données est loin d'être préservés)	Ford (1976)
Procédure de Modélisation					
Maximum Vraisemblance (Maximum likelihood)	Les paramètres sont estimés par les Données disponibles et les valeurs manquantes sont estimées en fonction des paramètres	Lors que les données Observés sont tirés d'une distribution normale mutlivariée	Augmentatio n de la précision si Le modèle est correct	Les hypothèses de la distribution Exigé par la technique sont relativement strictes.	Desarbo et al.(1990),Lee and Chiu (1990)

Maximum d'experance (Expectaton Maximization)	An iterative process that continues until there is convergence in the parameters estimates	Lorsque les hypothèses de répartition sont remplies	Augmentation de la précision si le modèle est correct	L'algorithme prend du temps à converger et, est trop complexe	Laird(1988), Little and Rubin(2002), Malbotra (1987),Ruud(1991)

Tableau II.2 : les différentes méthodes d'imputation [25]

II.4.2.3 Données incomplètes et bases de données

Dans le domaine des bases de données, l'expression « valeur nulle » est plus couramment utilisée que celle de valeur manquante. Dans [21], l'auteur propose de remplacer les valeurs nulles sur chaque colonne par « valeur spéciale » (ou default-value). De cette façon, ces valeurs spéciales seront traitées comme de nouvelles modalités. Dans [21], le fondateur de l'algèbre relationnelle adopte une vision plus pragmatique et indique qu'il existe deux sortes de valeurs manquantes : celle manquante mais applicable et celle manquante mais qui cache en réalité une valeur inapplicable. Sans chercher à trouver les valeurs réelles, il propose deux nouvelles valeurs de remplacement : « A-mark » (Missing-but-applicable) et « I – mark » (inapplicable).

Ces deux nouvelles valeurs seront par la suite employées pour remplacer les valeurs nulles.D'autres travaux ont étudié la problématique des valeurs manquantes dans le cadre des dépendances fonctionnelles, citons à titre d'exemple [21].

II.4.2.4 Les méthodes supervisées

Les méthodes supervisées sont utilisées pour prédire la classe d'appartenance d'un objet.

L'idée repose sur l'utilisation d'exemples déjà classés pour apprendre un modèle puis de l'utiliser afin de déterminer la classe de tout nouvel exemple. L'étape d'apprentissage, qui consiste à construire le modèle des exemples supervisés; l'étape de test, de classement ou de décision, qui valorise ce modèle pour classer les exemples, dont la classe est inconnue. Pour exhiber un modèle, l'objectif consiste à trouver des relations dans les données permettant d'expliquer leur appartenance à une classe spécifique. Le modèle utilisé pour la prédiction peut se présenter sous différentes formes : règles d'association, arbres de décision, réseaux de neurones, etc. Dans ce qui suit, nous

présentons les travaux qui ont abordé la problématique des valeurs manquantes dans le cadre de l'apprentissage supervisé [21].

II.4.2.4.1 Réseaux de neurones

Les réseaux de neurones artificiels pour imputer des valeurs appartiennent à la troisième catégorie. Dans ce cas, une valeur manquante est remplacée par la sortie calculée par le réseau. Les autres valeurs corrélées décrivant l'observation sont fournies en entrée du réseau. Dans un premier temps, on apprend au réseau à reconnaître la variable de sortie. Le réseau apprend donc les liens implicites reliant l'ensemble des variables d'entrée à cette variable de sortie. Puis on effectue un test pour vérifier l'apprentissage en définissant au hasard un échantillon de la population des sujets. On vérifie ainsi que le réseau a bien codé les liens et qu'il est capable d'effectuer la reconnaissance. On en déduit alors que cette variable de sortie est bien liée aux autres variables d'entrée, qu'on peut retrouver sa valeur en cas d'absence en utilisant le codage interne au réseau.

- **Procédure d'imputation par réseau de neurones**

En pratique, la méthode d'imputation des données manquantes par réseaux de neurones consiste à réaliser les étapes suivantes [27]:

1. Vérifier l'adéquation de la méthode aux besoins. Cette méthode n'est applicable que dans le cas où l'on dispose de peu de sujets présentant des non-réponses à certains items alors que l'on veut disposer de données complètes avec un fort taux de certitude.
2. On sépare un échantillon présentant des sujets avec des données complètes.
3. On identifie l'item présentant des données manquantes et on le qualifie comme variables de sortie.
4. On construit un réseau de type perceptron auquel on apprend par apprentissage supervisé à reconnaître les valeurs de la variable de sortie.

Cet apprentissage nécessite la plupart du temps l'utilisation d'une retro propagation des erreurs. Il convient de faire attention au sur apprentissage.

Dans certains cas en effet, les taux chutent brutalement au cours d'une opération de retro propagation des erreurs. Il faut alors retrouver la configuration du réseau précédent qui devient alors le modèle optimal. On le sauvegarde alors soigneusement. On calcule alors le taux de réussite en apprentissage. En entrée de ce même réseau, on présente ensuite les réponses aux items d'un sujet présentant une non-réponse sur

l'item sélectionné (cf. le point 3). Le réseau prédit alors la réponse avec le taux de réussite calculé au point 4.

II.4.2.4.2 Clustering sur des données incomplètes

Selon [31], l'analyse des données incomplètes peut se faire grâce au clustering fou. Cette méthode nécessite toutefois de traiter les données incomplètes différemment selon l'origine des valeurs manquantes.

La première étape consiste donc à analyser pour un ensemble de données les raisons de la présence de valeurs manquantes. Dans un deuxième temps, on recherche les corrélations dans la base. Comme pour le clustering, l'objectif du clustering fou est de diviser un ensemble de données en un ensemble de clusters tels que la similarité intra-classes est nettement supérieure à la similarité interclasses.

Cependant, le but est de pouvoir traiter les données qui pourraient appartenir à plusieurs groupes en même temps. On introduit un degré d'appartenance aux différents clusters, calculé en fonction de la distance entre cette donnée et le cluster. Chaque cluster peut donc être considéré comme un sous ensemble fou.

Cette méthode de traitement des données incomplètes par clustering fou permet de compléter des valeurs manquantes à chaque itération de l'algorithme de clustering, sur le même principe que la complétion proposée par l'algorithme Expectation-Maximization [31]. De plus, pour tenir compte de la différence entre données complètes et incomplètes, on réduit le degré d'appartenance des données incomplètes.

II.4.2.4.3 Règles d'association, valeurs manquantes et complétion

Dans le cadre des techniques de description, des travaux ont été proposés pour la recherche de règles d'association. [31] présente un algorithme afin de prendre en compte les données incomplètes lors de l'extraction des règles, par omission partielle et temporaire de ces enregistrements. Ces règles peuvent ensuite être utilisées afin de compléter les valeurs manquantes. [31] met en œuvre un système d'approximation probabiliste dans lequel une valeur manquante peut prendre plusieurs valeurs lors de la découverte des règles. Ces méthodes approximatives permettent d'extraire des règles proches de celles qui devraient être obtenues sur la base complète. En fin d'autres méthodes utilisent les règles d'association et certains indices de confiance afin de compléter les valeurs manquantes [31].

II.4.2.4.4 Réseaux bayésiens

De nombreuses méthodes d'apprentissage de structure de réseaux bayésiens ont vu le jour ces dernières années. Alors qu'il est possible de faire de l'apprentissage de paramètres de réseaux bayésiens à partir de données incomplètes et que l'inférence dans les réseaux bayésiens est possible même lorsque peu d'attributs sont observés [32], les algorithmes d'apprentissage de structure avec des données incomplètes restent rares.

Lorsque les données sont incomplètes, il est possible de déterminer les paramètres et la structure du réseau bayésien à partir des entrées complètes de la base. Comme les données manquantes sont supposées l'être aléatoirement, nous construisons ainsi un estimateur sans biais. Néanmoins, dans l'exemple d'une base de 2000 cas sur 20 attributs, avec une probabilité de 20% qu'une mesure soit manquante, nous ne disposerons en moyenne que de 23 cas complets. Les autres données à notre disposition ne sont donc pas négligeables et il serait donc préférable de faire l'apprentissage en utilisant toute l'information à laquelle nous avons accès.

Un avantage des réseaux bayésiens est qu'il suffit que seules les variables Xi et Pa(Xi) soient observés pour estimer la table de probabilité conditionnelle correspondante.

La recherche de structure de réseaux bayésiens peut utiliser des bases de données incomplètes, par exemple par le biais d'un échantillonnage de Gibbs ou encore en utilisant une approche comme l'algorithme EM [32]. D'autres travaux utilisent des techniques plus originales, comme [33] qui effectuent l'apprentissage de sous-structures locales ou encore [34] qui utilise une méthode à base de recherche d'indépendances conditionnelles.

II.5 Comparaison

Dans cette partie, nous allons ressortir un petit tableau comparatif de quelques méthodes supervisés en se basant sur les points suivants : l'acquisition, représentation et utilisation des données.

Connaissances	**Arbre de decision**	**Réseaux de neuronnes**	**Réseaux bayésiens**
Acquisition			
Mixte	Avantageux	Plus avantageux	Avantageux
Données seulement	Avantageux	avantageux	Plus avantageux
Incrémental (tps réel)		avantageux	Plus avantageux
Généralisation	Avantageux	Plus avantageux	Avantageux
Données incomplètes		avantageux	Plus avantageux
Représentation			
Incertitude			Plus avantageux
Lisibilité	Avantageux		Plus avantageux
Facilité	Plus avantageux	avantageux	
Homogénéité			Plus avantageux
Utilisation			
Requête élaborée			Plus avantageux
Utilité économique		avantageux	Plus avantageux
Performance		Plus avantageux	

Tableau II.3 : comparaison de quelques méthodes supervisés de traitement des *données*

II.6 Choix de la Méthode

Avant de réfléchir sur la méthode de traitement à appliquer aux données manquantes, un chercheur doit tout mettre en place pour ne pas avoir des valeurs manquantes [25], puisqu'il n'existe pas de méthode totalement efficace pour traiter le problème des données manquantes. Comme l'affirme [25], la meilleure méthode de traitement c'est de ne pas avoir de valeurs manquantes. Les méthodes permettent, au mieux, de réduire les biais induits par la présence de ces données manquantes.

Dans le chapitre suivant, l'accent sera mis sur les méthodes de réseaux bayésiens que nous utilisons dans notre étude, nous ne prétendons pas couvrir l'ensemble des méthodes, mais nous évoquons les plus courantes, celles que nous avons incluses dans notre travail.

Les principaux objectifs que l'on peut vouloir poursuivre sont les suivants :

- **Précision de l'étape d'analyse** : dans notre contexte, la phase d'analyse correspond à la construction d'un modèle de classification supervisée. Un des objectifs est alors de maximiser la performance du classificateur.

- **Temps réel :** d'apprendre la structure de façon incrémentale, c'est-à-dire apprendre la structure au fur à mesure de l'acquisition de données.

- **Précision de la substitution** : la valeur de remplacement doit être aussi proche que possible de la vraie valeur.

Le réseau bayésien a été choisi à causes de ces caractéristiques offertes qui sont très important dans notre contexte.

II.7 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art des principales méthodes dédiées au traitement des valeurs manquantes. Nous nous sommes d'abord intéressés aux modèles d'imputation simple et d'imputation multiple. Ensuite, nous avons présenté des travaux faisant partie du domaine de bases de données, des ensembles approximatifs ainsi que de l'apprentissage supervisé.

Cette étude bibliographique montre l'importance de cette problématique d'une part et la richesse des contributions d'autre part. Les méthodes auxquels nous nous sommes intéressés de près sont ceux des réseaux de neurones, clustering, réseaux bayésiens etc…

Après avoir présenté ces différentes méthodes, on a présenté un petit tableau comparatif et la méthodologie de choix de notre méthode.

Dans la suite de ce travail, nous nous focaliserons sur les méthodes des réseaux bayésiens pour le traitement des données manquantes.

CHAPITRE III

LES RÉSEAUX BAYÉSIENS

Chapitre III : Les Réseaux bayésiens

III.1 Introduction

Les réseaux bayésiens sont nés dans les années 80 d'un besoin de gérer efficacement les incertitudes dans les systèmes experts à base de règles.

Mais rapidement, ceux-ci sont devenus l'un des outils les plus populaires de la communauté Intelligence Artificielle pour gérer les incertitudes. Leur succès est certainement dû au fait qu'ils allient une représentation compacte des probabilités et des mécanismes de calcul efficaces. En particulier, ils permettent d'inférer l'impact d'une nouvelle information sur une loi de probabilité (calcul de probabilités a posteriori) [35].

Les Réseaux Bayésiens sont des modèles graphiques qui représentent les relations probabilisées entre un ensemble des variables [36]. Un modèle graphique est une famille de distribution de probabilités définie en termes de graphe orienté ou non. Il est constitué de nœuds qui représentent des variables aléatoires et des arcs représentant les relations de dépendances entre ces variables.

On distingue deux formes de modèles graphiques probabilistes : les modèles orientés et les modèles non orientés, basés respectivement sur les graphes acycliques orientés connus sous le nom de réseaux bayésiens (RB) et les graphes non orientés, exemple les champs de Markov (HMM) pour ces modèles graphiques probabilistes, non seulement ils bénéficient des avantages des modèles probabilistes, mais aussi ils représentent les avantages liés à leur représentation graphique. En effet ils permettent de visualiser la structure et les propriétés de dépendance conditionnelles du modèle probabiliste correspondant. Ainsi, les RB sont une union entre la théorie des probabilités et la théorie des graphes. Ils constituent une technique d'acquisition, de représentation et de manipulation de connaissance et on les utilise, surtout, pour leur capacité d'effectuer des inférences dans un contexte d'incertitude et aussi pour leurs algorithmes d'apprentissages [37].

III.2 Définition d'un Réseau Bayésien

Les RB sont des modèles qui permettent de représenter des situations de raisonnement probabilistes basé sur le théorème de Bayes exprimé par la formule suivante, et ce à partir de connaissances incertaines.

$$P(A/B) = P(B) \times P(B/A)/P(A) \quad (1)$$

Ainsi, les RB associent une partie qualitative que sont les graphes et une partie quantitative représentant les probabilités conditionnelles associées à chaque nœud du graphe relativement au parent [37]. La partie qualitative exprime des indépendances conditionnelles entre variables et des liens de causalités et ce grâce à un graphe orienté acyclique dont les nœuds correspondent à des variables aléatoires. La partie quantitative est constituée de tables de probabilités dans le cas discret ou de distribution gaussiennes dans le cas continu. Un réseau bayésien $B = \{G, P\}$ est donc défini par un graphe dirigé, un espace probabiliste et un ensemble de variables aléatoires. Le graphe est sans circuit $G = (X, E)$ où X est l'ensemble des nœuds (ou sommets) et E, l'ensemble des arcs. L'espace probabiliste est tel que (Ω, P) où Ω est l'univers des probabilités et P l'ensemble de variables aléatoires X= {X1,…, Xn} associées aux nœuds du graphe et tel que :

$$P(X1,\ldots, Xn)=\prod_{i=1}^{n} P(Xi/Pa(Xi)) \quad (2)$$

Dans cette expression, *Pa (Xi)* est l'ensemble des parents du nœud Xi dans G [37].
La construction d'un réseau bayésien consiste donc à :

- Définir le graphe du modèle
- Définir les tables de probabilités de chaque variable, conditionnellement à ses causes.

Le graphe est aussi appelé la "structure" du modèle, et les tables de probabilités ses "paramètres". Généralement, la structure est définie par des experts et les tables de probabilités calculées à partir de données expérimentales [37].

Dans notre travail pour les tests, nous allons utiliser un réseau dont la structure est déjà connue. Il provient de l'exemple de diagnostic de la dyspnée *Asia*, qui a été introduit par [4] Figure III.1. Les huit nœuds sont binaires et on peut noter que pour l'arc entre A et T, qui nous dit que le fait d'avoir séjourné en Asie modifie le risque d'avoir contracté une tuberculose, la probabilité *a priori* est très faible et l'influence de A sur T est également très faible [38].

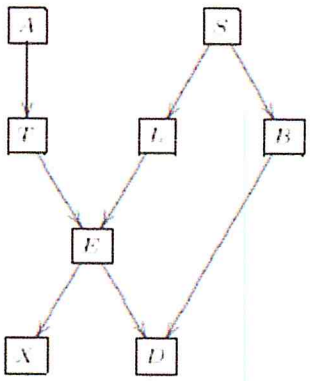

Figure III.1 : Le réseau Asia

Il existe plusieurs variantes des RB telles que [39]: les RB multi agents, les RB de niveaux deux, les RB orientés objets, les diagrammes d'influence, les RB dynamiques (temporels), les RB multi entités, les filtres bayésiens : qui sont des RB dynamiques particuliers et les RB adaptés à la classification tels que ; les RB naïf, les RB naïf augmentée etc.

En classification, particulièrement, les RB sont larges. Dans ce cas, le nœud parent est considéré comme une variable non observée précisant à quelle classe appartient chaque objet alors que les nœuds enfants sont des variables observées correspondant aux différents attributs caractérisant cet objet. Plusieurs modèles sont conçus dans ce but. Parmi ces réseaux, on peut citer le RB naïf qui est le plus simple, le réseau bayésien augmenté par n'importe qu'elle structure ou par une structure arborescente et autres [37].

III.3 Méthodes d'apprentissage des Réseaux bayésiens

III.3.1 Apprentissage de la structure

Différents facteurs peuvent venir perturber l'apprentissage de la structure du RB, le plus fréquent d'entre eux est la présence de données non observées, on parle alors de données manquantes. Dans ce cas deux approches sont possibles, la première consiste à compléter les données, avant l'apprentissage, par une inférence statistique. La deuxième solution réside dans l'utilisation d'algorithmes d'apprentissage gérant la présence de données manquantes.

Depuis plus de 20 ans de nombreuses stratégies ont été mises en place afin d'approcher au plus près la solution optimale, celles-ci sont généralement classées en deux catégories.

- **Recherche d'indépendances :** L'approche la plus naturelle consiste à rechercher toutes les relations d'indépendances entre les variables et d'obtenir ainsi les dépendances directes et donc la structure du réseau. Différentes mesures d'indépendance et stratégies ont été déployées dans cet objectif.
- **Maximisation d'un score** : L'utilisation d'un score mesurant la vraisemblance du modèle sachant les observations représente la seconde approche. Dans ce cas les heuristiques développées s'attachent à rechercher le réseau maximisant ce score.

Une troisième catégorie de méthodes dites "hybrides" ou "mixtes" combinent les deux approches précédentes [40].

La recherche de structure de réseaux bayésiens peut utiliser des bases de données incomplètes, par exemple par le biais d'un échantillonnage de Gibbs ou encore en utilisant une approche comme l'algorithme EM. D'autres travaux utilisent des techniques plus originales, comme [33] qui effectue l'apprentissage de sous-structures locales ou encore [34] qui utilise une méthode à base de recherche d'indépendances conditionnelles.

Elle utilise des tests statistiques afin de déterminer les indépendances entre les variables dans le réseau. On peut citer l'algorithme de « PC » pour Peter et Clark, « IC » algorithme de principe similaire, pour Inductive Causation « BNPC » Nommée BN-PC-B pour Bayes Net Power Constructor et B, car les mêmes auteurs introduisent deux algorithmes, le premier et BN-PC-A, « GS » comme Greedy Search ou recherche gloutonne et la « Méthode de recherche de l'arbre de recouvrement de poids maximal » (Maximal Weight Spanning Tree ou MWST). Cette méthode s'applique à la recherche de structure d'un réseau bayésien en fixant un poids à chaque arête potentielle $Xi \rightarrow Xj$ de l'arbre [37].

Ces algorithmes sont tous basés sur un principe identique :

- construire un graphe non dirigé contenant les relations entre les variables, à partir de tests d'indépendance conditionnelle,
- détecter les V-structures (en utilisant aussi des tests d'indépendance conditionnelle),
- « propager » les orientations de certains arcs,
- prendre éventuellement en compte les causes artificielles dues à des variables latentes

La caractéristique principale de toutes ces méthodes réside dans la détermination à partir de données des relations d'indépendance conditionnelle entre deux variables quelconques conditionnellement à un ensemble de variables [41].

- **PC** : L'algorithme PC utilise un test statistique tel le test du x^2 pour évaluer s'il y a indépendance conditionnelle entre deux variables. Pour cela la méthode débute par un graphe complet et étudie un à un tous les arcs de ce graphe, pour chacun d'eux, il teste dans un premier temps l'indépendance a priori (sans aucune variable de conditionnement) afin d'éliminer une première vague d'arcs entre deux nœuds déclarés indépendants. Puis il réexamine tous les arcs restant en conditionnant successivement le test d'indépendance avec chaque voisin d'un des deux nœuds, permettant ainsi d'éliminer d'autres relations. L'algorithme continue en augmentant progressivement la taille de l'ensemble de conditionnement, tant que le nombre de voisins des nœuds considérés le permet. Le graphe obtenu après cet étape est un graphe non orienté dû au caractère symétrique du test employé, s'en suit alors une orientation partielle du graphe en déterminant les v-structures à partir des indépendances conditionnelles ce qui permet également de déterminer par propagation d'autres orientations afin d'obtenir un graphe partiellement orienté [40].
- **IC :** L'algorithme IC (Inductive Causation), développé par Pearl [42], est basé sur le même principe, mais construit le graphe non orienté en ajoutant des arêtes au lieu d'en supprimer. Il faut noter que Pearl [43] a présenté en 1991 un algorithme IC différent qui prend en compte les variables latentes. Cet algorithme, renommé IC* dans [42].
- **BN-PC-B** : L'algorithme BN-PC-B [41] est le plus général des deux. Le principe de cet algorithme est simple et se décompose en trois phases : (1) utiliser l'arbre de recouvrement maximal (MWST), arbre qui relie les variables de manière optimale au sens de l'information mutuelle comme graphe non dirigé de départ, puis (2) effectuer un nombre réduit de tests d'indépendance conditionnelle pour ajouter des arêtes à cet arbre, et (3) finir avec une dernière série de tests pour supprimer les arêtes inutiles et détecter les V-structures.

Ces algorithmes consistent à parcourir tous les graphes possibles puis associer un score à chaque graphe. Le graphe qui possède le plus grand score va être sélectionné. Cependant, cette méthode est applicable seulement pour des problèmes à taille limitée

(quelques centaines de variables). Il existe deux types de score : les scores locaux et les scores globaux [37].

Pour que ces approches à base de score soient réalisables en pratique, nous verrons que le score doit être décomposable localement, c'est-à-dire s'exprimer comme la somme de scores locaux au niveau de chaque nœud. Se pose aussi le problème de parcours de l'espace B des réseaux bayésiens à la recherche de la meilleure structure. Comme une recherche exhaustive est impossible, les algorithmes proposés travaillent sur un espace réduit (espace des arbres, ordonnancement des nœuds), ou effectuent une recherche gloutonne dans cet espace [41].

- **MWST** : Apprendre un RB dont la structure est celle d'un arbre se rapproche de la recherche d'un arbre couvrant de poids maximum (MWST pour Maximum Weighted Spanning Tree). Chow and Liu propose en 1968 pour cela d'utiliser l'information mutuelle entre chaque paire de variable afin de pondérer l'arête correspondante et d'appliquer un algorithme de recherche de l'arbre couvrant de poids maximum sur la matrice ainsi créée. Le résultat est un arbre non orienté optimal maximisant l'information mutuelle par paire. D'autres mesures peuvent être utilisées pour la pondération, Heckerman propose en 1995 la variation d'un score décomposable provoqué par l'ajout de chaque arête. L'orientation de l'arbre obtenu s'effectue par propagation à partir d'une variable racine (généralement choisie aléatoirement) vers les variables les plus éloignées de cette racine. Cette orientation ne crée donc aucune v-structure rendant tout arc inversible ce qui assure de conserver le même score indépendamment du choix de la variable racine. La principale conséquence est donc l'impossibilité de déduire une quelconque causalité dans les réseaux produits [40].
- **K2** : [13] décrit l'algorithme K2 évoluant dans un espace des DAG contraint à un ordre donné. Sachant un ordre complet sur les variables, l'algorithme autorise uniquement pour chaque variable l'ajout des parents qui la précède dans cet ordre. Concrètement l'algorithme construit pour chaque variable l'ensemble de ses parents en ajoutant à chaque étape le parent autorisé au regard de l'ordre qui maximise le score sachant les parents précédemment ajoutés. La méthode telle que décrite par ses auteurs traite les variables suivant l'ordre fourni, cependant chacune des variables peut être traitée indépendamment des autres. En effet l'ordre imposé assure l'acyclicité des structures reconstruites ou lorsque le score

employé est décomposable seule la contrainte d'acyclicité empêche de traiter chaque variable indépendamment.

Les performances de cette méthode sont donc fortement liées au choix de cet ordre. Cependant le problème de recherche d'un ordre optimal est également NP-difficile. Cette recherche devient ainsi un problème en soit, une classe de méthodes développe des heuristiques afin de parcourir l'espace des ordres où l'algorithme K2 est employé en tant que routine à chaque évaluation d'un ordre. Il faut cependant noter que l'algorithme K2 n'est pas optimal bien que contraint dans son espace de recherche, l'ajout d'un seul parent à chaque étape de l'algorithme empêche la détection d'effets d'interaction de plusieurs variables qui, prisent indépendamment l'une de l'autre, n'améliorent pas le score. Ce problème est récurent pour les algorithmes qui n'appliquent qu'une seule opération élémentaire à chaque étape tout en refusant de dégrader le score. De plus l'algorithme K2 n'applique que des opérations d'ajout d'arc sans jamais considérer de suppression (les inversions sont également proscrites du fait de l'ordre) pourtant dans le cas d'effets d'interaction la présence d'un parent précédemment ajouté peut être remise en cause, après l'ajout d'un nouveau parent, dû aux changements induit sur les probabilités conditionnelles [40].

- **Méthodes hybrides**

D'autres approches, symétriques aux précédentes, vont profiter des avantages des méthodes à base de score pour aider les algorithmes d'apprentissage de structure par recherche d'indépendance conditionnelle. [34] part du fait que l'algorithme PC est sensible, tout d'abord, aux heuristiques employées pour ne pas parcourir tous les ensembles de conditionnement, et ensuite au seuil du test statistique utilisé. Ils suggèrent alors un parcours aléatoire de l'espace de ces deux paramètres (ordre permettant de limiter les ensembles de conditionnement ainsi que le niveau de signification du test) en se servant d'un score bayésien pour comparer les réseaux obtenus. Sur le même principe général, [34] présente un nouveau test d'indépendance conditionnelle Hybrid Independence Test se servant de certains avantages des approches à base de score comme l'ajout possible d'a priori et le recours à l'algorithme EM pour prendre en compte les données incomplètes.

III.3.2 Apprentissage des paramètres à partir des données incomplètes

Dans les applications pratiques, les bases de données sont très souvent incomplètes. Certaines variables ne sont observées que partiellement ou même jamais [36].

De nombreuses méthodes tentent d'estimer les paramètres d'un modèle à partir de données MAR. Citons par exemple le sequential updating, l'échantillonnage de Gibbs, et l'algorithme expectation maximisation (EM). Plus récemment, les algorithmes bound and collapse et robust bayesian estimator cherchent à résoudre le problème quel que soit le type de données manquantes.

L'application de l'algorithme itératif EM aux réseaux bayésiens a été proposée dans [44] et [45] puis adaptée aux grandes bases de données dans [46].

La méthode d'estimation de paramètres avec des données incomplètes la plus couramment utilisée est fondée sur l'algorithme itératif EM (Expectation-Maximisation) proposé par Dempster [47].

Soit :

- Xv= $\{Xv^{(l)}\}$ l=1.....N l'ensemble des données observées (visibles).
- q(t) = $\{\theta^{(t)}$ i, j, k$\}$ les paramètres du réseau bayésien à l'itération *t*.

L'algorithme EM s'applique à la recherche des paramètres en répétant jusqu'à convergence, les deux étapes Espérance et Maximisation sont décrites ci-dessous :

- Espérance : estimation des Ni,j,k manquants en calculant leur moyenne conditionnellement aux données et aux paramètres courants du réseau

$$N_i^* = E[Ni,j,k] = \sum_{i=1}^{n} P(Xi = k/Pa(Xi = C_i, X_l^{(v)}, \Theta^{(t)}) \quad (3)$$

Cette étape revient à réaliser une série d'inférences (exactes ou approchées) en utilisant les paramètres courants du réseau, puis à remplacer les valeurs manquantes par les probabilités obtenues par inférence. Dans la formule précédente Ni,j,k est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état X_k et ses parents sont dans la configuration C_j.

- Maximisation : en remplaçant les Ni,j,k manquants par leur valeur moyenne calculée précédemment, il devient possible de calculer de nouveaux paramètres $\theta^{(t+1)}$ par maximum de vraisemblance

$$\Theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^{*}}{\sum_k N_{i,j,k}^{*}} \quad (4)$$

L'algorithme EM :

-Initialiser $\theta^{(0)}$, t=0

- Répéter

- t=t+1
- calculer $N_{i,j,k}^{*}$
- calculer $\Theta_{i,j,k}^{(t+1)}$
- tant que $|\theta^{(t)} - \theta^{(t-1)}| \geq \varepsilon$

De nombreuses heuristiques ont été conçues pour accélérer ou améliorer la convergence de l'algorithme EM [45]. Citons par exemple, l'ajout d'un moment, proposé par Nowlan [48] qui permet d'accélérer la convergence si le paramètre est bien réglé :

$$\Theta_{i,j,k}^{(t+1)} \leftarrow \Theta_{i,j,k}^{(t+1)} + \gamma\ \Theta_{i,j,k}^{(t)} \quad (5)$$

Dans le cadre du réseau bayésien, pour l'apprentissage des paramètres, il suffit de remplacer le maximum de vraisemblance de l'étape M par un maximum (ou une espérance) a posteriori. Nous obtenons dans le cas de l'espérance a posteriori :

$$\Theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^{*} + \alpha_{i,j,k}}{\sum_k N_{i,j,k}^{*} + \alpha_{i,j,k}} \quad (6)$$

De nombreuses implémentations et adaptations de cette méthode ont été proposées à cause du succès de la méthode EM.

➢ **EM généralisé**

Choisir θ_{n+1} en maximisant $\Delta(\theta|\theta n)$ permet de maximiser l'augmentation de la vraisemblance $L(\theta)$ à chaque étape et donc de « converger » plus rapidement.
Néanmoins, lorsqu'il est difficile d'obtenir le maximum de $\Delta(\theta|\theta n)$, il est possible de se contenter de choisir θ_{n+1} tel que $\Delta(\theta_{n+1}|\_\theta n) > \Delta(\theta|\theta n)$ .Dans ce cas, l'augmentation de la vraisemblance n'est pas optimisée à chaque itération mais la convergence vers un point de stationnarité existera toujours (même preuve que précédemment) et il devra alors être effectué plus d'itérations pour y arriver [49].

Cette méthode est connue comme étant l'algorithme EM généralisé. Même, si en théorie, cette méthode converge moins rapidement (en nombre d'itérations), en pratique elle peut être plus rapide (en temps de calcul). En effet, il est possible d'économiser bon nombre de calculs à l'étape de maximisation, puisqu'au lieu d'évaluer toutes les valeurs pour choisir la plus grande, il est possible de se restreindre à un sous-ensemble (par exemple un voisinage) ou simplement, prendre la première valeur qui augmente la vraisemblance pour l'itération suivante.

- **EM pour la classification**

L'algorithme CEM, introduit par [50], est une adaptation de l'algorithme EM pour la classification. Celui-ci maximise alors la vraisemblance classifiant plutôt que la vraisemblance.

- **EM incrémental**

L'inconvénient de l'algorithme EM est qu'il doit être effectué hors-ligne. Or si de nouveaux exemples deviennent disponibles, il devient nécessaire de ré effectuer le calcul.

[51] a introduit une méthode basée sur l'algorithme EM appelée Voting EM permettant un apprentissage de paramètres de réseaux bayésiens de manière incrémentale.

- **Monte-Carlo EM**

Lorsque la méthode EM demande des calculs d'intégrales qui sont parfois insurmontables (ou de sommes contenant un nombre exponentiel de termes dans le cas discret ici présenté), il est possible d'utiliser des méthodes de Monte-Carlo pour leur évaluation.

La méthode Monte-Carlo Expectation Maximization (MCEM) introduite par [52] est une extension de la méthode EM qui utilise une méthode de Monte-Carlo pour évaluer ces intégrales/sommes. En pratique, il faut que la base soit suffisamment grande pour que l'estimation soit fiable.

- **EM variationnel**

Comme dans le cas précédent, lorsque l'évaluation des intégrales de l'algorithme EM est difficile, il est possible d'utiliser des méthodes variationnelles pour effectuer une estimation.

Nous sommes alors en présence d'un algorithme dits d'EM variationnel comme l'on introduit [49].

III.4 Méthodes d'inférence pour les Réseaux bayésiens

L'inférence est le calcul de la probabilité de n'importe qu'elle variable d'un modèle probabiliste à partir de l'observation d'une ou de plusieurs autres variables. Il consiste à propager une ou plusieurs informations au sein de ce réseau, pour en déduire comment sont modifiées les croyances concernant les autres nœuds [37].

En partant d'un réseau bayésien défini par un graphe et la distribution de probabilité associée (G,θ). Supposons que le graphe soit constitué de n nœuds, notés $X=\{X1, X2,..., Xn\}$. Le problème général de l'inférence est de calculer $p(X_i|Y)$, où $Y \subset X$ et $X_i \in Y$. La complexité de ce problème dépend de la structure du réseau [36].

On peut distinguer deux catégories d'algorithmes d'inférence : l'inférence exacte et l'inférence approchée. Plusieurs méthodes ou algorithmes conçues spécialement pour les problèmes d'inférence exacte pour les réseaux bayésiens. On peut citer l'algorithme de passage de messages de Pearl, dont le principe est le suivant, à chaque nœud est associé un processeur qui peut envoyer des messages à ses voisins, jusqu'à ce qu'un équilibre soit atteint, en un nombre fini d'étapes. Le problème de l'algorithme de passage de messages de Pearl est qu'il ne s'applique qu'aux arbres, l'algorithme d'arbre de jonction, qui est une généralisation de l'algorithme de passage de messages de Pearl permet de faire de l'inférence sur n'importe quel type de graphe. Cette méthode est divisée en cinq étapes qui sont : la moralisation du graphe, la triangulation du graphe moral, la construction de l'arbre de jonction, l'inférence dans l'arbre de jonction en utilisant l'algorithme des messages locaux et la transformation des potentiels de clique en lois conditionnelles mises à jour. A rappeler qu'une clique est un sous graphe du graphe G dont tous les nœuds sont connectés deux à deux. On peut citer encore l'Algorithme d'élimination des variables ou l'algorithme d'élimination de Bucket qui consiste à marginaliser la distribution de probabilité jointe d'un réseau, en procédant variable par variable. Chaque marginalisation sur une variable *Xi* donne lieu à une somme des probabilités de cette variable. Parfois, cette somme vaudra 1, ce qui conduit à l'élimination de la variable *Xi*. On procédera alors à la marginalisation sur une des variables restantes et ainsi de suite jusqu'à ce que la distribution soit marginalisée. Le problème de cet algorithme est que l'ordre dans lequel les variables sont éliminées détermine la quantité de calcul nécessaire pour marginaliser la distribution de probabilités jointe et donc la complexité de l'algorithme. Dans le cas de la classification on peut avoir soit des variables discrètes ou bien un mélange de variables discrètes

représentant les classes et les variables continues représentants les caractéristiques du modèle. Dans les deux cas le problème de l'inférence revient à calculer les probabilités *à posteriori* suivantes, ainsi la classe choisie est celle qui maximise ces probabilités :

$P\ (Ci/x)=P(Ci)\ \prod_{j=1}^{n} P(xj/Pa(xj),Ci)$ Si *Xj* admet un parent. (7)

$P\ (Ci/x)=P(Ci)\ \prod_{j=1}^{n} P(xj/Ci)$ si non. (8)

Avec :

$P(xj/Pa(xj),Ci)$: est la fréquence d'apparition de *xj* en connaissant *Pa (x j)* dans la classe *ci* si cas de variables discrètes donnée par l'équation.

$P(xj/Ci)$: est une distribution gaussienne donnée par l'équation.

Dans le cas où les probabilités conditionnelles ne sont pas exactes, pour effectuer une inférence exacte avec ces valeurs approximatives n'est plus probante. Il est alors intéressant d'effectuer une inférence approchée, en utilisant d'autres algorithmes tels que l'Algorithme de Métropolies Hastings [53], l'échantillonneur de Gibbs, Loopy belief propagation [54].

III.5 Conclusion

Les Réseaux Bayésiens sont souvent utilisés car ils ont pas mal d'avantages par rapport à autres techniques. En particulier, ils sont capables d'effectuer des raisonnements probabilistes à partir de données incomplètes alors que peu de méthodes sont actuellement capables d'utiliser les bases d'exemples incomplètes pour leur apprentissage [32].

Dans le cadre de notre travail, nous étudions la mise à jour séquentielle de la structure des réseaux bayésiens en tenant compte des données incomplètes et on s'attend à ce que la structure change dans le but de l'utiliser pour l'inférence afin de pouvoir imputer les données incomplètes.

CHAPITRE IV

APPROCHE PROPOSÉE ET EXPÉRIMENTATION

Chapitre IV : Approche proposée et expérimentation

IV.1 Introduction

Après avoir fait une étude sur les réseaux bayésiens, ces principales méthodes d'apprentissage de la structure, des paramètres et de l'inférence afin de connaitre ces étapes pour le traitement des données manquantes.

Ce chapitre sera consacré à l'approche proposée et son expérimentation.

IV.2 Contexte du travail

La décennie 2010 a vu une explosion des volumes de données informatiques mondialement produits [55]. Cette explosion provienne de l'utilisation croissante des outils de télécommunications électroniques, multiplication de capteurs connectés, production des données par les utilisateurs, mise en place des données publiques etc.

L'ensemble des données qui devient tellement gros qu'il en devienne difficile à travailler avec des outils classiques de gestion de base de données est ce qu'on appelle Big Data.

Les enjeux de Big Data présentent les caractéristiques suivantes :

- Des volumes se comptant en Téraoctets, voir en Exoctets
- Une production continue ou des actualisations fréquentes
- Une structure faible voir absente
- Une très grande hétérogénéité des sources, des formats, des informations associées (tags ou méta-tags).

IL peut arriver que ces données contiennent un certain nombre de données incomplètes, dans ce cas une importante partie de ces informations reste dormante par manque d'outils pour les traitements.

Ces données manquantes doivent être travaillées si cela n'a pas encore été fait c'est-à-dire procédé au nettoyage des données manquantes.

Dans le traitement de big data vue le volume des données, une manière de limiter le temps de calcul et les besoins en mémoire est de construire le modèle au fur et à mesure de l'arrivée des données en utilisant un algorithme d'apprentissage en temps réel et de façon continue. C'est pour tous ces raisons qu'on a opté pour les réseaux bayésiens.

IV.3 Processus de traitement des données manquantes basé sur les réseaux bayésiens

Les réseaux Bayésiens sont capables d'effectuer des raisonnements probabilistes à partir des données incomplètes alors que peu de méthodes sont actuellement capables d'utiliser les bases d'exemples incomplètes pour leur apprentissage [32].

Un avantage des réseaux bayésiens est qu'il suffit que seules les variables Xi et Pa(Xi) soient observés pour estimer la table de probabilité conditionnelle correspondante. Dans ce cas, il est alors possible d'utiliser tous les exemples (même incomplets) où ces variables sont observées. La recherche de structure de réseaux bayésiens peut utiliser des bases de données incomplètes, par exemple par le biais une approche comme l'algorithme EM.

Vu que l'utilisation des réseaux bayésiens pour la modélisation passe inévitablement par les trois étapes suivantes : la première est la détermination de la structure. Notre approche proposée pour l'apprentissage du modèle est celui de Friedman and Golszmidt qui consiste à apprendre la structure d'une façon séquentielle. L'approche est utile pour les masses de données et permet également de prendre en charge les données incomplètes pour l'apprentissage du modèle.

La seconde consiste à utiliser le modèle sortant pour l'estimation des paramètres et enfin à la dernière étape, une méthode d'inférence sera utilisée et ainsi l'information contenue dans ce réseau peut être calculée en n'importe quel nœud du réseau afin de pouvoir nettoyer les données manquantes.

Les phases de prétraitement des données dans notre contexte se résument en

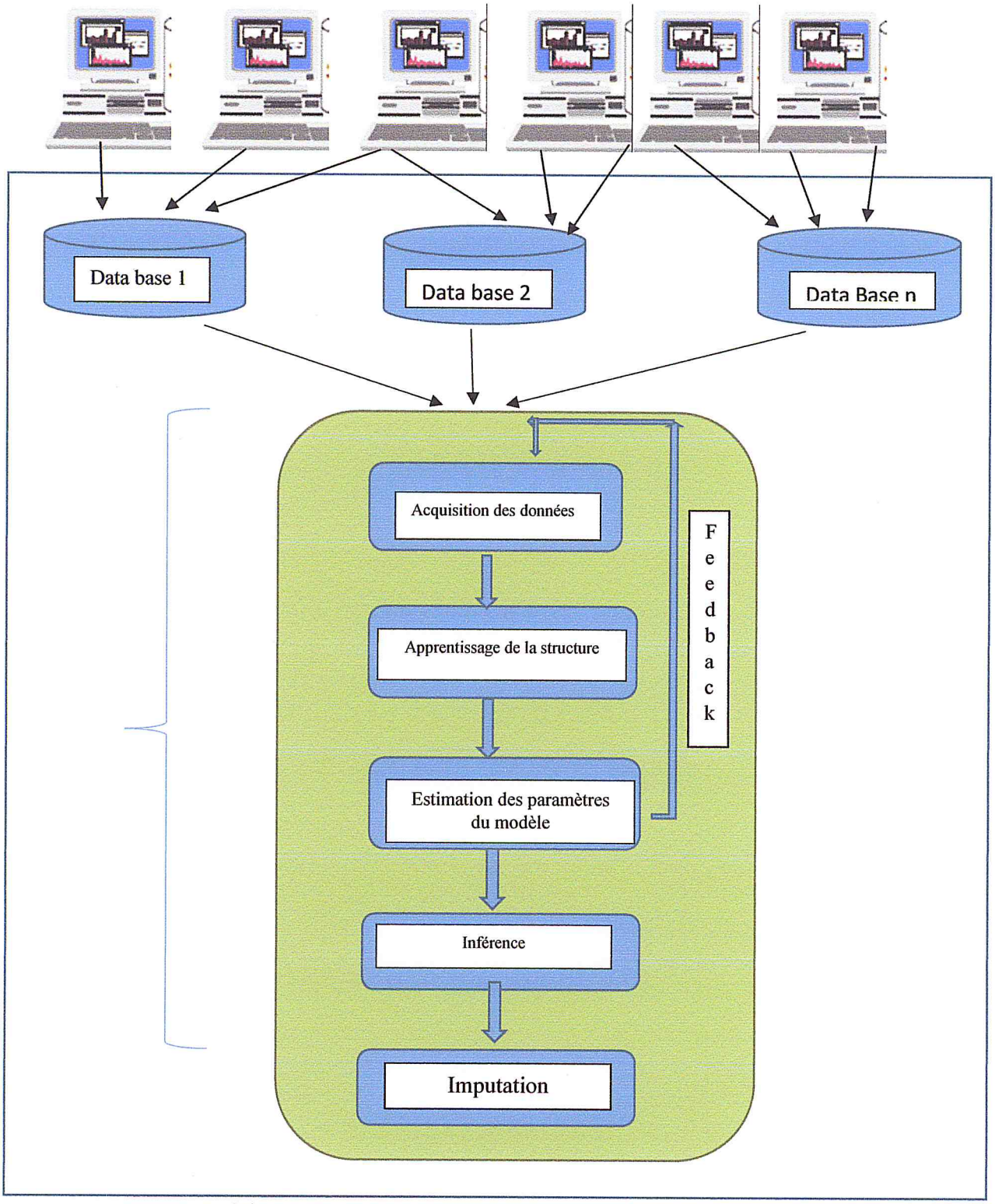

Figure IV.1 : Phase de prétraitement des données

IV.3.1 Apprentissage incrémental de la structure du Réseau Bayésien

La mise jour séquentielle des structures reste un problème d'apprentissage en ligne. A chaque itération, le procédé reçoit un nouvel enregistrement et produit la prochaine hypothèse sur le prochain enregistrement. Ce dernier est alternativement

employé pour mettre à jour le modèle et ainsi de suite. Le modèle pourrait changer après le rassemblement d'un certain nombre d'enregistrement.

Dans le cadre de ce travail, nous nous focalisons sur la description de la procédure de mise à jour séquentielle définissant le procédé d'apprentissage incrémental.

IV.3.1.1 Procédé d'apprentissage incrémental

Le procédé a comme composant de base un module qui met à jour un ensemble S d'enregistrement de statistiques suffisantes. Ces enregistrements vont permettre au procédé de mise à jour de choisir parmi un ensemble de réseaux possibles pour la mise à jour. Nous introduisons les notations nécessaires pour la bonne compréhension de l'approche [56].

$U=\{X_1, X_2,.....X_n\}$: ensemble des variables discrètes

$Val(X_i)$: ensemble des valeurs de X_i

X, Y, Z : Noms des variables

x, y, z : valeurs des variables

X, Y, Z : ensemble des variables

x, y, z : ensemble des valeurs ou Val(**X**)

P : la distribution de probabilité jointe de U

$N_X(x)$: Nombre des instances dans D où X=x

$N_x\hat{}$: vecteur de nombres des instances pour chaque valeur x de X appelé les statistiques suffisantes.

Suff(G) : $\{N_{Xi}\hat{}, pa(X_i)\ 1<i<n\}$: nombre de variables} : statistiques suffisantes pour G

S : statistique suffisante pour évaluer les voisins

Nets(S) : RBs qui peuvent être évalué en utilisant S / {Suff(G) € S}

F : ensemble des réseaux voisins (frontière), une méthode de recherche sera appliquer pour la recherche de la frontière.

Pour délibérer sur le choix entre deux structures G et G', nous devons mettre à jour l'ensemble Suff(G) et Suff (G') afin d'évaluer les deux structures dans le but d'appliquer le scoring.

Pour la généralisation à de plus grand ensemble qui couvre un sous ensemble considérable du rappel de l'espace de recherche que la procédure greedy hill climbing search fonctionne en comparant son candidat courant G à ses voisins. Dans notre cas les voisins sont obtenus par inspiration sur la méthode greedy hill climbing search en appliquant les trois opérations suivantes : Ajout, suppression et inversion d'arc [56].

Dans notre contexte, nous allons utiliser la formule K2 [57] pour l'affectation du score aux différents réseaux candidats dans le but de choisir la meilleure structure.

$$f(i,\pi i) = \prod_{j=1}^{q_i} \frac{(d_i-1)!}{(\alpha_{ij}+d_i-1)!} \prod_{k_i}^{d_i} \alpha_{ijk} \; ! \quad (1)$$

i : nom variable

di : nombre de modalités de i

qi : nombre de parent de i

α_{ijk} : nombre de cas où i et parent de j prennent K

α_{ij} : somme de de 1 à qi α_{ijk}

La procédure d'attribution de score aux structures se résume comme suit : la formule précédente sera affectée au score d'une variable et le score de la structure sera la somme des scores de toutes les variables constituant la structure.

Ainsi S se compose de toutes les statistiques suffisantes de G et ses voisins. Les réseaux contiennent beaucoup de réseaux qui sont plus simples, que l'approche peut être appliquée à n'importe quel procédé de recherche qui s'occupe de la recherche de frontière.

Après réception d'un nouveau exemple ou d'un certain nombre d'exemple, le procédé emploie les statistiques suffisantes dans S pour évaluer et choisir le réseau qui a le meilleur score dans la frontière F et une fois ce choix fait, il appelle le procédé de recherche pour déterminer la prochaine frontière et la mise à jour de S en conséquence.

L'algorithme de la procédure est décrit comme suit :

G un réseau initial.

F initialisé à la recherche de la frontière de G.

S= Suff (G) U $U_{B'\epsilon F}$Suff (G').

Forever

Read data Un.

Mettre à jour chaque enregistrement dans S utilisant Un.

If n mod k = 0 alors

G = $argmax_{G'\epsilon Net(S)}$ S (G|S)

Mettre à jour la frontière F (utilisant une procédure de recherche)

Placer S à Suff (G) U $U_{s'\epsilon F}$Suff (G').

Calculer les paramètres optimaux θ pour G de S.

Sortie (G,θ).

Algo IV.1 : Procédure d'apprentissage incrémentale [56]

La procédure a comme objectif utilisé le juste assez d'information dans le but de maintenir ses ressources pour prendre la prochaine décision dans l'espace de recherche. A chaque étape k, la procédure prend une décision et à chaque prise de décision, le procédé réapproprie ses ressources en vue de la prochaine itération ce qui implique le retrait pour l'ajout de certaines statistiques suffisantes de S.

IV.3.1.2 Adaptation de l'approche aux données manquantes

Précédemment, nous avons fait la supposition que les données sont complètes dans le sens que chaque exemple Un de données contient des valeurs pour chacun des variables dans U, ce qui n'est pas le cas malheureusement dans beaucoup d'application de vie réelle. L'une des sources de difficulté dans l'apprentissage des données incomplètes est le calcul en nature. Nous ne pouvons plus décomposer la probabilité des données. Ceci signifie que les scores MDL ou bayésien ne peuvent pas être écrits comme de termes locales mesurant à quel point nous nous modelons la probabilité de chaque variable donné, ses parents.

Dans l'article que nous utilisons, il se concentre sur la procédure EM .Afin de l'adaptation au problème de mise à jour séquentielle, nous devons l'étendre de deux restrictions et heureusement deux méthodes récentes traitent chacune de ces restrictions.

Dans la première étape, les paramètres actuels Θ sont utilisés pour calculer la valeur prévue de toutes les statistiques suffisantes des enregistrements appropriés, Suff(G).

Dans la seconde, Θ est remplacé par les paramètres Θ' estimés à partir de ces statistiques prévues .Elle est essentiellement équivalente à apprendre des paramètres des données complètes.

Dans cette approche, les nouveaux cas de données entrantes sont utilisés pour recalculer continuellement les statistiques suffisantes.

La justification intuitive derrière cette approche est que n'importe quel ordre suffisamment long de des échantillons d'i.i.d est semblable.

La procédure est décrite comme suit :

B = (G,θ) un réseau initial.

Pour tout Nxϵ Suff (G)

N (x) = N0 * P_B (x)

Forever

Read data instance y

Pour tout NxϵSuff (B)

N (x) = N (x) * α+ P_B (x/y)

Mettre à jour les paramètres de θ à partir de nouvelles statistiques suffisantes.

Sortie B

Algo IV.2 : Adaption aux données manquantes [56]

Dans cette procédure N_0 indique la confiance dans le réseau initial et α est un paramètre d'affaiblissement qui a une valeur inférieure à 1. Habituellement nous

utilisons α pour être tout à fait près de 1 par exemple 0,99. En utilisant ce paramètre d'affaiblissement, nous diminuons graduellement des anciens échantillons.

La deuxième restriction d'EM standard est qu'elle traite seulement avec l'apprentissage des paramètres dans une structure fixe. Friedman prouve que si on utilise les statistiques suffisantes prévues pour évaluer les structures alternatives utilisant le score MDL et choisit les structures qui sont affectées de plus grand score que le modèle actuel.

Ils suivent alors que nous pouvons employer les statistiques suffisantes prévues dans notre procédure de recherche pour évaluer de nouveaux modèles. Une modification simple de notre approche qui traite des données incomplètes est obtenue en combinant les deux techniques.

B un réseau initial.

F initialisé à la recherche de la frontière de B.

S= Suff (B) U $U_{B' \epsilon F}$Suff (B').

Pour tout Nxϵ S

N (x) = N0 * P_B (x)

Forever

Read data instance y

Pour tout Nx ϵ Suff (B)

N (x) = N (x) * α + P_B (x/y)

If n mod k = 0 alors

G = $argmax_{B' \epsilon Net(S)}$ S (B|S)

Mettre à jour la frontière F (utilisant une procédure de recherche)

Placer S à Suff (B) U $U_{B' \epsilon F}$Suff (B').

Calculer les paramètres optimaux θ pour G de S.

Sortie B.

Algo IV.3: Procédure d'apprentissage incrémentale adapté aux données manquantes [56]

IV.3.2 Apprentissage des paramètres à partir des données incomplètes

Dans les applications pratiques, les bases de données sont très souvent incomplètes. Certaines variables ne sont observées que partiellement ou même jamais [36].

De nombreuses méthodes tentent d'estimer les paramètres d'un modèle à partir de données MAR. Citons par exemple le sequential updating, l'échantillonnage de Gibbs, et l'algorithme expectation maximisation (EM). Plus récemment, les algorithmes bound and collapse et robust bayesian estimator cherchent à résoudre le problème quel que soit le type de données manquantes.

L'application de l'algorithme itératif EM aux réseaux bayésiens a été proposée dans [44] et [45] puis adaptée aux grandes bases de données dans [58].

La méthode d'estimation de paramètres avec des données incomplètes la plus couramment utilisée est fondée sur l'algorithme itératif EM (Expectation-Maximisation) proposé par Dempster [47].

Soit :

- Xv= $\{Xv^{(l)}\}$ l=1..N l'ensemble des données observées (visibles).
- q(t) = $\{\theta^{(t)}i, j, k\}$ les paramètres du réseau bayésien à l'itération *t*.

L'algorithme EM s'applique à la recherche des paramètres en répétant jusqu'à convergence les deux étapes Espérance et Maximisation décrites ci-dessous :

- Espérance: estimation des Ni,j,k manquants en calculant leur moyenne conditionnellement aux données et aux paramètres courants du réseau

$$N_i^* = E[Ni,j,k] = \sum_{i=1}^{n} P(Xi = k/Pa(Xi = C_i, X_l^{(v)}, \Theta^{(t)}) \quad (2)$$

Cette étape revient à réaliser une série d'inférences (exactes ou approchées) en utilisant les paramètres courants du réseau, puis à remplacer les valeurs manquantes par les probabilités obtenues par inférence.

Dans la formule précédente Ni,j,k est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état x_k et ses parents sont dans la configuration c_j.

- Maximisation : en remplaçant les Ni, j, k manquants par leur valeur moyenne calculée précédemment, il devient possible de calculer de nouveaux paramètres $\theta^{(t+1)}$ par maximum de vraisemblance

$$\Theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^{*}}{\sum_k N_{i,j,k}^{*}} \quad (3)$$

L'algorithme EM :

-Initialiser $\theta^{(0)}$, t=0

- Répéter

- t=t+1
- calculer $N_{i,j,k}^{*}$
- calculer $\Theta_{i,j,k}^{(t+1)}$
- tant que $|\theta^{(t)} - \theta^{(t-1)}| \geq \varepsilon$

Algo IV.4 : Maximization Expectation EM [36]

De nombreuses heuristiques ont été conçues pour accélérer ou améliorer la convergence de l'algorithme EM [45]. Citons par exemple, l'ajout d'un moment, proposé par [48] qui permet d'accélérer la convergence si le paramètre est bien réglé :

$$\Theta_{i,j,k}^{(t+1)} \leftarrow \Theta_{i,j,k}^{(t+1)} + \gamma \ \Theta_{i,j,k}^{(t)} \quad (4)$$

Dans le cadre du réseau bayésien, pour l'apprentissage des paramètres, il suffit de remplacer le maximum de vraisemblance de l'étape M par un maximum (ou une espérance) a posteriori. Nous obtenons dans le cas de l'espérance a posteriori :

$$\Theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^{*} + \alpha_{i,j,k}}{\sum_k N_{i,j,k}^{*} + \alpha_{i,j,k}} \quad (5)$$

IV.3.3 Inférence

L'inférence dans un réseau bayésien est le calcul de la probabilité conditionnelle des variables de requête, donnant un ensemble de variables d'évidence comme connaissance au réseau. Il consiste à propager une ou plusieurs informations au sein de ce réseau, pour en déduire comment sont modifiées les croyances concernant les autres nœuds. L'inférence dans un réseau bayésien peut être exacte ou approximative. Généralement l'inférence exacte est le NP dur [59]. En exploitant la structure du réseau,

beaucoup d'algorithmes ont été proposé dans le but de rendre l'inférence exacte pratique pour un large éventail d'applications. La complexité des algorithmes d'inférence exacte augmente excessivement avec la densité du réseau, la largeur des cliques et le nombre d'états des variables aléatoires [60].

Dans notre travail, nous allons utiliser l'inférence proposée par Jensen.

Dans une première phase, celui-ci construit une structure secondaire non orientée appelée arbre de jonction et dans un deuxième temps, il effectue des calculs probabilistes dans cette structure.

- **Algorithme de Jensen**

L'algorithme d'inférence de Jensen, Lauritzen et Olesen que nous appellerons plus simplement « algorithme de Jensen ».

La structure du graphe du réseau bayésien n'est plus directement utilisé mais il emploie une structure secondaire appelée arbre de jonction.

IV.3.3.1 Génération de l'arbre de jonction

La généralisation de l'arbre de jonction passe par les trois étapes suivantes :

1- Moralisation et suppression des orientations des arcs.
2- Triangulation
3- Construction d'un arbre à partir des cliques du graphe triangle.

La phase de moralisation consiste à reformuler la décomposition de la loi jointe sous forme d'un graphe non orienté. Elle découle de la propriété suivante de l'indépendance conditionnelle : soient X, Y, Z des variables aléatoires ou des groupes de variables alors X est indépendante de Y conditionnellement à Z si et seulement si les expressions suivantes, qui sont équivalentes, sont vérifiées.

1) $P(X, Y, Z)=P(X|Z)\ P\ (Y, Z)=P(X|Z)\ P(Y|Z)\ P(Z)$; (6)
2) $P(X, Y, Z)$ est de la forme $a(X, Z)\ b\ (Y, Z)$, où a et b sont des fonctions quelconques(7).

Lorsqu'on utilise la première propriété, on obtient une décomposition « orientée » dans le sens où elle fournit des probabilités conditionnelles (les parents des nœuds sont les variables à droite des barres de conditionnement).

Ainsi **P(V)= P(T\H,R) P(S\G,H)P(R\F)P(H\D)P(G\C)P(F\B)P(D\A,B) P(C\A) P(B)P(A)**(8)

En revanche, si l'on applique la propriété 2 à la même loi jointe, on ne peut plus obtenir un graphe orienté, simplement parce que les fonctions a et b de la propriété ne contiennent pas de barre de conditionnement. L'idée consiste alors à relier entre elles toutes les variables dont dépendent ces fonctions. Ainsi, une application de la propriété 2 similaire à celle de la propriété 1, qui avait abouti à la décomposition de la loi jointe ci-dessus, permet d'obtenir la décomposition suivante :

P(V) = a(T,H,R)b(S, G,H)c(R, F)d(H,D)e(G,C) f(F,B)g(D, A,B)h(C,A)i(B)j(A) (9)

Ce qui permet d'en déduire le graphe moralisé. Pour synthétiser, la moralisation consiste simplement à relier tous les parents d'un même nœud et à supprimer les orientations (les arcs deviennent donc des liens).

Remarquons que les cliques du graphe moralisé correspondent aux ensembles de variables des fonctions a, b, c, etc. (après absorption, telle que celle de j(A) par h(C, A) par exemple, chaque fois que c'est possible). Une clique est un ensemble maximal de nœuds tel qu'il existe un lien reliant tout couple de variables de cet ensemble. L'arbre de jonction est précisément formé à partir de ces cliques :

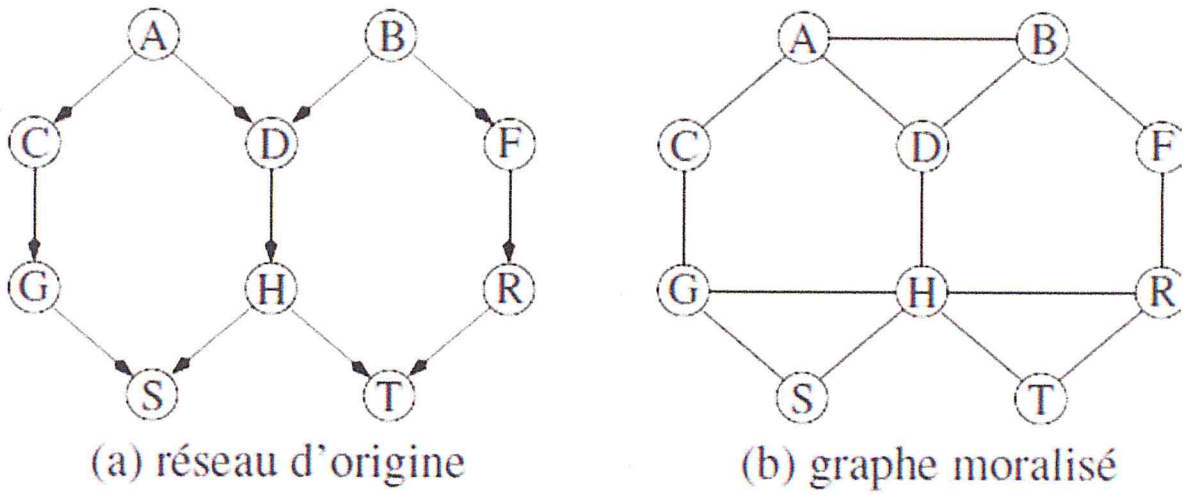

Figure IV.2 : La phase de moralisation [59]

Définition (arbre de jonction) : Soit G = (V; E) un graphe non orienté, et soit C l'ensemble des cliques de G. Un arbre de jonction est un arbre dont les nœuds sont les éléments de C et dont les liens vérifient la propriété d'intersection courante : Si C1 et C2 sont des cliques de C alors, dans toute chaîne reliant C1 et C2, toutes les cliques de la chaîne contiennent C1 ∩ C2.

Malheureusement, du graphe moralisé on ne peut pas systématiquement déduire un arbre de jonction. En effet, le graphe de jonction obtenu peut encore contenir des cycles. L'algorithme d'inférence de Jensen ne fonctionne que lorsque le graphe de jonction est un arbre. C'est pourquoi, après la moralisation, on effectue une phase de triangulation. En effet, pour un graphe G donné, il existe un arbre de jonction correspondant si et seulement si G est triangulé [60].

Un graphe G = (V; E) est triangulé si et seulement si tout cycle de longueur 4 ou plus possède une corde, c'est-à-dire que le cycle contient un couple de nœuds non adjacents dans ce cycle et reliés par un lien dans E. Comme l'a montré [61], trianguler le graphe moralisé revient à appliquer l'algorithme suivant :

Algorithme (triangulation) : Soit G = (V; E) un graphe non orienté, où V = {X1,…...Xn}. Soit σ une permutation de {1,…, n}. On dit qu'on élimine un nœud Xi de G lorsque l'on rajoute à ε des liens entre tous les voisins de Xi (Xi et ses voisins forment alors une clique) et qu'ensuite on supprime Xi ainsi que ses liens adjacents de G.

Trianguler G consiste :

1) à éliminer successivement les nœuds $X\sigma(1),\ldots, X\sigma(n)$, et à créer ε_T l'ensemble des liens ajoutés successivement à G lors de ces éliminations ;

2) à former le graphe $G_T = (V; \varepsilon \cup \varepsilon_T)$.

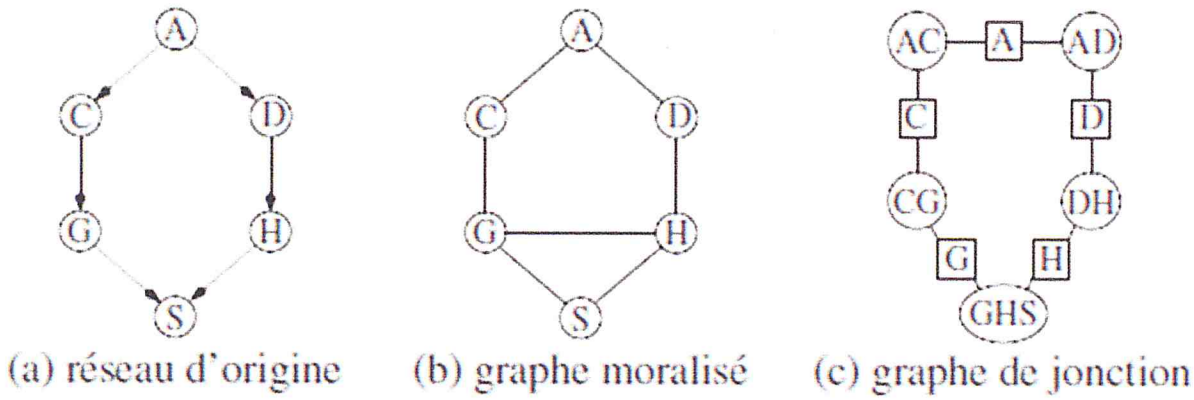

Figure IV.3 : Graphe moralisé et graphe de jonction [59]

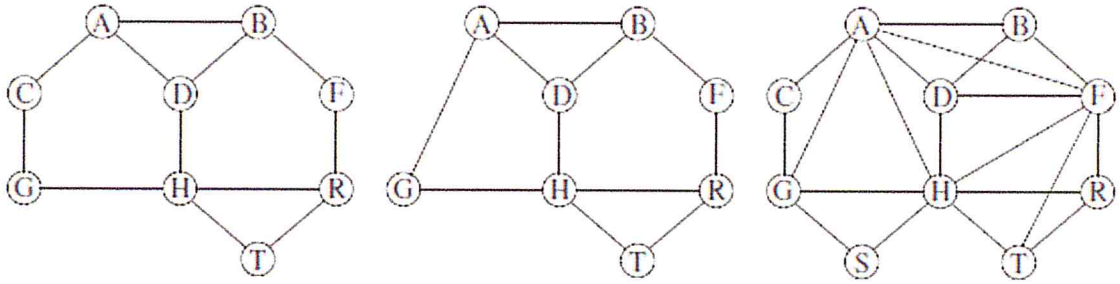

Figure IV.4 : La phase de triangulation [59]

À titre d'exemple, reprenons le graphe moralisé de la **figure IV.2 b** et appliquons la séquence d'élimination S, C, G, R, T, B, D, H, F, A. L'élimination du nœud S ne rajoute aucun lien puisque les voisins de S, les nœuds G et H, sont déjà reliés entre eux. À l'issue de cette élimination, on obtient le graphe de la **figure IV.4 a**. Lorsque l'on élimine C, on doit rajouter le lien (A, G) car A et G sont les voisins de C et ne sont pas encore reliés. À l'issue de cette élimination on obtient alors le graphe de la figure 4b, et ainsi de suite. Enfin, le graphe triangulé GT obtenu en rajoutant au graphe moralisé les liens de triangulation est représenté sur la **figure IV.4 a**.

L'arbre de jonction est obtenu en prenant l'ensemble des cliques du graphe triangulé G_T, et en les reliant de manière à respecter la propriété de l'intersection courante. Pour cela, le lecteur pourra se reporter à [60]. Notons que pour un graphe triangulé, il peut exister plusieurs arbres de jonction. Cela dit, comme le montre [62], tous ces arbres possèdent les mêmes cliques et les mêmes séparateurs. À titre d'illustration, la **figure IV.5** montre un arbre de jonction correspondant au graphe triangulé de la **figure IV.4 a**.

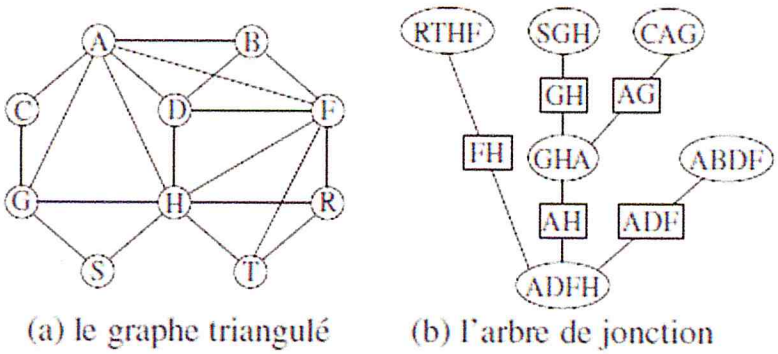

Figure IV.5 : L'arbre de jonction [59]

IV.3.3.2 Inférence dans l'arbre de jonction

Une fois l'arbre de jonction construit, Jensen associe à chaque clique et à chaque séparateur un potentiel, c'est-à-dire une fonction de l'ensemble des variables de la clique. Celles-ci correspondent, dans l'esprit, aux fonctions a, b, etc., que nous avions obtenues dans l'équation (9). Celles-ci correspondent, dans l'esprit, aux fonctions a, b, etc., que nous avions obtenues dans l'équation (9). Nous ne décrirons pas ici l'algorithme permettant d'initialiser ces potentiels et nous laisserons le lecteur se reporter à [63].Simplement, notons qu'à l'issue de cette phase d'initialisation les potentiels contiennent les probabilités jointes de toutes les variables des cliques et séparateurs auxquels ils sont associés. Dans la suite, nous noterons $\emptyset$ les potentiels des séparateurs et ψ ceux des cliques.

Le problème auquel nous allons nous intéresser est celui du calcul des probabilités marginales a posteriori, celui des probabilités a priori n'étant pas en soi extrêmement intéressant dans la mesure où il n'est effectué qu'une seule fois (pendant la phase d'initialisation). Considérons donc la portion d'arbre de jonction de la figure 6.

Supposons qu'à l'issue de la phase d'initialisation, les potentiels de Ci, Cj et S sont respectivement ψ **(Ci) = P(Ci),** ψ **(Cj) = P(Cj) et** $\emptyset$ **(S) = P(S).** Puisque S est un séparateur, S=Ci∩Cj et donc $\emptyset(S) = \sum_{Ci\setminus S} P(Ci) = \sum_{Ci\setminus S} P(Cj)$. On dit alors que l'arbre a atteint un état d'équilibre dans la mesure où tous les potentiels des cliques et des séparateurs sont cohérents entre eux.

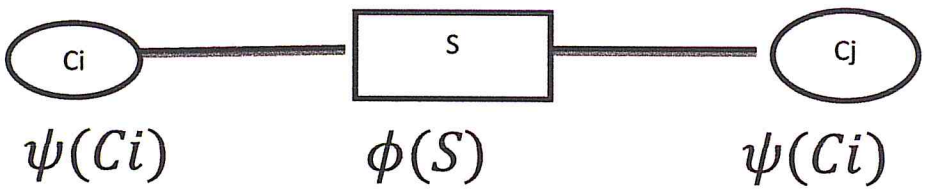

Figure IV.6 : Algorithme d'inférence de Jensen [59]

Maintenant, on apprend une nouvelle information et celle-ci a pour effet de modifier ψ (Ci) en ψ *(Ci). Ce peut être, par exemple, une information e_X indiquant qu'une variable X de Ci, qui pouvait a priori prendre les valeurs $x_1,\ldots..,x_k$, ne peut plus prendre la valeur x_1, ou bien encore que l'on vient d'observer qu'une variable Y avait pris la valeur y. Dans tous les cas, nous supposerons que toute information e_X relative à un nœud X est indépendante conditionnellement à X des autres variables du réseau et

des informations relatives à ces variables. Revenons à notre nouvelle information qui a modifié ψ (Ci) en ψ *(Ci). L'équilibre est rompu puisque $\phi(S) \neq \sum_{Ci\setminus S} \psi^*(Ci)$ .Pour rétablir ce dernier, il suffit d'appliquer la double opération suivante, que l'on appelle une absorption de Ci par Cj :

1. $\phi(S) = \sum_{Ci\setminus S} \psi^*(Ci)$ (10), remplace le potentiel ∅ **(S)** associé à S ;

2. $\psi^*(Cj) = \psi(Cj) \times \frac{\phi^*(S)}{\phi(S)}$ (11) ,Remplace le potentiel ψ (Cj) associé à Cj

Il est alors facile de voir qu'à l'issue de ces deux étapes, l'arbre est à nouveau dans un état d'équilibre. L'algorithme d'inférence de Jensen consiste à utiliser de manière systématique la procédure d'absorption pour rétablir l'équilibre dans tout l'arbre de jonction. Pour assurer que la totalité du graphe sera traitée correctement, [63] montre qu'il suffit de choisir au hasard une clique C, la racine, et de lui appliquer successivement les deux fonctions Collecte et Distribution, ci-dessous. Une fois le graphe à nouveau dans un état d'équilibre, la probabilité marginale a posteriori de chaque variable Xi peut être déterminée simplement en choisissant une clique C (resp. un séparateur S) contenant la variable Xi et en calculant $\sum_{C\setminus\{Xi\}} \psi^*(C)$

(resp. $\sum_{C\setminus\{Xi\}} \phi^*(S)$)

Algorithme (Collecte sur une clique Ci) : pour toutes les cliques Cj adjacentes à Ci excepté, si elle existe, la clique qui a appelé la Collecte de Ci, faire :

(1) appeler Collecte sur Cj,

(2) effectuer l'absorption de Cj par Ci.

Algorithme (Distribution sur une clique Ci) : pour toutes les cliques Cj adjacentes à Ci excepté, si elle existe, la clique qui a appelé la Distribution de Ci, faire :

(1) effectuer l'absorption de Ci par Cj,

(2) appeler Distribution sur Cj.

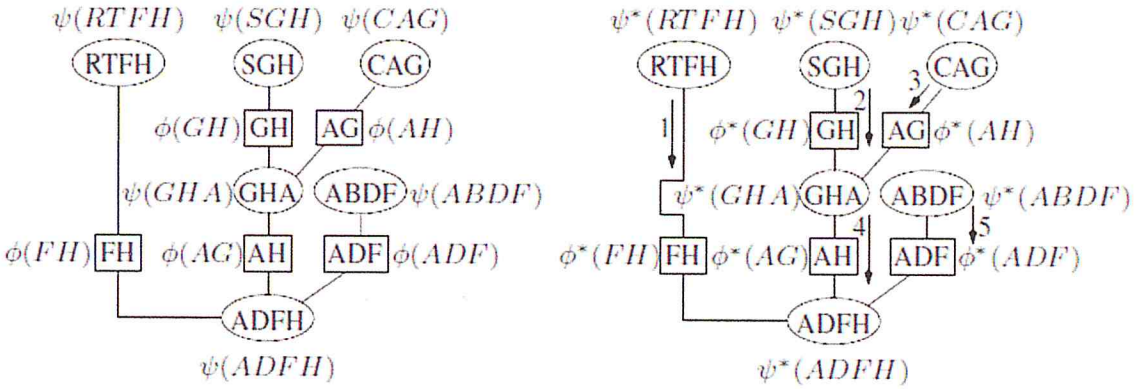

Figure V.7 : Illustration de l'algorithme d'inférence de Jensen [59]

Illustrons cette méthode sur l'exemple de la **figure IV.6** : sur la partie gauche de la figure sont représentés les potentiels à l'issue de l'initialisation de l'arbre de jonction. De nouvelles informations e_T et e_C relatives à T et C sont apparues. Pour propager ces informations dans tout l'arbre de jonction, nous avons choisi arbitrairement la clique ADFH comme racine. Nous appelons donc la fonction Collecte sur cette clique. Celle-ci appelle alors Collecte sur la clique RTFH. Cette clique n'ayant pas d'autre voisin, elle met à jour sa probabilité $\psi*(\text{RTFH}) = P(R, T, F, H, eT)$ et la clique ADFH « absorbe » cette information : $\phi(\text{FH})$ est remplacée par $\phi*(\text{FH}) = P(F, H, e_T)$ et (ADFH) est remplacé par $\psi(\text{ADFH}) \times \phi*(\text{FH})/\phi(\text{FH})$. La clique ADFH exécute alors la fonction Collecte sur GHA qui, à son tour, l'exécute sur les cliques SGH puis CAG. SGH n'ayant reçu aucune information, $\psi*(\text{SGH}) = \psi(\text{SGH})$. La clique AGH absorbe alors les informations de SGH puis de CAG, elle a alors le potentiel $\psi*(\text{GHA})$, qu'elle transmet à ADFH pour absorption. Enfin, cette dernière exécute Collecte sur ABDF et absorbe ses informations de manière à obtenir le potentiel $\psi*(\text{ADFH})$. Dans une deuxième étape, on appelle la fonction Distribution sur ADFH et l'équilibre est rétabli.

IV.4 Expérimentation

Pour vérifier la praticabilité de méthodes proposées dans ce document, nous avons implémenté les algorithmes présentés. Nous avons principalement examiné l'exactitude et l'efficacité de l'apprentissage de la structure, la convergence de l'inférence et finalement le nettoyage de données incomplètes.

Dans notre expérience nous avons adopté le réseau bayésien classique Asia. On fait le test avec le réseau Asia modifié de la figure suivante.

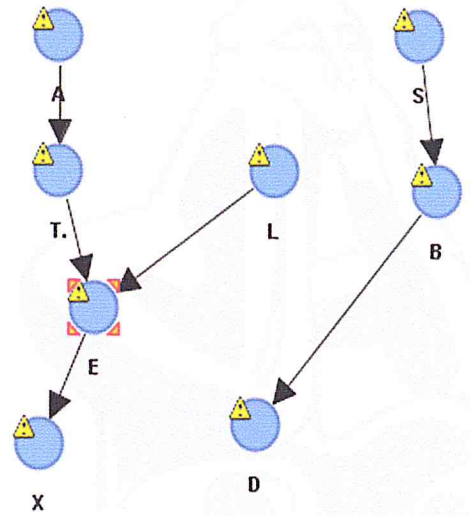

Figure IV.8: Réseau original Asia modifié

En un premier test on va le test avec 50 enregistrements lus et examiné l'état de la structure au fur et à mesure tout en examinant le nombre d'arcs de plus et le nombre d'arcs de moins par rapport au réseau original et en deuxième, recommencé la même procédure mais cette fois ci avec 100 enregistrements lus. Les résultats sur les deux figures respectivement.

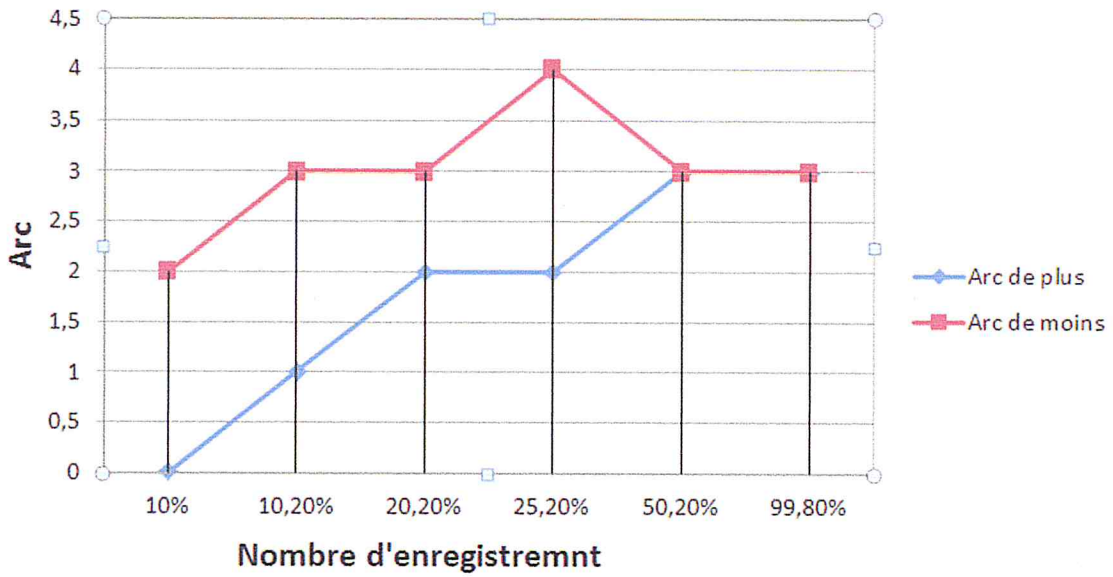

Figure IV.9 : Apprentissage de la structure avec 50 enregistrements lus

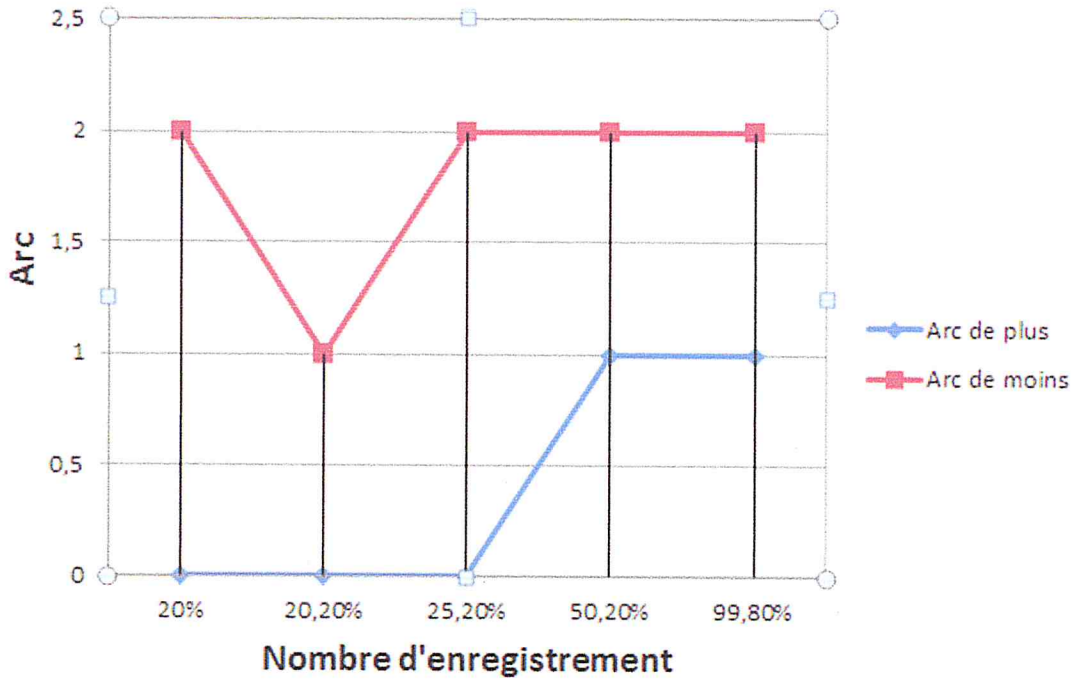

Figure IV.10 : Apprentissage de la structure avec 100 enregistrements lus

Les arcs qui existent dans la structure originale et qui n'existent pas la structure courante sont appelés les arcs de moins et les arcs qui n'existent pas dans la structure originale et qui existent dans la structure courant appelés les arcs de plus. Par rapport à ces deux tests on peut en conclure que plus le nombre d'enregistrement lus au départ est grand plus la structure est précise.

Pour la précision du nettoyage des données incomplètes à l'aide de l'inférence, d'abord nous avons comparé la valeur la plus possible prévue par l'inférence à l'aide de logiciel Bayes Server à la vraie valeur dans l'ensemble de données original. Nous allons utiliser 1 et 0 pour dénoter respectivement le cas où la valeur prévue est la même avec la vraie valeur, le cas où la valeur prévue n'est pas la même avec la vraie valeur. La précision moyenne a été définie par moyen de 0 ou de 1 pour tous les tuples inachevés. On calcule le nombre de la vraie valeur trouvée et ensuite la précision à travers le nombre total de tuples incomplètes moins le nombre de la vraie valeur trouvée. La figure IV.13 nous montre la précision moyenne des valeurs les plus possibles obtenues avec l'inférence à l'aide de bayes Server avec différentes proportions de données incomplètes.

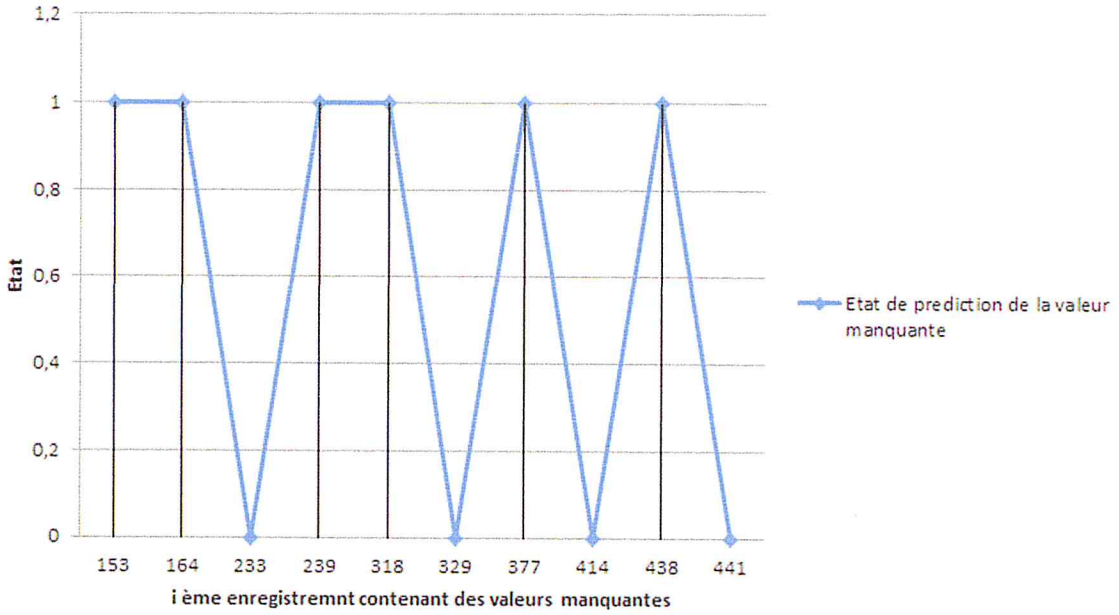

Figure IV.11 : Inférence sur 10 enregistrements contenant des valeurs manquantes

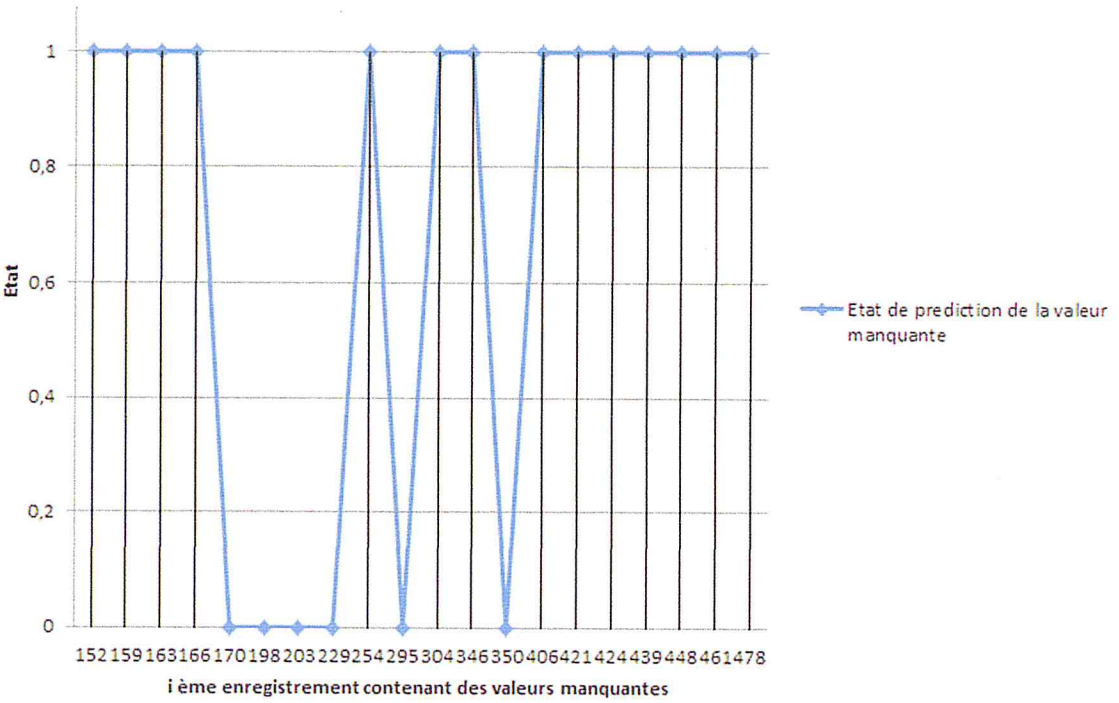

Figure IV.12 : Inférence sur 20 enregistrements contenant des valeurs manquantes

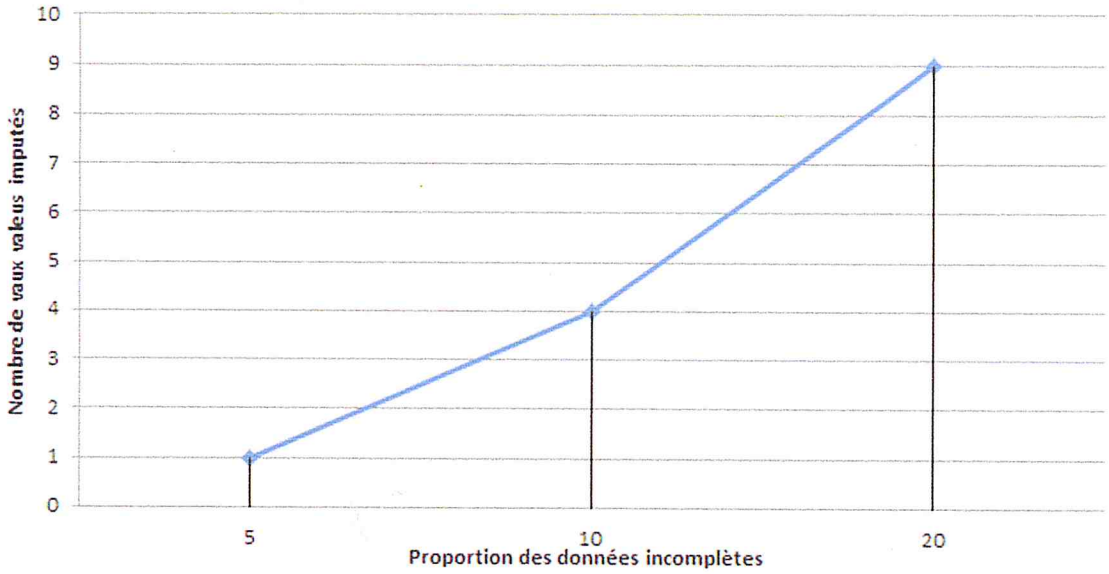

Figure IV.13 : Précision de nettoyage des données incomplètes

Nous pouvons constater sur le graphe que plus la proportion des données incomplètes augmente, plus le nombre de la valeur prévue différent de la vraie valeur augmente et l'on peut conclure que précision sera diminuée lentement avec l'augmentation des proportions des enregistrements incomplets.

Ainsi, l'exactitude de notre inférence pour le nettoyage des données incomplètes est principalement déterminée par la structure du réseau avec une certaine partie d'enregistrement inachevé.

IV.5 Conclusion

L'adaptation des méthodes existantes à la prise en compte de données manquantes est importante pour pouvoir traiter des problèmes réels. Dans ce chapitre nous avons utilisé la méthode proposée par Friedman and Goldszmidt permettant de faire la mise à jour séquentielle de la structure en faisant un survol sur la méthode d'inférence de Jensen.

Conclusion générale

1. Conclusion

La qualité des données est avant tout un problème métier, pas seulement un problème informatique.

Au cours de la réalisation de ce mémoire, on est passé d'une généralisation à une spécialisation. En premier lieu nous avons abordé la notion Data Quality et son importance en se focalisant sur les principales étapes de prétraitement de données. De ces étapes de prétraitement de données, nous avons mis l'accent sur Data Cleaning en évoquant ces principales tâches et méthodes et pour finir on s'est fixé comme objectif le traitement des données incomplètes.

Il est plus en plus fréquent d'être confronté à des applications où les données sont nombreuses mais incomplètes et que les utilisateurs aimeraient exploiter de manière à extraire le plus d'information possible.

Au cours de ce projet, nous avons étudié les différentes méthodes de traitement de données incomplètes et faire une étude comparative de ces différentes méthodes ce qui nous a amené à opter pour le choix de l'imputation à travers les réseaux bayésiens.

Naturellement, l'imputation est réellement utile si l'analyse de la base complétée par des méthodes statistiques ou des procédures de data mining donne des résultats fiable et les réseaux bayésiens sont capables d'effectuer des raisonnements probabilistes à partir de données incomplètes alors peu de méthodes sont actuellement capables d'utiliser les bases d'exemple incomplètes pour leur apprentissage.

Une mauvaise qualité de données coute cher et conduit à des ruptures dans les processus, à des décisions métiers moins pertinents et a une gestion médiocre de la relation client.

Plus la qualité sera incorporée aux habitudes e à la culture de nos entreprises, plus la démarche qualité progressera, et Paradoxalement son succès résidera dans sa banalisation.

2. Perspectives

Et comme perspectives, nous envisageons de prendre en compte des Noisy Data, les données inconsistantes, les doublons pour le traitement des données.

Adapter l'approche de Friedman et Goldszmidt (56).que nous avons utilisé, procédure d'estimation des paramètres et les différentes procédures de l'inférence à la technique de MapReduce dans le but de réduire le temps d'exécution.

Bibliographie

Bibliographie

*[1]. **Kamel, Magdi.**Data Ware housing and Mining. s.l. : IGI Global, 2009.*

*[2]. **A. F., Elgamlal*, N.A. , Mosa et N.A., Amasha.**Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse. International Journal of Soft Computing and Engineering (IJSCE) , January 05, 2014, pp. Volume-3, Issue-6,pp 226.*

*[3]. **Livre Blanc.**Qualité de données,Quelles(s) vérité(s) dans les entreprises. Juin 2011.*

*[4]. **Optimind.**les dossiers techniques d'information, La donnée en assurance. Septembre 2012.*

*[5]. **Informatica, Livre Blanc de JEMM research.**Des données de qualité,Exploitez le capital de votre organisation. Janvier 2008.*

*[6].**Fatimah, Sidi, et al..**Data Quality:A Survey of Data Quality Dimensions. 978-1-4673-1090-1/12/, Malaysia : IEEE, 2012.*

*[7]. **Batini C., Cappiello C., Francalanc C. i, Maurino A.**Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR) .200 .vol. 41. p. 16.*

*[8]. **Batini C. and Scannapieca M.**Data quality: Concepts, methodologies and techniques. s.l. : Springer-Verlag New York Inc. 2006.*

*[9]. **Rye, Kyung-seok.** A Study on Data Quality Management Maturity Mode. International Conference on Advanced Communication Technology.2005*

*[10]. **Bagchi S., Xue Bai and Kalagnanam J.** "Data Quality Management using Business Process Modeling". IEEE International Conference on Services Computing, 2006*

*[11]. **Jianjun Cao, Xingchun Diao,Guoquan Jiang, et al** Data Lifecycle Process Model and Quality Improving Framework for TDQM Practices . International Conference on E-Product E-Service and E- Entertainment. 2010*

*[12]. **Lucas. A.** Corporate Data Quality Management Towards a Meta- Framework.. International Conference on Management and Service Science (MASS).2011*

*[13]. **Kuang Chen.** Usher: Improving Data Quality with Dynamic Forms, IEEE trans. Know-ledge and data engineering. vol. 23, No. 8,August 2011*

*[14]. **McGilvray, D .** Ten Steps to Quality Data and Trusted Information . USA .2008.*

Bibliographie

*[15]. **Goulet, Martin.**Hiérarchiser les dimensions de la qualité des données : analyse comparative entre la littérature et les praticiens en technologies de l'information. Québec, Canada. février 2012.*

*[16].**Batini, C, et al..**Methodologies for Data Quality Assessment and Improvement. s.l. : ACM Computing Surveys. juillet 2009. Vol. vol. 41 no 3, p. 16:1-16:52.*

*[17]. **Batini, C.** Data Quality Concepts, Methodologies and Techniques. New York .s.l: Springer. 1998.*

*(18). **Madnick, S.E, Wang, R.Y et Zhu, H.** Overview and Framework for Data and Information Quality Research : ACM Journal of Data and Information Quality. juin 2009. Vol. 1.*

*[19]. **Institut canadien d'information sur la santé.** Le cadre de la qualité des données de l'ICIS 2009.7 décembre 2010.*

*[20]. **Mengjie Chen, Meina Song et Jing Han, Haibong E.** Survey of quality of data. Chine.World Congress on Information and Communication Technologies. p. 4., 2012.*

*[21]. **Leila , Ben Othman Amroussi.**Conception et validation d'une méthode de complétion des valeurs manquantes fondée sur leurs modèles d'apparition. Université de Caen.2011.*

*[22]. **Carpenter, J. R et Kenward, M. G.** Missing data in randomised controled trials. 2008.*

*[23]. **Stéphane , Paquin .**Comparaison de quatre méthodes pour le traitement des données manquantes au sein d'un modèle multiniveau paramétrique visant l'estimation de l'effet d'un programme de prévention. Université de Montréal .2010.*

*[24]. **Graham, J. W.**Missing Data Analysis: Making It Work in the Real World. 2009.*

*[25]. **ABDERRAZAK , BENNANE .**TRAITEMENT DES VALEURS MANQUANTES POUR L'APPLICATION DE L'ANALYSE LOGIQUE DES DONNEES À LA MAINTENANCE CONDITIONNELLE. UNIVERSITÉ DE MONTRÉAL.AOUT 2010.*

*[26]. **Sophie, O'PREY.**Mise en oeuvre de nouvelles procédures de redressement et comparaison de méthodes d'imputation. s.l. : Direction des statistiques demographiques et sociales. 2008.*

*[27]. **Benoît , Virole .** Etude prospective des applications possibles des réseaux de neurones formels dans le traitement des données psychométriques. . s.l: Editions du Centre de Psychologie Appliqué.Juin 2001.*

*[28]. **Antonio , ANSELMI, Paola , M. Chiodini et Flavio, Verrecchia.** Données manquantes et prévisions: Données manquantes et prévisions: Méthodes à imputation variable. Mars 2009.*

Bibliographie

*[29]. **Oumy , Niass, Abdou , Kâ Diongue et Aïssatou , Touré.** ETUDE DES DONNÉES MANQUANTES EN SÉRO-ÉPIDÉMIOLOGIE. 2012.*

*[30]. **Dr Roch , Giorg.**Imputation Multiple pour la Prise en Compte de Données Manquantes - Application à la Survie Relative. Marseille : LERTIM.2011.*

*[31]. **Céline , Fiot.**Extraction de séquences fréquentes : des données numériques aux valeurs manquantes. Université Montpellier II. septembre 2007.*

*[32]. **Olivier, François et Philippe , Leray.** Apprentissage de structure des réseaux bayésiens et données incomplètes. s.l. : INSA Rouen - Laboratoire PSI.*

*[33]. **Sebastiani , P et Ramoni , M.** Bayesian selection of decomposable models with incomplete data. s.l. : Journal of the American Statistical Association.2001. Vol. Vol. 96, No. 456.*

*[34]. **Dash , D et Druzdzel , M.** Robust independence testing for constraint- based learning of causal structure. s.l. : Proceedings of The Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03). 2003.*

*[35]. **Philippe, Leray et Olivier, François.**Réseaux Bayésiens pour la Classification Méthodologie et Illustration dans le cadre du Diagnostic Médical. s.l. : INSA Rouen/ PSI,FRE CNRS 2645 . 2002.*

*[36]. **NGUYEN , Trung Thanh .**Réseaux Bayésiens. s.l. : Institut de la Francophonie pour l'Informatique, 2005.*

*[37]. **ZAABOT , Zohra.**Les Réseaux Bayésiens. Application en Reconnaissance de Formes à partir d'Informations Complètes ou Incomplètes. UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU .2012.*

*[38]. **Olivier , François et Philippe, Leray.** Étude Comparative d'Algorithmes d'Apprentissage de Structure dans les Réseaux Bayésiens. s.l. : Laboratoire Perception, Systèmes, Information - FRE CNRS 2645. 2003.*

*[39]. **Smail, L.**Algorithmes pour les réseaux bayésiens et leurs extensions. Université de Polytech Nantes . 2004.*

*[40]. **Jimmy, VANDEL.**Apprentissage de la structure de réseaux bayésiens. Application aux données de génétique-génomique. UNIVERSITÉ DE TOULOUSE. Décembre 2012.*

*[41]. **Philippe , LERAY.**Réseaux bayésiens : apprentissage et modélisation de systèmes complexes. Université de Rouen .Mai 2010.*

*[42]. **Judea, Pearl.** Causality : Models, Reasoning, and Inference. s.l. : Cambridge University Press, 2000.*

Bibliographie

*[57]. **Ozge, Doguc et Josez, Emmanuel Ramirez-Marque.** A genericmethodforestimatingsystemreliabilityusingBayesiannetworks. U.S.A : Stevens Institute of Technology, 2009.*

*[58]. **Bo , Thiesson, Christopher , Meek et David , Heckerman.** Accelerating EM for large databases.*

*[59]. **Nam, Ma, Yinglong , Xia et Viktor , K. Prasanna.**Parallel Exact Inference on Multicore.*

*[60]. **Lauritzen, S. L. et Spiegelhalte, D. J.** , "Local computation with probabilities and graphical structures and their application to expert systems. s.l. : Royal Statistical Society B. 1988, Vol. vol. 50.*

*[61].**Olfa , Ben Naceur-Mourali et Christophe , Gonzales.** Une unification des algorithmes d'inférence. s.l. : RSTI - RIA, 2004.*

*[62]. **COWELL , R., et al., et al.**Probabilistic Networks and Expert Systems. s.l. : Springer-Verlag. 1999.*

*[63]. **ROSE, D.** Triangulated Graphs and the Elimination Process . s.l. : Journal of Mathematical Analysis and Applications.1970. Vol. vol. 32.*

*[64]. **JENSEN , F.** Optimal junction trees-Proceedings of the Tenth Conference on Uncertainty in Articial Intelligence. 1994.*

*[65)]. **JENSEN , F.** An Introduction to Bayesian Networks. North America .s.l: Springer-Verlag.1996.*

Bibliographie

*[43]. **Judea, Pearl et Tom , Verma.** A theory of inferred causation. San Mateo, California . 1991.*

*[44]. **Robert, Cowell, et al.** Probabilistic Networks and Expert Systems-Statistics for Engineering and Information Science., s.l. : Springer-Verlag.1999.*

*[45]. **Radford , Neal et Geoffrey, Hinton.**A view of the EM algorithm that jus- tifies incremental, sparse and other variants. Kluwer Academic Publi- shers, Boston, : In Michael Jordan, editor. 1998.*

*[46]. **Joe, Suzuki.** Learning Bayesian belief networks based on the MDL principle : An efficient algorithm using the branch and bound technique. E82-D(2), s.l. : IEICE Transactions on Information and Systems, 1999.*

*[47]. **Dempster, A., Laird, N. et Rubin , D.** Maximum likelihood from incomplete data via the EM algorithm. s.l. : Journal of the Royal Statistical Society. 1977.*

*[48]. **Nowlan, S.** Soft competitive adaptation : Neural Network Learning Algorithms based on Fitting Statistical Mixtures. Carnegie Mellon Univ.1991.*

*[49]. **François , Olivier.**De l''identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes. Institut National des Sciences Appliquées de Rouen .novembre 2006.*

*[50]. **Celeux, C. et Govaert, G.**A classification EM algorithm for clustering and two stochastic versions-Computational Statistics and Data Analysis. 1992.*

*[51]. **Cohen, I, Bronstein, A et Cozman, F.** Adaptive online learning of bayesian network parameters. 2001.*

*[52]. **Wei, G et Tanner, M.**A monte-carlo implementation of the em algorithm and the poor man's data augmentation algorithms. s.l. : Journal of the Americam Statistical Association. 1990.*

*[53]. **Mazet, V.** Introduction aux méthodes de Monte Carlo par chaînes de markov. Université Henri Poincaré : CRAN.CNRS UMR 7039. Mai 2003.*

*[54]. **Tanaka, K., Imoune, J. I et Titterington, D.M.** loopy belief propagation and probabilistic image processing. 1999.*

*[55]. **François Stephan , Stephan et Renaud, Bonnet.** Big Data, tirer parti de l'explosion. s.l. : crip(Club des Responsables d'Infrastructures et de Production). Janvier 2012.*

*[56]. **Nir , Friedman et Moises, Goldszmidt.** Sequential Update of Bayesian Network Structure. s.l. : University of California,SRI International.*