Une nouvelle approche d'estimation pour les entrepôts de données multi-granulaires incomplètes

Nestor Koueya*, Sandro Bimonte**,
Engelbert Mephu Nguifo***,****

*Laboratoire d'informatique, Université de Dschang, Cameroun
koueya@gmail.com
**Irstea, TSCF, Clermont Ferrand
sandro.bimonte@irstea.fr
***Université Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND
****CNRS, UMR 6158, LIMOS, F-63173 AUBIERE
mephu@isima.fr

Résumé. Les entrepôts de données spatiales (EDS) sont caractérisés par une forte corrélation des données. De ce fait, les méthodes d'interpolation spatiales et temporelles sont très utilisées pour estimer les faits manquants. Ces méthodes ignorent souvent la présence éventuelle des mesures agrégées. Ce qui entraîne un biais sur l'agrégation. Nous proposons une approche qui adapte les fonctions d'estimation existantes pour la prise en compte des des mesures agrégées connues.

1 Introduction

Les Systèmes d'Aide à la Décision (SAD) sont des systèmes d'informations flexibles et interactifs qui aident les décideurs dans l'extraction d'informations utiles pour identifier et résoudre des problèmes et pour prendre des décisions. Parmi les SAD, les systèmes d'entrepôts de données sont probablement les plus utilisés dans le monde académique et industriel (Bimonte, 2007). Un entrepôt de données est une « collection de données orientées sujet, intégrées, non volatiles et historisées, pour l'aide à la décision » (Inmon, 1996). Ces données sont analysées en utilisant les opérateurs OLAP qui permettent l'exploration en ligne des données entreposées selon le modèle multidimensionnel. Les opérateurs OLAP intègrent des fonctions d'agrégation qui permettent la visualisation des données à différents niveau de détails ou granularités. Au niveau des granularités fines on retrouve les données détaillées ou micro données, alors que les données agrégées sont retrouvées au niveau des granularités élevées. Les données ou mesures agrégées résultent des calculs (somme, moyenne, etc.) opérés sur les données détaillées. Elles sont souvent stockées pour faciliter la navigation dans les Bases de Données Multidimensionnelles (BDM).

Cependant, les valeurs incomplètes sont endémiques aux bases de données (Dyreson et al., 2003). Cette assertion est valable pour les BDM. Leur présence peut influencer négativement la qualité des mesures agrégées (décisionnelles), puisque les résultats des analyses

fondées sur des données incomplètes peuvent êtres inexacts (Dyreson et al., 2003). Les recherches sur les informations incomplètes ont été intensives dans le contexte des bases de données relationnelles, déductives et orientées objets (Dyreson, 1997). Ces recherches couvrent les valeurs manquantes qui représentent la forme la plus connue des informations incomplètes. L'estimation est une approche de traitement des valeurs manquantes très utilisée dans le contexte des BDM. Les méthodes d'estimation pour ce qui est des valeurs ou faits détaillés se regroupent en deux catégories : méthodes horizontales et descendantes. En effet, les méthodes horizontales estiment les faits manquants sur la base des relations de similarités, des ressemblances/dissemblances, des corrélations et autres. Dans le contexte spatiotempotel-multidimensionnel, l'interpolation [1] est la méthode d'estimation horizontale la plus répandue. Ces méthodes horizontales estiment les faits manquants sans tenir compte d'une existence éventuelle des mesures agrégées, conséquence l'agrégation peut présenter un biais même si l'estimation est bonne. Les méthodes d'estimation descendantes elles utilisent les mesures agrégées comme seul paramètre pour l'estimation des mesures détaillées. Elles ne se soucient pas alors des corrélations qui existent entre les faits de la base et ainsi, un fait estimé peut être très différent des autres faits connus.

Au delà, des problèmes liés aux informations incomplètes, une autre problématique importante émerge : «la modélisation des mesures à différentes granularités», c'est-à-dire, la modélisation des mesures qui ne sont pas toujours disponibles aux niveaux les plus fins des dimensions (par exemple, pour un polluant la valeur de pollution peut être renseignée au mois et non au jour) (Bimonte et al., 2014). Il n'y a pas de solution pour ces types de modèle, il faut soit estimer les faits soit les supprimer.

Dans cet article, nous restons dans le contexte spatio-temporel pour proposer une approche qui adapte les méthodes d'estimation existantes pour la prises en comptes des caractéristiques principales des BDM spatiales qui sont les corrélations et la multigranularité. Notre proposition s'articule autour de deux contributions. La première est un prédicat de sélection qui aide les fonctions d'estimation à prendre en compte seulement les données sémantiquement utiles en utilisant l'organisation hiérarchique des données dans les BDM. La seconde est une méthode d'ajustement des valeurs estimées aux mesures agrégées connues.

Cet article est organisé comme suit : la section 2 présente les travaux existants, la section 3 motive brièvement notre proposition, la section 4 présente quelques définitions préalables, la section 5 présente l'approche d'estimation, la section 6 présente notre implémentation. Enfin, les conclusions et les travaux futurs sont abordés dans la section 7.

2 Etat de l'art

La présence de valeurs manquantes dans les bases de données est un problème ancien qui s'est toujours posé lors de l'exploitation de données réelles (Rubin, 1976 ; Rioult, 2005). Ce problème est abordé dans de nombreux domaines tels que la statistique, les bases de données, la fouille de données etc., et de nombreuses solutions sont proposées (Wohlrab L. et Furnkranz J., 2011 ; Eekhout I, et al., 2012). Notons que l'estimation ou la prédiction occupe une part importante dans ces solutions. Dans le contexte des BDM, plusieurs travaux visent à réutiliser les solutions existantes dans d'autres domaines. Shoshani (1997) compare les BDM aux modèles

1. L'interpolation est une méthode inductive qui consiste à estimer la valeur en un point x_0 à partir des observations aux points $x_1; \ldots; x_n$

de données statistiques. Les travaux, tels que Abdelbaki et al. (2012), Rabasèda et al. (2011) s'orientent vers le couplage des cubes de données OLAP à la fouille de données, afin de tirer avantage des méthodes éprouvées de prédiction disponibles. Dans la proposition de Xintao et al. (2002), les modèles logistiques et log linéaires sont combinés pour estimer les valeurs manquantes à partir des exemples connus. Les travaux de Ahmed et Miquel (2005), Ahmed et al. (2009) s'inscrivent dans le contexte spatio-temporel et utilisent les fonctions d'interpolation spatiales et temporelles pour l'estimation des faits manquants. Ces méthodes de prédiction sont inductives, car partant des exemples connus pour prédire les faits manquants : ce sont les méthodes d'estimation horizontales. Parce que issues pour la plupart d'autres domaines, elles ne tiennent pas fondamentalement compte de la présence des mesures agrégées stockées garant du temps de réponse réduit lors des analyses. Et en intégrant les valeurs estimées, l'agrégation peut présenter un biais par rapport aux valeurs connues d'avance.

Une autre famille des travaux part plutôt des mesures agrégées stockées pour prédire/estimer les faits détaillés manquants : ce sont les méthodes d'estimation descendantes. Dans ce registre, Camossi et al. (2006) proposent deux fonctions d'estimation descendantes ou fonctions de raffinement (restr et split). Dans Palpanas et al. (2005), les auteurs estiment les faits détaillés à partir des mesures agrégées en utilisant le principe de l'entropie maximale et un algorithmique itératif d'ajustement proportionnel. Dans Xintao et al (2002b) les auteurs transforment le problème d'estimation des faits détaillés à partir des mesures agrégées en un système linéaire, et proposent de résoudre le système par la méthode de décomposition selon les valeurs singulières (SVD). En effet, l'approche descendante n'est pas adaptée dans le contexte des données géographiques marqué par une forte corrélation des données. Car une valeur prédite par cette approche peut être trop différente des autres valeurs connues de la base.

Si beaucoup de méthodes horizontales sont bonnes pour les BDM spatiales, aucune au meilleur de notre connaissance ne prend en compte la présence éventuelle des mesures agrégées qui est une caractéristique essentielle des BDM dans le processus d'estimation. Et en intégrant les valeurs estimées, elles entraînent des incohérences sur l'agrégation par rapport aux données stockées et connues d'avance.

3 Motivations

Dans cette section, nous illustrons avec un exemple simple l'idée générale de notre approche. Considérons l'entrepôt dont le schéma est donné à la figure 1.

Ce schéma est constitué de trois dimensions : une dimension spatiale (Magasins) représentant les données géographique des ventes avec les niveaux ville, région et pays ; une dimension temporelle classique et une dimension thématique représentant les produits.

Un exemple de données factuelles représentant l'instanciation de la figure 1 avec les données aux niveaux les plus fins des dimensions (trimestre, produit, ville) est présenté au tableau 1. Certaines mesures ne sont pas disponibles au niveau ville mais seulement au niveau région de la dimension spatiale, par exemple le montant de vente du produit Milk dans la ville de Yakima est manquante alors que le montant total de vente est présent pour l'État de Californie.

Supposons que le décideur estime la vente du «Milk» de la ville de Yakima à l'aide d'une fonction horizontale (estimateur moyenne) et obtient 15,66. Si cette valeur est considérée, la

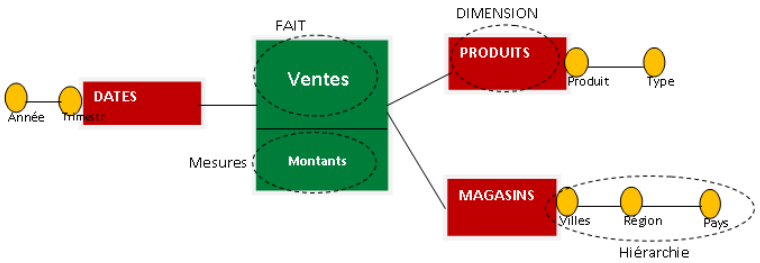

FIG. 1 – *Modèle multidimensionnel correspondant aux ventes des produits.*

somme ne sera plus 105 mais 109,66 ce qui est incohérent. Alors, il est important de prendre en compte les mésures agrégées prédéfinie dans le processus d'estimation.

Type produit
Valeurs agrégées(sommes)
Région
Magasins
Valeurs agrégées(sommes)

Trimestre 1				
CA	Alcoholic Beverages	Beverages	Milk	AGG1
Bellingham	0	5	2	7
Bremerton	18	52	19	89
Seattle			22	162
Spokane		39	28	82
Tacoma	55	75	20	150
Walla	3	7	3	13
Yakima	16	27		54
AGG2	139	313	105	557

TAB. 1 – *Instance des faits détaillés et agrégés (les cellules du cube correspondent à la mesure Montants).*

4 Prérequis

Dans ce qui suit, nous présentons les définitions des notions abordées dans ce papier et relatives aux entrepôts de données multi-granulaires incomplètes. Les quatres premières définitions montrent les concepts de base et parmi ces concepts de base, les deux premiers reprennent certaines notations définies dans Giacometti et al. (2008) et Negre (2009).

Cubes et dimensions : Un cube de dimension N est un tuple $C = \langle D_1, \ldots, D_N, F \rangle$ où :

- Pour $i \in [1, N]$, D_i est une dimension de schéma $sch(D_i) = \{L_i^0, \ldots, L_i^{d_i}\}$. Pour chaque dimension $i \in [1, N]$, chaque attribut L_i^j décrit un niveau d'une hiérarchie, j étant la profondeur de ce niveau.
- F est une table de faits de schéma $sch(F) = \{L_1^0, \ldots, L_N^0, m\}$, où m est l'attribut de mesure.

Exemple : On considère la figure 1
$C = \{PRODUITS, MAGASIN, TEMPS, VENTES\}$, alors
- $sch(PRODUITS) = \{Produit,\ Tous\ les\ produits\}$
- $sch(MAGASINS) = \{Ville,\ Region,\ Pays,\ Tous\ les\ magasins\}$
- $sch(TEMPS) = \{Trimestre,\ Annee,\ All\}$
- $sch(VENTES) = \{Produit,\ Ville,\ Trimestre,\ Montants\}$

Soit $dom()$ une fonction qui associe à tout attribut son domaine de définition. Notons $dom(D_i) = \cup_{j=0}^{d_i} dom(L_i^j), i \in [1, N]$, l'ensemble de tous les membres de la dimension D_i.

Référence de cellule (référence ou fait). Soit C un cube de dimensionn N, une référence de cellule est un N-uplet $\langle r_1, \ldots, r_N \rangle$ où $r_i \in dom(D_i)$ pour tout $i \in [1, N]$. Nous notons $ref(C)$ l'ensemble de toutes les références de C.

Exemple : On considère le tableau 1. $r = (Bremerton,\ Milk,\ Trimestre\ 1)$.

Mesure d'une référence.

Soit C un cube de donnée et $r \in ref(C)$, alors $mes(r)$ est la valeur numérique associée à r. Dans la suite, nous assimilerons fait à référence.

Exemple : On considère le tableau 1. $mes(Bremerton,\ Milk,\ Trimestre\ 1) = 19$.

Connaissant la définition de la mesure d'un fait, nous pouvons définir un fait manquant.

Fait manquant.

Soit C un cube et $r \in ref(C)$ un fait. Le fait r est manquant si sa mesure est manquante. Autrement dit $mes(r) = \varnothing$.

Exemple : On considère le tableau 1 et soit $r = (Yakima,\ Milk,\ Trimestre\ 1)$. Alors, $mes(Yakima,\ Milk,\ Trimestre\ 1) = \varnothing$. Donc r est un fait manquant.

Ces concepts de base vont permettre de définir les notions telles que «granularité d'un fait», «fait manquant» et autres. Mais avant, définissons d'abord le concept «granularité d'un membre».

Granularité d'un membre. Soit un cube C et l'une de ses dimensions D_i, la granularité $granu$ d'un membre $m1 \in dom(D_i)$ est la distance (nombre de niveaux) qui sépare ce membre du niveau Bottom (niveau de granularité le plus bas) de la hiérarchie.

Exemple : On considère la hiérarchie engendrée par la dimension $Magasins$ de la figure 1 et le tableau 1. Alors, $granu(Bremerton) = 0$.

Nous proposons d'additionner la granularité des membres afin d'obtenir la granularité d'une référence.

Granularité d'une référence. Soit un cube C de dimension N et soit $r = (r_1, \ldots, r_n)$ une référence telle que $r_i \in dom(D_i)$ et $i \in [1, N]$, la granularité $Granu$ de r est :

$$Granu(r) = \sum_{i=1}^{N} granu(r_i)$$

Exemple : On considère le tableau 1, et soit *r1=(Bremerton, Milk, Trimestre 1)*.
La granularité de $r1$ est :

$$\begin{aligned} Granu(r1) &= granu(Bremerton) + granu(Milk) \\ &\quad + granu(Trimestre\ 1) \\ &= 0 + 1 + 0 \\ &= 1 \end{aligned}$$

A partir de la granularité des faits, on peut définir les concepts «fait détaillé» et «fait agrégé».

Fait détaillé (respectivement fait agrégé).

Soit C un cube et $r \in ref(C)$ un fait. Le fait r est détaillé (respectivement agrégé) si seulement si $Granu(r) = 0$ (respectivement $Granu(r) > 0$).

Exemple : On considère le tableau 1 et soient
$r1 = (Bremerton,\ Milk,\ Trimestre\ 1)$ et $r2 = (CA,\ Milk,\ Trimestre\ 1)$.
Alors, $granu(Bremerton,\ Milk,\ Trimestre\ 1) = 0$ et
$granu(CA,\ Milk,\ Trimestre\ 1) = 1$. Donc $r1$ est un fait détaillé et $r2$ est un fait agrégé.

A partir des définitions ci-dessus, nous pouvons alors définir les notions d'ensembles telles que : «ensemble de faits détaillés manquants», «ensemble de faits détaillés complets ou connus» et «ensemble de faits agrégés».

Ensemble de faits détaillés manquants.

Soit C un cube, l'ensemble des faits détaillés manquants $EFmissing$ est
$EFmissing = \cup r_i\ ; r_i \in ref(C)\ ;\ Ganu(r_i) = 0\ et\ mes(r_i) = \varnothing.$

Ensemble de faits détaillés complets ou connus.

Soit C un cube, l'ensemble des faits détaillés complets $EFcomplete$ est :
$EFcomplete = \cup r_i\ ; r_i \in ref(C)\ ;\ Ganu(r_i) = 0\ et\ mes(r_i) \neq \varnothing.$

Ensemble de faits agrégés.

Soit C un cube, l'ensemble des faits agrégés $EMagregate$ est :
$EMagregate = \cup r_i; r_i \in ref(C); Ganu(r_i) > 0$

Dans la suite de cet article, nous utiliserons les notations C pour désigner un cube de données, $EFmissing$ pour désigner l'ensemble des faits détaillés manquants, $EFcomplete$ pour l'ensemble des faits détaillés connus, $EMagregate$ pour l'ensemble des faits agrégés. La mesure sera souvent assimilée à un fait et vice-versa.

Les notions de distance suivantes nous permettront de définir les prédicats de sélection de faits similaires au sens organisation structurelle des BDM.

Distance entre membres basée sur la granularité

Soit un cube C et une de ces dimensions D_i, la distance $d_{gmembers}$ entre deux membres $m1$ et $m2 \in dom(D_i)$ d'une même hiérarchie est la valeur absolue de la différence de granularité des deux membres. Plus formellement,

$$d_{gmembers}(m1, m2) = |granu(m1) - granu(m2)|$$

Afin d'obtenir la distance entre références, nous proposons d'additionner les distances entre les membres $d_{gmembers}$.
N.B : Le cas des hiérarchies multiples n'est pas pris en compte.

Distance entre références basée sur la granularité

Soit un cube C de dimension N et soit $r1 = (r_1^1, \ldots, r_1^n)$ et $r2 = (r_2^1, \ldots, r_2^n)$ deux références telles que : $r_j^i \in dom(D_i)$, pour $i \in [1, N]$ et $j \in [1, 2]$, la distance entre deux références au sens distance de granularité est : $d_{granu}(r1, r2) = \sum_{(i=1)}^{N} d_{gmembers}(r_1^i, r_2^i)$.

Distance entre références basée sur le plus court chemin

Soit un cube C de dimension N et soit $r1 = \langle r_1^1, \ldots, r_1^n \rangle$ et $r2 = \langle r_2^1, \ldots, r_2^n \rangle$ deux références telles que : $r_j^i \in dom(D_i)$,pour $i \in [1, N]$ et $j \in [1, 2]$, la distance entre deux références au sens plus court chemin est : $d_{sp}(r1, r2) = \sum_{(i=1)}^{N} d_{members}(r_1^i, r_2^i)$, où $d_{members}(r_1^i, r_2^i)$ est la longueur du plus court chemin entre r_1^i et r_2^i (Negre, 2009).

Exemple : On considère le tableau 1 et soient $r_1 = (Yakima,\ Milk,\ Trimestre\ 1)$ et $r_2 = (Yakima,\ Milk,\ 1997)$. Alors, $d_{granu}(r_1, r_2) = 1$ et $d_{sp}(r_1, r_2) = 1$.

5 Approche d'estimation

Dans cette section nous décrivons la démarche proposée pour l'estimation des faits détaillés manquants. La première approche (descendante) est préconisée lorsque seules les mesures agrégées sont connues. La seconde approche (horizontale) est préconisée lorsque certains faits détaillés sont connus. L'ajustement permet d'adapter les valeurs estimées par l'approche horizontale aux mesures agrégées connues.

5.1 Approche d'estimation descendante

Elle exploite uniquement les mesures agrégées pour estimer/prédire les faits détaillés. Avec cette famille de méthode, on distribue les valeurs agrégées aux faits détaillés.

Définition 5.1 *(fonction d'estimation descendante)*
Une fonction $VerticalEstimation$ est une fonction d'estimation descendante si et seulement si elle estime les valeurs de $EFmissing$ uniquement à partir de $EMagregate$. Autrement dit :

$FEstimatedV = VerticalEstimation\ \ (EMagregate, EFmissing)$.

Exemple : On considère le tableau 1. Une simple application de la méthode *split* donne le tableau 2. Rappelons que la méthode *split* fait une distribution «uniforme» des mesures agrégées aux faits détaillés manquants.

Cette approche d'estimation descendante est recommandée lorsque seuls les faits agrégés sont connus. Mais lorsque certains faits détaillés sont connus, nous préconisons plutôt l'approche horizontale d'estimation.

5.2 Approche horizontale d'estimation

Selon cette approche, seuls les faits détaillés connus sont utilisés pour l'estimation des faits manquants.

Définition 5.2 *(fonction d'estimation horizontale)*

Trimestre 1				
CA	Alcoholic Beverages	Beverages	Milk	AGG1
Bellingham	0	5	2	7
Bremerton	18	52	19	**89**
Seattle	***23,5***	***108***	22	**162**
Spokane	***23,5***	39	28	**82**
Tacoma	55	75	20	**150**
Walla	3	7	3	**13**
Yakima	16	27	***11***	**54**
AGG2	**139**	**313**	**105**	**557**

Biaisées

TAB. 2 – *Estimation des valeurs manquantes par la méthode* ***split****.*

Étant donné une instance r_i^1 de $EFmissing$, $hEstimation$ est une fonction d'estimation horizontale si elle estime r_i^1 uniquement à partir de $EFcomplete$. Plus formellement, $\hat{r}_i^1 = hEstimation(r_i^1, EFcomplete)$.

- Si $hEstimation$ est une fonction d'interpolation temporelle alors, seule la dimension temporelle est considérée et on utilise les faits connus relevés à des dates différentes.
- Si $hEstimation$ est une fonction d'interpolation spatiale alors, seule la dimension spatiale est considérée et on utilise les faits connus de même date et spatialement proches.

Les fonctions d'interpolation à notre connaissance ne prennent pas en compte les classes sémantiques des membres d'une dimension et cela peut biaiser le résultat de l'estimation. Nous proposons alors un prédicat permettant de sélectionner les membres sémantiquement utiles dans le processus d'estimation comme cela est fait pour les recommandations des requêtes OLAP.

5.3 Sélection des faits

Cette section exploite les notions de distance définies à la section 4 pour définir un prédicat de sélection des faits.

Références frères. Soit un cube C de dimension N et soient $r1 = (r_1^1, \ldots, r_1^n)$ et $r2 = (r_2^1, \ldots, r_2^n)$ deux références de faits telles que : $r_j^i \in dom(D_i)$, pour $i \in [1, N]$ et $j \in [1, 2]$. $r1$ et $r2$ sont frères si leur distance d_{granu} est égale à 0.

Références frères proches. Les deux références frères $r1$ et $r2$ ci-dessus sont proches si seulement si $d_{sp}(r1, r2) <= k$. $k \in N$ étant la distance maximale de proximité. La valeur de k est choisie au gré de l'utilisateur. Si elle vaut 0, cela signifie que $r1 = r2$. Si elle est infinie alors, on a tous les faits du cube.

Exemple : En considérant les références du dernier exemple de la section 4, on a $d_{granu}(r_1, r_2) = 1$ et $d_{sp}(r_1, r_2) = 1$. Donc r_1 et r_2 ne sont pas proches car $d_{granu} \neq 0$.

On peut ainsi définir la fonction $Closed(ref1, ref1, k); ref1, ref2 \in ref(C)$ et $k \in N$ qui retourne vrai si $ref1$ et $ref2$ sont frères proches. Mieux encore on peut choisir la dimen-

sion de sélection avec $Closed(ref1, ref1, k, D)$ D étant la dimension du cube C concernée par la sélection. L'algorithme 1 (closed_friends_list) retourne la liste des faits connus proches pouvant participer à l'estimation d'un fait manquant.

Entrée : $EFcomplete$ (l'ensemble des faits connus)
m_fact (le fait manquant)
k (la distance maximale)
D (dimension concernée par la sélection)
Sortie : $friends\_list$ //Liste des références proches

```
friends_list = {}
pourTout fait r ∈ EFcomplete faire
    si Closed(r,m_facts,k,D) !=-1 alors
        friends_list = friends_list ⋃ r
    finsi
finPour
Retourner friends_list
```

Algorithme 1 : Closed_friends_list.

Exemple : On considère le tableau 1 et soit $r1$ la combinaison $(Yakima,\ Milk,\ Trimestre\ 1)$.

La mesure de ce fait est manquante. Ainsi,

$$friends\_list{=}Closed\_friends\_list(EFcomplete, r1, 2, Magasins)$$
$$= \{(Bremerton,\ Milk,\ Trimestre\ 1), (Seatle,\ Milk,\ Trimestre\ 1),$$
$$(Spokane,\ Milk,\ Trimestre\ 1), (Tacoma,\ Milk,\ Trimestre\ 1),$$
$$(Walla,\ Milk,\ Trimestre\ 1), (Yakima,\ Milk,\ Trimestre\ 1)\}.$$

En utilisant les données l'estimateur $moyennne$, on obtient $mes(r1) = 15,66$. Un raisonnement analogue sur les autres faits manquants du tableau 1 permet d'obtenir le tableau 3.

Trimestre 1				
CA	Alcoholic Beverages	Beverages	Milk	AGG1
Bellingham	0	5	2	7
Bremerton	18	52	19	**89**
Seattle	***18,4***	***34,16***	22	**74,56**
Spokane	***18,4***	39	28	**85,4**
Tacoma	55	75	20	**150**
Walla	3	7	3	**13**
Yakima	16	27	***15,66***	**58,66**
AGG2	**128,8**	**288,33**	**109,66**	

TAB. 3 – *Estimation des valeurs manquantes par la méthode de la moyenne.*

En effet, les méthodes horizontales estiment les faits manquants sans tenir compte des mesures dérivées éventuellement connues. En conséquence, si on fait un **Roll Up** sur les valeurs estimées, le résultat diffère éventuellement des valeurs connues. Pour remédier à cette situation, nous proposons dans la section suivante une méthode d'ajustement des valeurs estimées aux mesures agrégées prédéfinies.

5.4 Ajustement des valeurs estimées aux mesures agrégées prédéfinies

Dans cette section, nous décrivons la démarche utilisée pour ajuster les valeurs estimées aux mesures agrégées stockées.

L'exemple illustratif concerne le cas du tableau 3 où on constate effectivement que l'estimation a entrainé des biais sur les valeurs agrégées connues (tableau 1).

La démarche d'ajustement est la suivante :

Soit $B = Substract(EMagregate, EFcomplete)$. ($Substract$ est une fonction qui soustrait les valeurs connues aux mesures agrégées).

On définit une fonction φ de $ref(EFmissing) \times ref(EMagregate)$ vers *{0,1}* telle que :

$\varphi(x,b) = \{^{1 \ si \ x \ est \ utilise \ pour \ le \ calcul \ de \ l'agregat \ b}_{0 \ dans \ le \ cas \ contraire}$

$x \in EFmissing$ et $b \in B$

On en déduit que :

$$b_j = Agg(\bigcup_{i=1} \varphi(x_i, b_j)x_i) \tag{1}$$

Où Agg est une fonction d'agrégation.

Soit $r_i^1 \in EFmissing$. En effet, x_i est la mesure théorique de r_i^1 qui permet d'avoir l'effet escompté. En estimant cette valeur avec une méthode horizontale, on obtient $\hat{x}_i$.

Problème 1 : *Soit x la mesure théorique d'un fait manquant et $\hat{x}$ une estimation non nulle de cette valeur à partir d'une méthode horizontale. Le problème d'ajustement des mesures estimées aux valeurs agrégées connues revient à trouver le réel x' tel que $x = x'\hat{x}$.*

Ainsi, $x_i = x'_i\hat{x}_i$ ce qui ramène l'équation 1 à :

$$b_j = Agg(\bigcup_{i=1} \varphi(x_i, b_j)x'_i\hat{x}_i) \tag{2}$$

La linéarisation de l'équation 2 avec la fonction d'agrégation somme permet d'écrire l'équation suivante :

$$\hat{A}X' = B \tag{3}$$

où $\hat{A}$ est une matrice $m \times p$, X' un vecteur $m \times 1$ et B un vecteur $p \times 1$ tels que :

$$\widehat{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{\varphi(x_1, b_1)\hat{x}_1} & \cdots & \boldsymbol{\varphi(x_m, b_1)\hat{x}_m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\varphi(x_1, b_p)\hat{x}_1} & \cdots & \boldsymbol{\varphi(x_m, b_p)\hat{x}_m} \end{pmatrix}, \ \boldsymbol{X'} = \begin{pmatrix} \boldsymbol{x'_1} \\ \vdots \\ \boldsymbol{x'_m} \end{pmatrix} \text{ et } \boldsymbol{B} = \begin{pmatrix} \boldsymbol{b_1} \\ \vdots \\ \boldsymbol{b_p} \end{pmatrix}$$

Nous proposons une démarche de modélisation semblable à celle de Xintao et al. (2002b) à la différence que les valeurs de notre matrice d'entrée sont issues d'une estimation alors que

les siennes sont booléennes, de plus la résolution de son système est la phase terminale, alors que dans notre cas c'est une phase de préconditionnement dont le résultat permet d'ajuster nos valeurs. L'énumération ci-après décrit les étapes de notre méthode d'ajustement :

- *identifier les mesures agrégées dont certains faits détaillés sont manquants ;*
- *calculer la valeur agrégée $\hat{s}_i$, en ignorant les faits manquants (s_i est la valeur agrégée prédéfinie) ;*
- *calculer $b_i = s_i - \hat{s}_i$ où b_i est la mesure agrégée concernant uniquement les cellules vides ;*
- *pour toute mesure manquante x_i, trouver une estimation $\hat{x}_i$ à partir d'une méthode horizontale ;*
- *définir la matrice $\hat{A}$ en parcourant la liste des valeurs estimées tel que $a_{ij} = \varphi(x_i, b_j)\hat{x}_i$;*
- *résoudre l'équation $\hat{A}X' = B$, où B est le vecteur des b_i*
- *calculer x_i par : $x_i = \hat{x}_i \times x'_i$.*

En se basant sur l'exemple illustratif et en suivant la démarche ci-dessus jusqu'à la cinquième ligne, nous obtenons l'équation suivante :

$$\begin{pmatrix} 18{,}4 & 34{,}16 & 0 & 0 \\ 0 & 0 & 18{,}4 & 0 \\ 0 & 0 & 0 & 15{,}66 \\ 0 & 0 & 0 & 15{,}66 \\ 0 & 34{,}16 & 0 & 0 \\ 18{,}4 & 0 & 18{,}4 & 0 \end{pmatrix} \begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{pmatrix} = \begin{pmatrix} 140 \\ 15 \\ 11 \\ 11 \\ 108 \\ 47 \end{pmatrix} \tag{4}$$

La première composante de l'équation 4 représente la matrice $\hat{A}$, la deuxième représente le vecteur d'ajustement des valeurs X', et la dernière le vecteur B.

Une simple résolution de l'equation 4 permet d'obtenir le résultat suivant :

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{pmatrix} = \begin{pmatrix} 1{,}73 \\ 3{,}16 \\ 0{,}81 \\ 0{,}70 \end{pmatrix} \tag{5}$$

En exécutant la dernière étape (ligne 7) de la démarche ci-dessus on obtient le tableau 4.

Trimestre 1				
CA	Alcoholic Beverages	Beverages	Milk	AGG1
Bellingham	0	5	2	7
Bremerton	18	52	19	**89**
Seattle	***32***	***108***	22	**162**
Spokane	***15***	39	28	**82**
Tacoma	55	75	20	**150**
Walla	3	7	3	**13**
Yakima	16	27	***11***	**54**
AGG2	**139**	**313**	**105**	557

TAB. 4 – *Valeurs estimées ajustées.*

Entrée : $EFmissing$ //l'ensemble des faits détaillés manquant,
$EFcomplete$ //l'ensemble des faits détaillés complets,
$EMagregate$ //l'ensemble des mesures agrégées
Sortie : $FEstimated$ //l'ensemble des faits estimés

si *pas de fait détaillé connu* **alors**
  $FEstimated$ = $VerticalEstimation(EMagregate, EFmissing)$
finsi
sinon
  // Estimation horizontale
  pourTout *fait* $r \in EFmissing$ **faire**
    si *mes(r) est manquante* **alors**
      $friends\_list = Closed\_friends\_list(EFcomplete, r, k, ..)$
      Estimer $\hat{r}$ *en utilisant friends_list à l'aide de l'interpolation spatiale ou temporelle*
      $FEstimatedH = FEstimateH \bigcup \hat{r}$
    finsi
  finPour
  Remplacer les mesures manquantes par celles estimées
  Verifier si en agrégeant on retrouve $EMagregate$
  si *ce n'est pas le cas* **alors**
    Calculer $B = Substract(EMagregate, EFcomplete)$
    $FEstimated$ = *Ajustement de* $FEstimateH$ *à* B
  finsi
  sinon
    $FEstimated$ = $FEstimateH$
  finsi
finsi
Retourner $FEstimated$

Algorithme 2 : Pseudo algorithme de notre démarche.

5.5 Algorithme

Dans cette section, nous décrivons succinctement notre approche d'estimation (cf. algorithme 2).

Pour estimer les faits détaillés manquants à mesures agrégées connues, on vérifie d'abord l'existence des autres faits détaillés, s'ils n'existent pas du tout alors on applique une méthode descendante. Dans le cas contraire on utilise les méthodes d'interpolation spatiale ou temporelle. Puisque les mesures dérivées sont connues d'avance, il faut vérifier que les valeurs estimées respectent cette contrainte. Si ce n'est pas le cas, il faut les niveler ou les réajuster.

Dans le pseudo algorithme 2, on remarque bien que les fonctions d'estimations sont génériques. Puisqu'il existe une panoplie de méthodes d'estimation dans la littérature. On suppose que l'utilisateur en fonction du domaine étudié choisira sa propre méthode.

6 Implémentation

Notre implémentation s'est inscrite dans le cadre d'un processus d'implémentation d'un entrepôt de données.

Nous avons créé trois sources de données génériques. La première étant constituée des faits au plus fin niveau de granularité, et contenant les valeurs manquantes à estimer. Les deux autres sont essentiellement constituées des mesures agrégées. Ces dernières ont été précalculées et stockées dans les tables.

Le processus ETL a été développé grâce à l'outil Talend Open Studio [2]. Il était question à ce niveau d'extraire les données de nos sources génériques, de les normaliser, et puis d'estimer les faits manquants afin de charger dans l'entrepôt de données. Pour cela, nous avons intégré notre approche sous forme de routines Java dans Talend, afin d'utiliser notre algorithme pour l'estimation des faits manquants durant l'intégration des données.

Pour l'implémentation de l'entrepôt de données nous avons choisi le schéma en étoile (Kimball, 1996). Le schéma en étoile est constitué d'une table de faits et de tables de dimensions. Nous avons utilisé Mondrian comme serveur OLAP et JPivot comme client OLAP. Comme SGBDR nous avons utilisé PostreSQL et PostGIS pour le support de la composante spatiale.

La validation préliminaire de notre implémentation s'est effectuée en utilisant les données FoodMart. Pour le test nous avons choisi l'algorithme split défini dans Camossi et al. (2006) comme méthode verticale et la méthode d'interpolation temporelle (*MeanOverTime*) comme méthode horizontale. Nous utilisons la moyenne des RMSEs (racine carrée de la moyenne des carrés des erreurs) comme critère d'évaluation du système. Le résultat de l'estimation avec 33 % de valeurs manquantes est donné au tableau 5.

Les valeurs des RMSEs pour les faits détaillés estimés sont élevées pour ces trois approches, cela pourrait se justifier par le pourcentage elevé des valeurs manquantes. Nous remarquons que cette valeur est beaucoup plus élevée avec l'approche verticale. L'approche horizontale présente de meilleurs résultats, mais elle entraîne un biais élevé sur les valeurs agrégées. L'ajustement corrige considérablement le biais sur les valeurs agrégées mais augmente le biais sur les faits détaillés estimés. En considérant les biais sur faits détaillés et mesures agré-

2. http ://fr.talend.com/products/open-studio-di.php

RMSE	faits détaillés	faits agrégés
Approche verticale	0.39	0
Approche horizontale	0.11	0.13
Approche horizontale ajustée	0.16	0.064

TAB. 5 – *Analyse de performance.*

gées, l'approche horizontale ajustée est une des meilleures solution pour l'estimation des faits manquants. Nous mènerons davantage les expérimentations afin de valider ces hypothèses.

N.B : Les prédicats de sélections n'ont pas été utilisés lors des tests.

7 Conclusion

Dans cet article, nous proposons une nouvelle approche d'estimation pour les entrepôts de données multi-granulaires incomplètes. Cette approche adapte les fonctions d'estimation existantes pour la prise en compte des caractéristiques essentielles des BDM. Un prédicat de sélection des faits sémantiquement utiles au processus d'estimation est proposée avec un algorithme pour ajuster éventuellement les valeurs estimées aux valeurs agrégées prédéfinies.

Pour nos futurs travaux, nous prévoyons de tester cette démarche sur plusieurs algorithmes d'estimation, et ensuite de mener des expériences en incluant les prédicats de sélections sur les données réelles afin de valider la démarche quant aux approches classiques d'estimation des valeurs manquantes.

Références

Abdelbaki W., Sadok B. Y., et Messaoud R. B. (2012). Une approche connexionniste pour l'extension de l'olap à des capacités de prédiction. EDA, volume B-8 of RNTI, page 72-81. Hermann.

Ahmed T. et Miquel M. (2005). Multidimensional Structures Dedicated to Continuous Spatio-temporal Phenomena. In : JACKSON et al. 22th British National Conference on Databases, Sunderland, UK. Berlin Heidelberg : Springer, 29-40 p. LNCS 3567.

Bimonte S. (2007). *Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation*. Thèse de doctorat, Institut National des Sciences Appliquées de Lyon, Informatique et Information pour la Société.

Bimonte S., Boulil K., Pradel M., André G. et Chanet J.P. (2004). Analyse des indicateurs énergétiques des entreprises agricoles : une approche Spatial OLAP. *Revue internationale de géomatique*.

Camossi E., Bertolotto M. et Bertino E. (2006). A multigranular object-oriented framework supporting spatio-temporal granularity conversions. *International Journal of Geographical Information Science*, 20(5) :511-534.

Dyreson C. (1997). Using an incomplete data cube as a summary data sieve. *Bulletin of the IEEE technical committee on data engineering*, 20 :19-26.

Dyreson C. E, Pedersen T. B. et Jensen C. S. (2003). Incomplete information in multidimensional databases, In *Multidimensional Databases*, pages 282-309. Maurizio Rafanelli(ed), Idea Group Publishing.

Eekhout I., Boer M. R.,Twisk Jos W. R., Vet Henrica C. W. et Heymans M. W. (2012). Missing data : a systematic review of how they are reported and handled. *Epidemiology, September ;* 23(5) :729-32.

Giacometti A., Marcel P., et Negre E. (2008). A framework for recommending OLAP querie, In *DOLAP*, pp. 73-80.

Inmon W.H. (1996). Building the Data Warehouse 2nd Ed. John Wiley and sons, New York, NY, ISBN 0764599445.

Kimball R. (1996). The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses. New York, John Wiley & Sons.

Negre E. (2009). *Exploration Collaborative des cubes de données*. Thèse de doctorat, Université François Rabelais Tours.

Palpanas T. et Koudas N. (2005) Using datacube aggregates for approximate querying and deviation detection, *bulletin of the IEEE technical committee on data engineering*, 17(11) :1465-1477.

Rabasèda S. L., Boussaid O., Niemczuk A. B. et Messaoud R. B. (2011). Prédiction dans les cubes de données olap. *Conférence Méditerranénne sur l'Ingénierie Sur des Systèmes Complexes (MISC'11)*, Agadir, Maroc.

Rioult F. (2005). *Extraction de connaissances dans les bases de données comportant des valeurs manquantes ou un grand nombre d'attributs*. Thèse de doctorat, Université de Caen Basse-Normandie (spécialité Informatique).

Rubin D.B. (1976). Inference and Missing Data. *Biometrika, 29, 159-183*.

Wohlrab L. et Furnkranz J. (2011). A review and comparison of strategies for handling missing values in separate-and-conquer rule learning, *J. of Intelligent Information Systems, 36(1) :73-98*.

Xintao W. et Barbara D. (2002). Modeling and imputation of large incomplete multidimensional datasets. *Springer-Verlag Berlin Heidelberg*, pages 286-295.

Xintao W. et Barbara D. (2002b). Learning missing values from summary constraints, *SIGKDD Explorations, Volume 4, Issue 1, page 30*.

Summary

The spatial data warehouses are characterized by a strong correlation data. Therefore, methods of spatial and temporal interpolation are widely used to estimate the missing facts. These methods often ignore the possible presence of aggregate measures. This causes a bias on aggregation. Thus, we propose an approach that adapts existing estimation functions for taking into account the aggregated values in the estimation process.