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