FUSION ET GREFFES DE DONNEES

Gilbert Saporta

Conservatoire National des Arts et Métiers
Chaire de Statistique Appliquée-CEDRIC
saporta@cnam.fr

Nicolas Fischer

EDF Division Recherche et Développement et CNAM-CEDRIC
Nicolas.Fischer@edf.fr

Cet article fait partie des Actes des 5èmes JOURNEES MODULAD qui ont eu lieu les 16 et 17 novembre 2000 à E.D.F. (Clamart).

Résumé :

La fusion statistique de fichiers a pour but de compléter un fichier « receveur » où certaines variables ne sont pas renseignées (questions non posées) à l'aide d'un ou plusieurs fichiers « donneurs » portant sur d'autres individus. Le fichier donneur comprend bien sûr des variables communes ainsi que les variables d'intérêt renseignées pour tous les individus. Les remplacements de données manquantes se font soit par des méthodes d'imputation basées sur des proches voisins (injection) soit à l'aide de méthodes explicites de type régression.

Les greffes d'enquêtes poursuivent des objectifs proches, en ce sens qu'il s'agit par exemple de positionner des résultats d'un sondage (une analyse factorielle) sur ceux d'un autre en utilisant des variables passerelles, mais sans nécessairement chercher à estimer les données manquantes. Cet exposé présentera la problématique, les principales techniques utilisées, les critères de validation, ainsi que les dangers potentiels ....