SÉMINAIRE INED
LES RENCONTRES DE STATISTIQUE APPLIQUÉE

VARIATIONS SUR L'ANALYSE EXPLORATOIRE DE DONNÉES

Séance du vendredi 12 juin 2009 de 14h à 17h30

Avec le développement de l'informatique et la mise à disposition de larges bases de données, le statisticien se tourne de plus en plus vers des méthodes interactives, intuitives et robustes qui lui permettent d'appréhender la singularité ou la diversité de ses données. La statistique exploratoire, très présente aujourd'hui dans le Data Mining largement ancré dans les entreprises, s'est développé au cours des trente dernières années et est aujourd'hui intégrée dans les principaux logiciels de statistique (module Sas-Insight, module Stat-Studio, SPSS, Spad, logiciel R etc.). La démarche exploratoire, qui rehausse le graphique et la robustesse comme éléments centraux d'appréciation d'une analyse statistique, met le statisticien au centre de ses données et lui permet d'en comprendre la structure et la complexité.

Statistique exploratoire et statistique confirmatoire (modélisation) ont souvent été opposées, pourtant ces deux démarches sont complémentaires, c'est cette articulation que nous essaierons notamment de mettre en évidence au cours de cette séance.

RÉSUMÉS DES INTERVENTIONS

14h15 – **La place de John W. TUKEY au sein de l'analyse exploratoire.**

Monique LE GUEN • (CES – Matisse, Université Paris 1)

Notre exposé a pour objectif de mieux faire connaître John Wilder TUKEY (1915-2000), figure emblématique de l'Analyse Exploratoire des Données et du « Data Analysis ».
Après un bref aperçu de la vie de TUKEY (Princeton, Harvard, Stanford, AT&T Bell Laboratories, etc.), nous survolerons ses 60 années de travaux statistiques appliqués à tous les domaines de la connaissance.
Nous montrerons son influence, sur un grand nombre de ses élèves et collaborateurs, dont les développements techniques actuels en visualisation (graphiques), résistance, robustesse, procédures, logiciels, enseignement, permettent d'apprécier son rayonnement et sa clairvoyance.

15h00 – **La statistique visuelle.**

Eugène HORBER • (Université de Genève)

Afin de mettre en évidence le rôle essentiel des outils graphiques et d'une démarche visuelle interactive du chercheur explorateur, cette présentation-démonstration proposera d'examiner un outil particulier, le nuage des points et ses multiples facettes et variantes, nées dans le sillage de l'école de Tukey. L'objectif de ce tour guidé sera de présenter différents outils (matrices de nuages, co-plots,..), d'insister sur l'apport essentiel de l'interactivité (gestes de la souris pour interroger/modifier), voire de l'animation, tout en montrant à travers ces exemples la mise en pratique des principes de l'exploration, principes qui nous mènent vers une pratique dans laquelle la démarche visuelle (re) trouvera une place de choix.

15h50 – Apports de la statistique exploratoire dans le cadre de la régression linéaire.
Josiane CONFAIS • ISUP

Toute étude statistique est une analyse de statistique***s***, et donc c'est une analyse de données ! Cette lapalissade met en évidence le fait qu'il n'est pas possible de bâtir un modèle statistique quel qu'il soit, sans avoir au préalable analysé les données sur lesquelles il s'appliquera. Cette phase préliminaire d'analyse est celle où le statisticien explore les données pour en découvrir les particularités. La statistique exploratoire est donc l'outil privilégié pour débuter l'étude. Mais elle est également utile dans la phase de validation du modèle construit.
L'exposé montrera comment, à partir de données concrètes sur lesquelles la construction d'un modèle de régression linéaire est envisagée, l'analyse exploratoire aide le statisticien dans la définition puis la validation du modèle.

Le logiciel SAS®/Insight illustrera la démarche.

16h20 – Méthodes de classification sur séries temporelles. Application à la prévision.
Dominique LADIRAY • INSEE (Chef du Département Statistiques de Court Terme)

L'analyse des données "à la française" et l'analyse des séries temporelles ont chacune une longue histoire mais curieusement leurs chemins ne se sont que rarement croisés, au moins jusqu'à un passé récent. Dans les vingt dernières années, avec la mise à disposition d'énormes bases de données temporelles, il y a eu une explosion d'intérêt pour l'exploration de ces données. Des centaines d'articles ont alors proposé des méthodes et algorithmes pour classer, indexer, segmenter et discriminer les séries temporelles.

De nombreuses méthodes de classification, mesures de similarité et algorithmes ont été développés au cours des ans, essentiellement pour des données d'enquêtes. Malheureusement, la plupart de ces mesures de similarité ne peuvent être directement utilisées sur des données temporelles. De nouvelles distances et de nouvelles stratégies ont été définies, certaines d'entre elles étant basées sur des outils ou résultats récents de l'analyse des séries temporelles : coefficients cepstrum, transformée par ondelettes, modèles markoviens cachés etc.

La première partie de l'exposé sera consacrée à la présentation des méthodes les plus utilisées en classification de séries temporelles.

La seconde partie de l'exposé montrera, à travers un exemple, comment ces méthodes exploratoires peuvent être utilisées, et avec bonheur, avec des méthodes plus confirmatoires. Une méthode de construction d'un modèle économétrique de prévision basée sur la classification de séries sera ainsi proposée et comparée aux méthodes habituellement utilisées (approche « General to specific » ou modèles factoriels dynamiques) ;

Arnaud Bringé et Bénédicte Garnier

Institut national d'études démographiques • Établissement public scientifique et technologique
133, bd Davout, Paris 20e • Standard : 01 56 06 20 00 • Fax : 01 56 06 21 92
Le Service Méthodes Statistiques de l'Ined organise un séminaire trimestriel se proposant de couvrir un domaine de Statistique Appliquée traitant de l'application d'une méthodologie. Ce séminaire est ouvert à tous, sans frais de participation.
Pour tout renseignement contacter Mme Chafika Mekhazni, secrétaire du service, au 01 56 06 20 91.