

J. P. FENELON

L'école d'été du C.N.R.S. sur l'analyse des données

Mathématiques et sciences humaines, tome 70 (1980), p. 69-75

http://www.numdam.org/item?id=MSH_1980__70__69_0

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1980, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (<http://msh.revues.org/>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

L'ECOLE D'ETE DU C.N.R.S. SUR L'ANALYSE DES DONNEES

J.P. FENELON *

L'objet de l'Ecole ne peut pas se décrire en catégories de programme au sens scolaire du terme. Cependant on peut assurer que, du 1er au 17 septembre 1980, chaque stagiaire aura parcouru les 10 points chauds de l'Analyse des Données dans sa pratique réelle (voir paragraphe I), aura abordé les 8 points théoriques (avec les formulaires et dépouillements impliqués) du paragraphe II, et enfin, sera autonome dans la manipulation d'une série de six types de logiciels mettant en jeu plus de 30 programmes principaux et plus de 150 sous-programmes (voir paragraphe III).

I - LES DIX POINTS CHAUDS DE L'ANALYSE DES DONNEES

Il n'apparaît nulle part, ni dans les ouvrages, ni dans les stages habituels, que la pratique réelle de l'Analyse des Données est une espèce de "jeu de l'Oie" constituée de dix cases entre lesquelles le praticien doit se déplacer en permanence en fonction de ses hypothèses, ses essais, ses résultats, ses réussites partielles, ses succès à confirmer, ses données à valider.

Il s'agit en bref des dix points suivants :

1. Le dossier d'étude, de recherche, doit être élaboré, constitué informatiquement, consulté statistiquement, enrichi de résultats, mis à jour parfois ... en tenant compte de multiples contraintes.

* C.N.R.S. -

C R E D O C, 142, rue du Chevaleret 75013 PARIS Tél. : 584 14 20.

2. Les données aussi bien du type I-lignes que du type J-colonnes, doivent être extraites et gérées informatiquement avec certains statuts définis par la stratégie statistique (filtres, pondérations, codes sans réponses, etc...).
3. Les codages portent tantôt sur des variables, tantôt sur des tableaux. Ils sont de nature diverses (dédoublément, formes disjonctives complètes, bornes d'amplitudes ou d'effectifs égaux, etc...).
4. Les tableaux mis à l'étude, extraits du dossier et souvent fruits d'un recodage, sont de nature très diverses : Contingences, Fréquence, Mesures, Logiques, de BURT, de Préférences, de Régression, de Discrimination, etc...
5. En fonction de la nature du tableau, des mesures de distance, des indices de similitude, des indicateurs de proximités sont à calculer entre individus, entre variables, entre éléments, entre classes ...
6. En liaison avec le point 5, on peut élaborer des représentations algébriques ou géométriques, notamment sous forme de nuages ...
7. On doit mettre en oeuvre des techniques de réduction (facteurs, classes, etc...).
8. On doit organiser des sorties, dépendant de la nature statistique et du support informatique : nuages étiquetés, facteurs documentés, classes qualifiées, hiérarchies indicées, etc... sur des imprimantes, des traceurs, des disques, des écrans, etc...
9. On doit très souvent retourner au tableau étudié pour reconstituer des données manquantes, étudier des éléments illustratifs ou supplémentaires, simuler des résultats différents pour valider des synthèses, etc...
10. On doit presque toujours retourner au dossier pour enrichir des synthèses, effectuer des régressions, rechercher des discriminations, comparer des méthodes, calculer des aides à l'interprétation des résultats, etc...

REMARQUE

En dynamique réelle, chacun de ces points peut renvoyer à n'importe lequel des dix autres. C'est dire que l'analyse des données est plutôt un art de la pratique des cheminements entre ces dix points, plutôt qu'une technique linéaire et figée. Les stagiaires s'en persuaderont : groupés par trois avec un enseignant plein temps, ils auront le loisir de parcourir ces dix points, à la fois sur les "figures imposées" préparées à l'avance par le Comité Scientifique de l'Ecole, et à la fois sur leurs propres données. L'optimum semble être de conserver un équilibre entre ces deux types d'activités : figures imposées et figures libres.

II - LES HUIT POINTS THEORIQUES

On peut regrouper les points théoriques en huit mots clefs : tableaux, nuages, caractères, distances, synthèses géométriques, algébriques, stratégies, validité.

1. Tableaux statistiques : types de tableaux analysables : mesures, probabilités, rangs, logiques, pavés, etc..., informations endogènes et exogènes à un tableau, sens des nuages, élaboration des codages et recodages, etc...
2. Nuages statistiques : espaces multidimensionnels, représentations spatiales des points, des masses, des profils, des lois, etc...
3. Caractères statistiques multidimensionnels : moyennes, centre de gravité, dispersion, inertie, etc...
4. Distances disponibles : euclidiennes, distributionnelles ; indices de distance ; indicateurs de proximités, de corrélation, etc...
5. Synthèses géométriques : propriétés d'inertie, axes principaux, facteurs, formulaires et interprétations, aides au dépouillement des analyses en composantes principales et des correspondances.

6. Synthèses algébriques : classes, arbres, hiérarchies, partitions, aides au dépouillement des classifications et contributions mutuelles entre classifications et analyses factorielles.

7. Stratégies statistiques : descriptions, synthèses, typologies, segmentations, régressions, discriminations, ... autant de mots qui furent longtemps autant de cases étanches entre elles. Aujourd'hui, plus que de méthodes et de programmes différents, il s'agit plutôt d'objectifs différents mettant en oeuvre des outils universels.

8. Significativité et validité : en dehors de quelques cas particuliers, la validité et la significativité en Analyse des Données sont des questions à poser en terme de simulations et de stabilité des formes.

Les huit points ci-dessus ne sont pas abordés comme dans un cours, d'une manière linéaire à progression continue, mais chaque enseignant plein temps veillera à ce que chacun des trois stagiaires de son groupe de travail ait parcouru, à propos des données traitées, chacun de ces huit points.

Cela sera fait indifféremment à propos des données préparées à l'avance ou à propos des données proposées par le stagiaire.

III - LES SIX POINTS INFORMATIQUES DE L'ANALYSE DES DONNEES

L'Analyse des Données utilise l'informatique de plusieurs manières. Chacun des six points suivants donnera lieu à au moins quatre heures de travaux pratiques (et à au plus huit heures) dans chaque groupe de travail.

1. L'utilisation du Centre de Calcul, la saisie et l'organisation des données à traiter.
2. La gestion des données en vue de leur analyse. L'exemple des systèmes OSIRIS, SPSS, BMDP.
3. La statistique descriptive classique : programmes des bibliothèques fortran françaises (A.D.D.A.D. et C.E.S.I.A.), et des systèmes disponibles

durant le stage : OSIRIS, SPSS, BMDP, HARWELL, IMSL.

4. Les programmes d'analyses factorielles disponibles : bibliothèques françaises : ADDAD, CESIA, YAGOLNITZER ; leurs divers types de programmes des composantes principales et des correspondances.
5. Les programmes de classification, qu'il s'agisse des partitions (type nuées dynamiques de DIDAY) ou des hiérarchies (CAH) sont également disponibles avec leurs aides à l'interprétation et au dépouillement.
6. Les procédures de simulation et de validation donnent lieu à des exercices de programmation.

Les cadres de l'Ecole veilleront en permanence à ce que si, dans un groupe donné, tel ou tel de ces six points étaient sous-représentés compte tenu du type des données apportées par tel stagiaire, il soit dupliqué des listages complets d'autres groupes (cartes de commandes, de paramètres, d'appels aux bibliothèques et résultats). Le critère a priori : chaque stagiaire quitte l'école avec une documentation équilibrée entre ces six points.

LE PUBLIC VISE PAR L'ECOLE

On croit trop souvent que l'Analyse des Données "c'est de l'informatique" ; - "c'est de la statistique" - ; - "c'est des mathématiques appliquées". Il s'agit là de "lectures" erronées du concept d'Analyse des Données, lectures qui justifieraient une sélection par niveaux de connaissance.

On croit trop souvent que les techniques de l'Analyse des Données dépendent des matières analysées ; c'est faux : tel exposé d'un éthologue sur un combat d'épinoches fournira à un économiste un modèle pour le traitement de ses données chronologiques. Ce qui laisse loin de nous l'idée d'effectuer une sélection par discipline !

L'Analyse des Données est pour beaucoup un des seuls points de contact entre la réalité et la discipline abordée. C'est dire que l'Analyse des Données concerne aussi bien l'ingénieur que le chercheur, aussi bien le statisticien que l'étudiant ! Elle porte ses fruits aussi bien en économie qu'en musicologie, en psychologie qu'en océanographie, en marketing qu'en écologie !

Alors s'il n'y a pas sélection par niveaux, s'il n'y a pas sélection par discipline, comment s'effectuera la sélection des candidats à cette école ? Uniquement sur ce critère : le candidat possède les données traitables statistiquement et il est bien décidé à acquérir une formation correcte à l'Analyse des Données. Correcte c'est-à-dire éloignée des deux écueils les plus répandus : faire confiance à un spécialiste, donc renoncer à prendre des décisions sur des points qu'il est pourtant dramatique de sous-traiter ; ou bien, fasciné par l'apparente automaticité de l'informatique, utiliser soi-même aveuglément un résultat, jugé fiable parce qu'imprimé par un ordinateur !!

Mais ce critère risque d'être satisfait à la fois par un polytechnicien imbu de statistiques, un technicien hautement compétent mais qui n'a pas étudié de mathématiques depuis son Brevet, par un artiste génial qui n'a jamais mis les pieds dans une école, par un agrégé de lettres pris de nausée face au moindre signe mathématique.

Faire subsister tous ces publics ne peut sembler une gageure qu'à ceux qui confondent l'Analyse des Données avec des domaines qu'elle utilise (mathématique, statistique, informatique), mais qui ne représentent pas dix pour cent de sa substance !!

Les évidentes questions de langage seront résolues par le fait qu'en permanence il y a un enseignant pour trois stagiaires, et l'ambition de traiter un "programme" aussi vaste en quinze jours est accessible puisqu'il y aura essentiellement travaux sur la pratique de l'Analyse des Données (tantôt sur des figures imposées : données des enseignants, tantôt sur des figures libres : données des stagiaires).

Le nombre très limité des places (une vingtaine), la compétence de l'encadrement (rompu aux pratiques informatiques, formé à la statistique, pratiquant l'Analyse des Données dans les domaines les plus divers) le taux

de cet encadrement (trois stagiaires pour un enseignant plein temps), la richesse de l'environnement (stage résidentiel autour d'un ordinateur), le prix des inscriptions (500 F par jour de stage, soit 7 200 F pour les quinze jours du 1er au 17 septembre 1980)...

Autant de points qui nécessitent que la sélection des candidatures s'effectue très tôt pour que se vérifie, lors d'un contact entre les organisateurs et vous-mêmes, que vous êtes bien la personne que nous recherchons pour passer quinze jours avec nous :

1. Vous possédez des données analysables.
2. Vous voulez apprendre à les analyser correctement.

Ce sont nos deux seuls critères de recrutement.

Retournez-nous très vite ce bon par courrier (aucun contact téléphonique).

J.P. FENELON

BON A RETOURNER

J.P. FENELON
CNRS
CREDOC :
142, rue du Chevaleret
75013 PARIS

NOM : _____

Prénom : _____

Adresse (lettres capitales) : _____

☐ Désire participer à votre prochaine réunion d'information sur l'Ecole d'Eté sur l'Analyse des Données, avant de poser sa candidature.

☐ s'estime assez informé pour demander un rendez-vous personnel sur ses données et poser sa candidature.