Qu'est–ce que l'Analyse des Données ?

Jean-Paul Benzécri

Barcelona, June 2003

La dernière lettre reçue de Pierre Bourdieu me posait une question trop difficile : c'était, ou presque : qu'est–ce que l'Analyse des Données ? Je lui fis une réponse élusive.

Maintenant que mon ami n'est plus de ce monde, je lui dois de rassembler mes esprits, sinon pour lui répondre, au moins pour attester que je n'ai pas de réponse à donner qui me satisfasse. . .

In the last letter I received from Pierre Bourdieu, he asked me too difficult a question, that went like this, or almost : "What is Data Analysis" ? I answered in an elusive way.

Now that my friend is no longer of this world, I owe it to him to collect my wits, if not to answer him, at least to testify that I cannot provide an answer that satisfies me. . .

Quand, sur le projet de la Traduction Automatique des Langues Naturelles, linguistique, logique et mathématique entreprirent de collaborer en ayant l'ordinateur pour outil. . . , il apparut que, dans la voie frayée par Louis Guttman† et Chikio Hayashi†, le principe d'équivalence distributionnelle, proposé par le linguiste Zelig Harris†, devait régler l'analyse des données statistiques.

When linguistics, logic and mathematics set out to collaborate on a project of Automatic Translation of Natural Languages using the computer as a tool, it became clear, that, following the path opened by Louis Guttman† and Chikio Hayashi†, the principle of distributional equivalence proposed by the linguist Zelig Harris† was meant to rule the statistical data analysis.

Alors en donnant forme géométrique à cette analyse, on aboutirait à la recherche des axes principaux d'inertie d'une nuage de points munis de masses ; problème classique en dimension 3, mais à traiter ici en une dimension n quelconque. Ce qui requiert impérativement des diagonalisations de matrices carrées $n \times n$, calcul infaisable sans une machine, dès que n dépasse 3 (ou 4...).

Thus, giving a geometric form to this analysis would lead to researching the principal axes of inertia of a weighted cloud ; this is a classical problem in 3–dimensional space but to be treated here in any dimension n. This absolutely requires the computing of eigenvalues and vectors of squared matrices $n \times n$, a calculation impossible to make without a machine as soon as n exceeds 3 (or 4...).

(De nos jours), il faut quelques minutes pour les algorithmes de Classification et d'Analyse factorielle...
Mais la conception des données, leur mise en forme, l'examen des résultats prennent non seulement des heures, mais des mois...

(Today) this takes only a few minutes for Classification and Factor Analysis algorithms...
But the conception of data, their organization, the examination of results, all that requires not only hours, but months...

Il n'y a plus à strictement parler de problème de calcul ; mais le problème même de l'Analyse des Données subsiste, d'autant plus vaste que, le calcul ne mettant point de borne à la recherche, on n'a point d'excuse pour s'arrêter dans la collecte des données et la méditation.

Relativement à 1960. . . , le rapport de difficulté entre projets intellectuels et calculs est inversé.

Strictly speaking there are no longer any calculation problems, but the problem of analysis is still with us, and has become all the bigger as, since calculation no longer sets a limit to research, there is no excuse to stop collecting data and meditating.

Compared to 1960. . . the balance between intellectual projects and calculations has been reversed.

Il s'en faut de beaucoup que les principes qui nous paraisssent s'imposer soient admis de tous.
La distinction entre qualitatif et quantitatif ne nous semble pas toujours être bien comprise. En bref, *il ne faut pas dire* :

There are principles that for us are imperative and these principles are far from being recognized by all.
The distinction qualitative vs quantitative does not appear to be always well understood.
In short, *one should not say* that :

grandeur numérique continue	$\approx$	donnée quantitative ;
grandeur à un nombre fini de modalités	$\approx$	donnée qualitative.
continuous numerical magnitude	$\approx$	quantitative data ;
finite number of modalities magnitude	$\approx$	qualitative data.

Car au niveau de l'individu statistique (e.g., le dossier d'un malade), une donnée numérique — l'âge ou même la pression artérielle ou la glycémie — n'est généralement pas à prendre avec toute sa précision, mais selon sa signification ; et, de ce point de vue, il n'y a pas de différence de nature entre âge et pression.

Indeed, at the level of a statistical individual (e.g. a patient's file), a numerical data — age or even blood pressure or sugar level — is not to be taken as a rule with its full accuracy but according to its meaningfulness ; and from this point of view, there is no difference in nature between age and pressure.

Et surtout, pour comparer un individu à un autre, il faut considérer, non deux ensembles de données primaires, par exemple deux ensembles de cent nombres réels, un point de $\mathbb{R}^{100}$, à un autre point de $\mathbb{R}^{100}$, entre lesquels des ressemblances globales ne se voient pas, mais la synthèse de ces ensembles, aboutissant à quelques gradations, ou à des discontinuités, à des diagnostics...

Above all, in order to compare an individual to another, one should not consider two sets of primary data (e.g. two sets of 100 real numbers or one point of $\mathbb{R}^{100}$ to another point of $\mathbb{R}^{100}$) between which overall similarities cannot be observed, but the synthesis of these sets so as to reach some gradations or discontinuities and diagnoses.

Il ne faut pas trop vite affirmer, méprisamment, que la pensée ne peut que devenir paresseuse quand l'outil devient plus puissant . . .
le problème pratique provoque (ou, du moins, aiguillonne) le développement des idées théoriques.

One should not too quickly assert, contemptuously, that thinking can only become lazy when tools become more powerful. . .
practical problems trigger (or at least, spur) the development of theoretical ideas.

Quant aux problèmes de l'avenir...
analyse des images ou des sons de la musique et de la parole...
j'ai passé des mois à me régaler, sur un MacPlus, avec un petit logiciel "soundcap" sur la parole ; il en est résulté un article

Now as far the future is concerned...
image analysis or music and speech sound analysis,
I have spent months enjoying myself on a MacPlus computer, playing with small software "soundcap" about speech ; a paper came out of this.

[SPECTRE. STAT. VOIX.], *Cahiers d'Analyse des Données*, Vol. XIII, 1988, n°1, pp. 99-130.

Ce qu'il faut (ce à quoi je me targue d'avoir quelque peu réussi)... c'est voir ce qui, dans les objets étudiés, en l'espèce des sons, est pertinent. Voilà ce dont on doit d'abord s'enquérir dans tout corpus de dimension et de complexité "astronomiques".

What is mandatory (and for which I claim to have been successful)... is to see what among the studied objects (here sounds) is relevant. This is what should first be looked for in any corpus of "astronomical" dimension and complexity.

Le statisticien doit être modeste. . . le travail de ma génération a été exaltant. . .
une nouvelle analyse est à inventer, maintenant que l'on a, et parfois à bas prix, des moyens de calcul dont on ne rêvait même pas il y a trente ans. . .

The statistician should be modest. The work of my generation has been exhilarating. . .
A new analysis remains to be invented, now that we have at our disposal, sometimes at low cost, means of calculation we did not even dream about thirty years ago.

A la mémoire de Pierre Bourdieu...
Moi qui dois tant à sa familière et indulgente amitié, et suis pénétré de respect pour l'exemple que nous laisse la leçon, faite le 27 Mars 2001, terme de son enseignement au Collège de France.

In memory of Pierre Bourdieu...
I who owe him so much to his congenial and understanding friendship, and in great respect for his last lesson (March 27^{th}, 2001) at the Collège de France.

Jean–Paul Benzécri

June 2003