REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Ébauche d'une théorie statistique des agrégats

Revue de statistique appliquée, tome 8, nº 3 (1960), p. 107-111
<http://www.numdam.org/item?id=RSA_1960__8_3_107_0>

ÉBAUCHE D'UNE THÉORIE STATISTIQUE DES AGREGATS

P. THIONET
Administrateur à l'Institut National de la
Statistique et des Études Économiques

I - INTRODUCTION -

D'un usage courant en comptabilité Nationale le mot "agrégat" n'est jamais employé en théorie statistique. D'autre part les "problèmes d'agrégation" sont absents des traités et manuels de statistique, mais sont étudiés en économétrie. Cette division du travail est un état de fait, plus ou moins rationnel.

La méthodologie statistique traite des moyennes. Si une population, d'effectif N, a relativement à la variable x, une moyenne $\overline{x}$, l'agrégat est $N\overline{x}$ c'est-à-dire Σ x. On s'intéressera soit à $\overline{x}$, soit à Σ x, suivant le domaine d'applications considéré.

Considérons à présent des variables ou caractères x y z .... prenant sur chaque unité statistique (i) les valeurs x_i, y_i, z_i, ..., leurs agrégats pour l'ensemble de la population seront désignés par X, Y, Z, ... Il est utile de savoir si Y, Z,... ne peuvent pas s'exprimer, au moyen de X comme variable et, sinon toujours sous forme de fonctions :

$$Y = f(X) \quad , \qquad Z = g(X), \; \ldots .$$

du moins avec des variables aléatoires $\eta\zeta$ qu'on négligera en première approximation pour une population suffisamment nombreuse,

$$Y = f(X) + \eta \; , \qquad Z = g(X) + \zeta \; , \ldots$$

étant entendu que f, g, ... ne dépendent que de $X = \Sigma x$ et non des autres caractéristiques de la distribution des x_i. Ceci constitue un problème d'agrégation en statistique.

En supposant que les (i) sont des ménages de revenus x_i, on peut voir dans y_i, z_i, ... certains groupes de dépenses des ménages, quand on suppose constants les prix p, q, ... C'est alors un problème d'économétrie ; il a été traité par Gorman (Econometrica - Janvier 1953) et Nataf (Thèse 1953) quand on s'abstient d'utiliser des aléatoires η, ζ ..

Nous nous proposons de traiter directement le problème en statistique. Nous ferons l'économie de la théorie des choix, des surfaces d'indifférence, des courbes d'Engel, etc., et nous suivrons un procédé purement géométrique. Il s'agit d'un essai pour rapprocher mathématiciens, statisticiens, économistes (1).

----

(1) Le problème a été traité aussi sous l'angle de l'économétrie par R. G. D. Allen, Mathematical Economics, chapitre 20 (The aggregation problem) - Londres 1957.

II - PROBLEMES D'AGREGATION : LES BESOINS -

Empruntons à l'économétrie les notions de microrelation et de macrorelation et distinguons les relations suivant que le temps t entre ou non en ligne de compte. Les conditions réelles peuvent être les suivantes :

II. 1 - Ou bien le statisticien possède les séries statistiques (disons annuelles) des agrégats :

$$X_1\, X_2 \ldots.\; X_t \ldots.\; X_n$$
$$Y_1\, Y_2 \ldots.\; Y_t \ldots.\; Y_n$$
$$Z_1\, Z_2 \ldots.\; Z_t \ldots.\; Z_n \ldots$$

et il constate que les points (X_t Y_t Z_t) ont tendance à se placer sur une certaine courbe simple ; le nombre n de ces points n'est jamais très grand en pratique. Par ailleurs les données individuelles (x_{it}, y_{it}, z_{it}, ..) font défaut ; et c'est à partir des observations sur les agrégats qu'on se fait une idée du comportement de l'individu, c'est-à-dire de relations qui pourraient exister entre x_i y_i z_i ... Quelles hypothèses raisonnables peut-on faire ?

II. 2 - Ou bien le statisticien possède une (ou plusieurs) coupes à un instant t (ce qu'on appelle cross-section en Amérique) c'est-à-dire, sinon un "recensement" de tous les (x_{it} y_{it} z_{it}) pour t fixé et invariant, mais du moins un "sondage" c'est-à-dire un échantillon des (i) en principe nombreux.

Et alors on veut se faire une idée de l'évolution future de Y, Z, ... moyennant une hypothèse sur l'évolution de X et à supposer que les comportements individuels soient assez stables.

II. 3 - Ou bien le statisticien dispose à la fois des séries statistiques passées sur les agrégats et de cross-sections, les séries étant longues et les coupes bien représentatives ; auquel cas il n'est pas impossible de concevoir qu'on parvienne à une bonne compréhension des mécanismes et des prévisions valables. Cependant on risque encore d'avoir des données incomplètes, douteuses, incohérentes.

Le Statisticien aura une tâche ardue d'ajustements des données et d'estimation des paramètres, mais on ne s'en préoccupe pas ici. Ce qu'on lui demande d'abord, c'est d'adopter un ensemble cohérent de micro et de macro-relations, de relations temporelles et hors du temps, de liaisons stochastiques et fonctionnelles. Pour commencer, tenons-nous en aux liaisons fonctionnelles.

III - UN RESULTAT SIMPLE -

Supposons l'existence d'une micro-relation $y_i = f(x_i)$. Il est banal qu'on fasse l'erreur de croire les agrégats $X = \Sigma x_i$, $Y = \Sigma y_i$ liés par la macrorelation $Y = f(X)$; (L. Klein l'a signalé à propos des lois à élasticité constante), et c'est inexact, à moins que f ne désigne une fonction linéaire. Mais en outre :

Enoncé : La relation entre X et Y est une fonction si et seulement si la relation entre x et y est linéaire ; dans le cas général, Y est une fonctionnelle de la distribution des x et ne dépend pas exclusivement de Σx_i (ce qui empêche de faire comme prévu en II. 1 et II. 2 si la distribution se déforme).

Ce résultat important s'établit en appliquant la théorie statistique des moyennes.

Les points de coordonnées $[x_i, y_i = f(x_i)]$ sont répartis sur la courbe $y = f(x)$ dont on va admettre qu'elle est sans inflexion. Le centre de gravité de ces points de coordonnées :

$$\bar{X} = \Sigma x_i/N \quad , \qquad \bar{Y} = \Sigma f(x_i)/N \quad ,$$

se trouve forcément situé dans la concavité de la courbe, ce qui entraîne :

$$\text{si } f''(x) > 0 \quad : \quad \bar{Y} > f(\bar{X}) \text{ ou } Y > Nf(\bar{X})$$
$$\text{si } f''(x) < 0 \quad : \quad \bar{Y} < f(\bar{X}) \text{ ou } Y < Nf(\bar{X})$$

Admettons que Y soit fonction de X ; Y devrait être invariable quand on modifie la distribution des x_i à l'intérieur de $X = \Sigma x_i$ (en supposant l'effectif N donné et fixe), en particulier Y ne devrait pas changer si l'on remplaçait tous les x_i par leur moyenne $\bar{X}$, ce qui impliquerait :

$$Y = Nf(\bar{X})$$

On peut compléter la démonstration en supposant $[y = f(x)]$ munie de points d'inflexion, mais en modifiant la distribution des x_i une seconde fois, de façon à rassembler les points sur un arc de courbe démuni d'inflexions ; ceci suppose pourtant $f''(\bar{X}) \neq 0$, restriction sans grande portée.

Que signifie exactement le théorème ? Il invite à préférer les hypothèses de structure linéaires. Toutefois : La fonction $f(x_i)$ ne doit pas dépendre du temps t ; celui-ci modifie seulement x_i. Si par hasard le temps ne modifiait pas Σx_i tout en changeant individuellement certains x_i (ou tous les x_i) l'agrégat Y ne serait pas modifié non plus (les changements survenus aux y_i se compensant) à plus de 2 dimensions :

Enoncé : Si l'on suppose $y_i = f(x_i)$, $z_i = g(x_i)$, etc. les agrégats Y, Z, ... sont fonctions de l'agrégat $X = \Sigma x_i$ sans dépendre autrement de la distribution des x_i, si et seulement si les fonctions f, g, etc. sont linéaires ; il est supposé que f, g, etc. ne dépendent pas directement de t.

Extension à des microrelations individualisées : Supposons à présent que les microrelations s'écrivent $y_i = f_i(x_i)$, avec possibilité d'avoir moins de f_i que d'unités de sondage N (cas des groupes sociaux homogènes). Les comportements sont supposés stables, l'unité (i) ne changera pas de courbe au cours du temps. Nous disons encore que :

Enoncé : Les agrégats Y, Z, ... sont fonctions de l'agrégat $X = \Sigma x_i$ (sans dépendre autrement de la distribution des x_i) si et seulement si les fonctions f_i g_i .... sont linéaires et de la forme :

$$f_i = a_i + \alpha x_i \quad ; \quad g_i = b_i + \beta x_i \; ; \; \ldots$$

En effet : Y ne doit pas changer si je bouleverse les x_i en conservant Σx_i ; donc je vais remplacer x_1 par $x_1 + h$ et x_2 par $x_2 - h$, ce qui donne, quel que soit h :

$$f_1(x_1 + h) - f_1(x_1) = - [f_2(x_2 - h) - f_2(x_2)]$$

Supposons f_1 f_2 dérivables, il en résulte que :

$$f_1'(x_1 + h) = f_2'(x_2 - h) = f_3'(x_3 - h) = \ldots.$$

quelles que soient les variables entre parenthèses ; c'est donc une constante α. Le reste en résulte.

IV - CAS DE L'ECONOMETRIE -

Dans un espace cartésien de coordonnées (y_i, z_i...) les microrelations avec x_i comme paramètre sont figurées par une droite (ou plutôt une demi-droite, $x_i\, y_i\, z_i$.... étant positifs).

Lorsqu'on change les microrelations, cette demi-droite change.

Ceci est très exactement le résultat de Gorman et Nataf, dont les courbes d'Engel sont rectilignes, mais se modifient quand le vecteur prix est modifié. De même :

Les droites d'Engel des divers consommateurs ont une direction commune ($\alpha\beta$ ...)

La différence essentielle serait, en définitive, que les achats d'un consommateur ne sont pas strictement indépendants ; celui-ci "boucle son budget", par une relation (où figurent les prix p. q)

$$py + qz + \ldots = x$$

alors que nous supposons y, z, ... fonctions de x seulement.

V - CAS DE LIAISONS STOCHASTIQUES -

Si l'on imagine des microrelations de la forme $y_i = f_i(x_i) + \eta_i$, $z_i = g_i(x_i) + \zeta_i$ où le point aléatoire ($\eta\zeta$ ...) suit une loi de probabilité possédant des moments d'ordre 1 nuls et des moments d'ordre 2 finis, les agrégats ($\Sigma\eta$, $\Sigma\zeta$) seront plus ou moins négligeables à côté de Σf, Σg ; et le cas de $f_i = a_i + \alpha x_i$, etc. continue à présenter un certain intérêt. On n'abordera d'ailleurs pas ici l'étude de l'article de Theil (Econometrica 1958).

VI - AGREGATS ET COURBE DE CONCENTRATION DE GINI. DIVERS -

Simple remarque finale, les "tranches" des distributions statistiques sont larges, et les praticiens aiment avoir non seulement la ventilation par tranche des effectifs mais aussi des agrégats (nombre d'exploitations de plus de 100 ha et superficie correspondante). La représentation d'une distribution par une courbe de Gini (effectifs cumulés en abscisses, agrégats cumulés en ordonnées) informe alors bien mieux qu'un histogramme, et est donc à inclure dans la statistique des agrégats.

On notera que cette même courbe de concentration est commode pour étudier certains problèmes théoriques, comme cette agrégation ou mélange des populations statistiques de manière à obtenir une population de la même famille étudiée par nous même en 1955-56 (Publ. de l'Institut de Statistique de l'Université de Paris VI, 1957).

Disons un mot de l'agrégation de microrelations $y = f(x)$ pour une population stratifiée suivant des tranches de valeurs de x. Du § III ci-dessus découle qu'on doit postuler qu'à l'intérieur de chaque tranche dy/dx (et dY/dX) a la valeur constante α, qui peut changer d'une tranche à l'autre, - si bien entendu on n'a pas d'information sur la distribution des x à l'intérieur de chaque tranche. Remarquons que les frontières de strate se modifient en même temps, les unités statistiques étant censées ne pas quitter leur strate.

Pseudo-agrégation - La pseudo-agrégation est une exception apparente aux règles précédentes. Elle consiste à formuler sur la distribution des x_i une hypothèse de structure telle qu'une "certaine" agrégation devienne possible. Supposer $Y = F(X)$, c'est imposer à la distribution des x_i la condition $\Sigma f_i(x_i) - N\, F\,(\Sigma x_i/N) = 0$; un cas particulier intéressant est celui où f et F ont la même forme analytique (et géométrique), et diffèrent par quelque paramètre.

Exemple : $y_i = x_i^2/a$, $Y = X^2/b$ si le coefficient de variation γ des x_i reste constant.

On a en effet : $b = a\, N(1 + \gamma^2)^{-1}$

<u>Variation des microrelations f avec le temps t</u> : Tout ce qui précède suppose t absent de f. Admettons au contraire $y_i = y_i^\circ f(t, x_i)$. A quelle condition a-t-on $Y = Y^\circ f(t, Z)$ avec $Z = Z(x_1 \ldots\ldots x_n)$. Une telle agrégation implique $f(t, x_i) \equiv g(t) + h(t) K(x_i)$, du moment qu'on admet f dérivable.

On voit en même temps que $K(Z) = \Sigma y_i^\circ K(x_i)/Y^\circ$.

En particulier on aura $Z = X$ si $y_i = y_i^\circ g(t) + x_i h(t)$.

A ce propos on commet souvent l'erreur d'attribuer aux y_i des taux de croissance diversifiés, $y_i = y_i^\circ e^{x_i t}$ et on veut trouver $Y = Y^\circ e^{Zt}$, ce qui est impossible quel que soit Z. D'ailleurs en pratique si les taux x_i sont voisins de 0 et la projection à 4 ou 5 ans, la formule agrégée n'est pas incorrecte avec $Z = X/N$.