MATHÉMATIQUES ET SCIENCES HUMAINES

S. RÉGNIER

À propos de l'analyse de dépendance

Mathématiques et sciences humaines, tome 25 (1969), p. 5-11
<http://www.numdam.org/item?id=MSH_1969__25__5_0>

M.S.H. (7e année, n° 25, 1969, p. 5-11)

A PROPOS DE L'ANALYSE DE DÉPENDANCE *

par

S. RÉGNIER * *

1. — INTRODUCTION.

L'objet de l'analyse de dépendance est de mesurer l'intensité des relations de *Causalité* qui peuvent intervenir dans une structure complexe, entre des grandeurs sociologiques observables dans différentes populations. Une telle structure apparaît par exemple, dans la célèbre étude du suicide, par Durkheim, la variable x_3 = taux de suicide est influencée par :

x_1 = âge,
x_2 = état-civil,

mais x_2 (mariage ou célibat) est elle-même influencée par x_1, ce qui peut se représenter par un graphe orienté[1] :

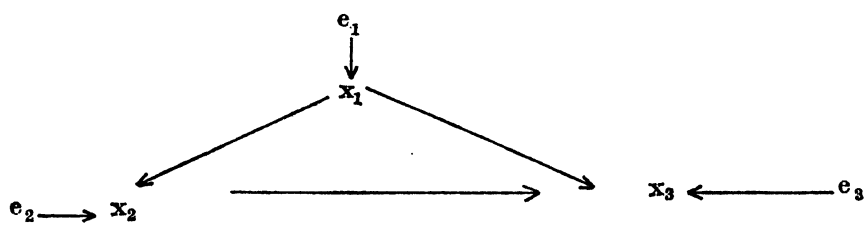

Fig. 1

On peut chercher à traduire ce graphe par un système de relations linéaires, mais il est alors indispensable d'ajouter des facteurs externes : il est évident que x_3 n'est pas *seulement* une fonction de x_1 et x_2. On écrira :

$$x_1 = e_1$$
$$x_2 = e_2 + a_2^1 x_1$$
$$x_3 = e_3 + a_3^1 x_1 + a_3^2 x_2$$

(où e_1 désigne l'action du milieu ambiant spécifiquement sur x_1, e_2 sur x_2, etc.).

Le problème est de déterminer les coefficients a_i^j alors que seules les grandeurs x_i sont mesurées. Nous exposerons d'abord la solution proposée par R. Boudon, avec le cheminement suivi par cet auteur, en explicitant plus complètement les démonstrations pour aboutir finalement à un point de vue un peu différent en 6 et 7.

---

* D'après R. Boudon : *L'Analyse mathématique des faits sociaux*, Plon, 1967, 30, 1, p. 88-134.

* * Centre de Mathématique appliquée et de Calcul de la Maison des Sciences de l'Homme.

1. L'âge influence l'état-civil, et non l'inverse. Cette « évidence » doit être acceptée dans la règle du jeu. Dès lors les dépendances considérées sont complètement dissymétriques à la différence des notions probabilistes ou statistiques usuelles de coefficients de corrélation, de régression, etc.

2. — MODÈLE MATHÉMATIQUE.

Pour nous exprimer en termes plus généraux, nous allons considérer un système de n variables réelles observées dans N populations :

$$\left\{ \begin{array}{ll} x_i^s & i = 1, 2 \dots n \\ & s = 1, 2 \dots N \end{array} \right.$$

soit N vecteurs $x^s = (x_1^s, x_2^s \dots x_n^s)$.

Pour alléger, on peut considérer x comme une grandeur statistique, comme un vecteur aléatoire, et *omettre l'indice s*. Nous nous donnons de plus un graphe de dépendance, c'est-à-dire un système de flèches liant certains couples de variables $x_i : \leftarrow x_j$; chaque flèche représente une relation de dépendance hypothétique, il s'agit de valuer ce graphe à l'aide d'une matrice carrée A. On devra avoir :

$$x = Ax + e, \text{ c'est-à-dire}$$
$$x_i = \Sigma\, a_i^j\, x_j + e_i$$

a_i^j ne pourra être $\neq 0$ que s'il existe une flèche de dépendance de x_j vers x_i, dans le cas contraire, l'absence de flèche de x_j vers x_i se traduit par la contrainte $a_i^j = 0$, que nous appellerons dans la suite un *zéro structurel* ; e_i représente l'influence des facteurs extérieurs au modèle, non mesurés par hypothèse.

Posons : $B = I - A$, nous aurons :

$$Bx = e \qquad (1)$$
$$Bxx' = e\,x' \text{ [1]}$$

$$E\,(B\,x\,x') = B\,E\,(x\,x') = E\,(e\,x') \qquad (2)$$

Nous supposerons les variables x centrées, (2) est une relation entre les covariances.

Sans hypothèse supplémentaire, nous ne pouvons pas déterminer A (et B). On peut trouver des matrices P telles que P B possède les mêmes zéros que B.

Alors $P\,B\,x = P\,e$, et P B constitue une autre solution du problème. Le choix de P est cependant limité par l'existence des *zéros structurels* dans la matrice B.

Si P comporte un terme non nul p_i^j, en dehors de la diagonale ($i \neq j$), dans P B, la ligne i est remplacée par une combinaison où intervient la ligne j de B, cette dernière doit donc présenter au moins les mêmes zéros (moins de termes non nuls).

3. — CONDITION POUR L'IDENTIFICATION.

Si nous imposons aux facteurs externes d'avoir deux à deux une corrélation nulle, et au graphe de dépendance d'être *transitif* et *sans cycles*, nous allons voir que la matrice moyenne : E (e x') présente, elle aussi, des zéros structurels :

$$b_i^j \neq 0 \Rightarrow b_j^i = 0 \Rightarrow E\,(e_i\, x_j) = 0.$$

et qu'alors une seule matrice B vérifie à la fois (1) et (2) en respectant à la fois les zéros structurels de B et ceux de E (e x').

Commençons par le dernier point · l'absence de cycle de longueur 1 (boucles) entraîne $a_i^i = 0$ et $b_i^i = 1$.

1. X' désigne le transposé du vecteur X, E (xx') désigne l'espérance mathématique de la matrice xx', soit ici simplement la moyenne arithmétique des N matrices indexées par *s*.

Si deux matrices B et PB vérifient (1) et (2), et si P est diagonale, $P = I$.

Si P n'est pas diagonale et qu'un terme p_i^j est non nul avec $i \neq j$, on a vu que la ligne j de B doit présenter au moins les mêmes zéros structurels que la ligne i. Et dès lors...

Puisque :

$$b_j^j = 1 \text{ (dans la ligne j)}$$
$$b_i^j \neq 0.$$

De même dans la matrice E (e x'), le terme diagonal E (e_j x_j) est non nul (c'est la covariance de x_j avec le facteur externe qui agit sur x_j et sur lui seul). Ce même raisonnement montre donc que dans la ligne i : $E(e_i x_j) \neq 0$, ce qui contredit que : $b_i^j \neq 0$.

Revenons maintenant au point délicat. Nous allons montrer que :

$$E(e_i x_j) \neq 0 \Rightarrow b_j^i \neq 0 \Rightarrow b_i^j = 0.$$

La deuxième implication exprime l'absence de cycle de longueur 2.

La première exprime qu'il n'y a corrélation entre un facteur explicite x_j et un facteur externe e_i que si x_j dépend du facteur explicite en relation directe avec e_i, c'est-à-dire x_i.

En effet :

$$e_i x_j = e_i (e_j + \Sigma a_j^k x_k)$$
$$E(e_i x_j) = E(e_i e_j) + \Sigma_k a_j^k E(e_i x_k)$$

Le premier terme est nul par hypothèse pour $i \neq j$.

Sous le signe somme, on doit isoler le terme $a_j^i E(e_i x_i)$, car la corrélation de e_i et x_i est toujours positive.

L'hypothèse : $E(e_i x_j) \neq 0$ implique donc :

$$\begin{cases} \text{ou bien } a_j^i \neq 0, \\ \text{ou bien } \exists k \neq i \quad a_j^k \neq 0 \text{ et } E(e_i x_k) \neq 0, \end{cases}$$

mais dans ce second cas, $E(e_i x_k) \neq 0$ implique encore :

$$\begin{cases} \text{ou bien } a_k^i \neq 0, \\ \text{ou bien } \exists l \neq i,\ a_k^l \neq 0 \text{ et } E(e_i x_l) \neq 0. \end{cases}$$

La chaîne complète des implications est infinie et peut s'écrire : $E(e_j x_j) \neq 0 \Rightarrow$

$$\begin{cases} a_j^i \neq 0, \\ \text{ou} \\ \exists k \neq i,\ a_j^k \neq 0 \quad \text{et} \begin{cases} a_k^i \neq 0, \\ \text{ou} \\ \exists l \neq i \quad a_k^l \neq \text{et} \begin{cases} \text{etc.} \end{cases} \end{cases} \end{cases}$$

$$\text{et} \begin{cases} a_m^i \neq 0, \\ \text{ou} \\ \exists n \neq i, \quad a_n^m \neq 0 \quad \text{et} \begin{cases} a_n^i \neq 0, \\ \text{ou} \\ \text{etc.} \end{cases} \end{cases}$$

Nous voyons qu'il existe une chaîne de variables de longueur finie ou non : $x_k\ x_l$ etc. $x_m\ x_n$ etc., telles que dans le graphe :

$$x_j \leftarrow x_k \leftarrow x_e \text{ etc.} \quad \leftarrow x_m \leftarrow x_n \text{ etc.}$$

et que si la chaîne est finie, elle se termine par la variable x_i, par exemple : $x_n \leftarrow x_i$. avec $a_n^i \neq 0$.

Mais le nombre de variables x_k étant fini, la chaîne ne peut être infinie sans présenter des cycles (en passant plusieurs fois par les mêmes variables).

Dans l'hypothèse où le graphe est sans cycle, nous devons conclure que :

$$E\,(e_i\,x_j) \neq 0 \Rightarrow$$

x_j dépend de x_i à travers une chaîne de variables : x_k, x_e, etc. x_m, x_n.

Si maintenant le graphe est transitif, il en résulte que x_j dépend aussi directement de x_i selon le schéma : $x_j \leftarrow x_i$, que par conséquent b_j^i n'est pas un zéro structurel, et que par contre b_i^j en est un.

4. — MÉTHODE DE CALCUL DES COEFFICIENTS b_i^j.

A tout coefficient b_i^j, qui n'est pas un zéro structurel, nous pouvons associer l'équation :

$$E\,(e_i\,x_j) = 0 = \sum_k b_i^k\, E\,(x_j\,x_k) \qquad (3)$$

qui concerne toute la ligne i de B.

Ces équations vont-elles admettre une solution unique ?

Considérons le sous-système d'équation relatif à la ligne i de B :

$$\begin{cases} \sum_k b_i^k\, E\,(x_j\,x_k) = 0, \\ \forall j \text{ tel que } b_i^j \neq 0 \text{ structurel} \end{cases} \qquad (4)$$

(c'est-à-dire tel que $x_i \leftarrow x_j$ dans le graphe).

Nous avons p équations à p inconnues : p est le nombre de flèches reçues par x_i. On peut supposer qu'elles proviennent de x_1, x_2, ..., x_p. Comme $b_i^i = 1$, nous pouvons isoler les termes $E\,(x_j\,x_i)$ comme second membre. Sous forme matricielle, et en mettant en facteur l'opérateur E d'espérance mathématique, le système (4) va s'écrire :

$$E \begin{bmatrix} x_1\,x_1, & x_1\,x_2 & \text{etc.} & x_1\,x_p \\ x_2\,x_1, & x_2\,x_2 & \text{etc.} & x_2\,x_p \\ & & & \\ x_p\,x_1, & x_p\,x_2 & \text{etc.} & x_p\,x_p \end{bmatrix} \begin{bmatrix} b_i^1 \\ b_i^2 \\ \\ b_i^p \end{bmatrix} = -E \begin{bmatrix} x_1\,x_i \\ x_2\,x_i \\ \\ x_p\,x_i \end{bmatrix}$$

Soit en condensant :

$$E\,(x\,x')\,B_i = -E\,(x\,x_i)$$

x désigne ici le vecteur des variables dont dépend x_i. Le déterminant du système ne peut être nul que si ces variables x_1, x_2, x_p sont liées par une relation linéaire stricte. Dans ce cas, l'une de ces variables s'exprime comme une combinaison linéaire stricte des $p - 1$ autres et devrait être retirée du modèle. Moyennant cette convention, *nécessaire également dans les études de régression*, nous voyons que le système (4) admet toujours une solution unique.

5. — PROPRIÉTÉ FONDAMENTALE.

Les équations :

$$\begin{cases} E\,(e_i\,x_j) = 0, \\ \forall j \text{ tel que } b_i^j \neq 0, \end{cases} \qquad (4)$$

expriment que la variance : $Q_i = E\,(e_i^2)$ est extremum.

En effet :

$$\frac{1}{2}\frac{\delta\,Q_i}{\delta\,b_i^j} = E\left(e_i\,\frac{\delta\,e_i}{\delta\,b_i^j}\right) = E\,(e_i\,x_j)$$

car :

$$e_i = \Sigma\,b_i^k\,x_k.$$

Considérons alors un graphe particulièrement simple où seule la variable x_n dépend des précédentes :

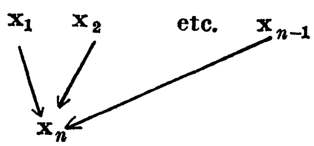

Fig. 2

Le système d'équation correspondant :

$$\begin{cases} x_1 = e_1 \\ x_2 = e_2 \\ \text{etc.} \\ x_n = e_n + a_n^1\,x_1 + a_n^2\,x_2 + \text{etc.} + a_n^{n-1}\,x_{n-1} \end{cases}$$

aura pour solution les coefficients : $a_n^i = -\,b_n^i$ qui rendent minima la variance de l'écart :

$$e_n = x_n - \Sigma\,a_n^i\,x_i.$$

Les coefflcients a_n^i seront alors les coefficients de régression classique.

Nous voyons ainsi que l'analyse de dépendance est une généralisation de l'analyse de régression multiple.

6. — RÉSUMÉ CRITIQUE.

Dans les cinq paragraphes précédents, j'ai tenté de suivre fidèlement l'exposé de Boudon, en précisant seulement les démonstrations et leurs conditions de validité. Il apparaît :

a) que la démonstration de l'énoncé « quand les corrélations entre facteurs implicites sont nulles, les coefficients de dépendance sont toujours identifiables », suppose le graphe sans cycle, et transitif ;

b) que le procédé proposé pour le calcul des coefficients de dépendance n'est univoque qu'au prix d'une convention, qui élimine du graphe tout facteur interne qui serait strictement une combinaison linéaire de certains autres.

Dès lors, il nous paraît difficile de conserver l'articulation de l'exposé de Boudon construit sur la notion d'identification. Les coefficients de dépendance seraient, selon lui, identifiables dans certaines conditions où les coefficients de régression ne le sont pas. C'est le contraire qui nous semble vrai. Les coefficients de régression sont uniques à partir de la seule convention *b*). Les coefficients de dépendance de Boudon demandent en plus, pour l'être, que le graphe des flèches de dépendance soit transitif sans cycle (condition a $\Rightarrow$ graphe d'une relation d'ordre) et que les facteurs externes soient sans corrélations mutuelles. Quand ils existent, les coefficients a_i^j du modèle sont les coefficients de régression de la variable interne x_i par rapport aux variables x_j dont elle reçoit des flèches (dont x_i dépend selon le graphe de dépendance).

CONCLUSION.

Nous proposons, devant un graphe de dépendance tel que celui de Durkheim, de le valuer à l'aide des coefficients de régression, en traitant tour à tour chaque variable comme un effet et comme une cause. Est-ce à dire que la distinction entre coefficients de régression et coefficients de dépendance soit vide de contenu ? La réponse doit être nuancée. On savait déjà les coefficients de régression porteurs de deux significations principales, ils donnent l'espérance mathématique conditionnelle de y/x_i $(i \in I)$, soit pour une loi normale de dimension $I + 1$, soit pour une certaine famille de lois unidimensionnelles dépendant des paramètres x_i. Ils serviront ici à mesurer l'intensité d'une influence causale, notion complètement dissymétrique. Le graphe pourra être non transitif et comporter des cycles : les valuations de deux flèches opposées seront deux coefficients complètement distincts comme dans l'exemple ci-après :

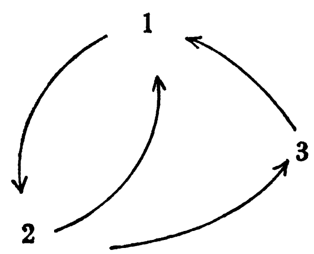

Fig. 3

$$x_1 = a_1^2\, x_2 + a_1^3\, x_3 + e_1$$

$$x_2 = a_2^1\, x_1 + e_2$$

$$x_3 = a_3^2\, x_2.$$

Nous devons cependant rejeter les graphes avec boucles ; une équation telle que :

$$x_1 = a_1^1\, x_1 + a_1^2\, x + e_1$$

conduirait à la solution triviale :

$$a_1^1 = 1 \quad a_1^2 = 0 \quad \text{variance } e_1 = 0$$

Pour conclure, il apparaît que les coefficients de régression constituent une solution possible au problème de l'analyse causale, et qu'on peut les nommer dans ce cas coefficients de dépendance.

Nota. — *A titre de contre exemple* concernant l'énoncé *a*), p. 31, voici un graphe non transitif, d'ailleurs sans cycle, où le coefficient b_2^1 n'est pas identifiable au sens de Boudon :

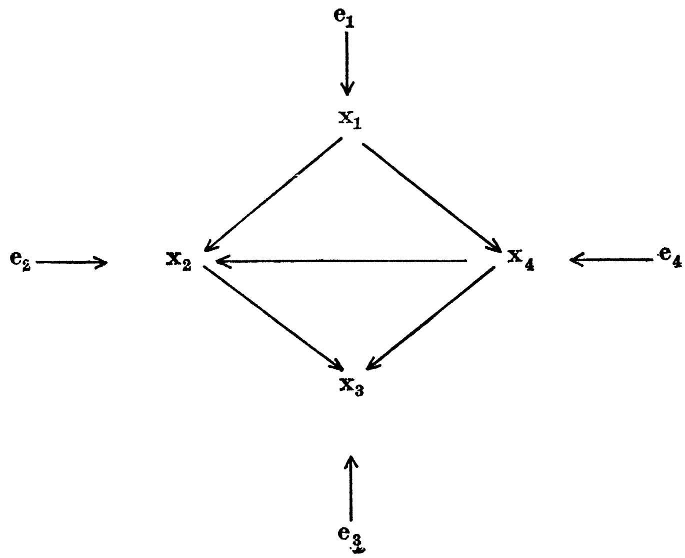

Fig. 4