JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MARIE BERRONDO-AGRELL
FRANÇOIS-XAVIER PACTEAU
STÉPHANE ROUY

Iphigénie, un procédé de typologie pour spécialistes et non-spécialistes

Journal de la société statistique de Paris, tome 134, n° 3 (1993), p. 41-48

<http://www.numdam.org/item?id=JSFS_1993__134_3_41_0>

IPHIGÉNIE, UN PROCÉDÉ DE TYPOLOGIE POUR SPÉCIALISTES ET NON-SPÉCIALISTES

par Marie BERRONDO-AGRELL
maître de conférence à l'Université Paris-Evry
(I.U.T., département Transport-Logistique)

par François-Xavier PACTEAU et Stéphane ROUY
étudiants à l'École Centrale de Paris

Résumé

Créer une partition dans un ensemble de données, est un des grands problèmes statistiques. La méthode correspondante peut être jugée selon les 3 critères suivants :

1. *Clarté et simplicité,* afin d'être explicable à des non-spécialistes et facile d'emploi pour les spécialistes.

2. *Flexibilité,* afin de pouvoir s'appliquer dans des cas très divers.

3. *Richesse en information,* afin de donner, comme nous le verrons plus loin, en moyenne, un nombre de groupes proche du nombre d'éléments par groupe.

Au vu de ces 3 critères, IPHIGÉNIE semble une méthode intéressante, une sorte de référence parmi les très nombreuses méthodes existantes.

Summary

How to create a partition in a set of data is one of the main statistical problems. The corresponding method may be judged with the 3 following criterias :

1. *Simplicity,* to be explained to non-specialists and easy to use by the specialists.

2. *Flexibility,* to be applied in very different cases.

3. *Quantity of information,* giving as an average a small difference between the number of groups and the number of elements per group, as we shall see further in this text.

Regarding theses 3 criterias, IPHIGENIA seems an interesting method, a kind of reference among all the other existing methods.

1. Présentation

Créer une partition dans un ensemble de données est un des grands problèmes statistiques. Il existe pour cela de nombreux procédés, souvent très sophistiqués, mais possibles néanmoins grâce à l'informatique. Nous avons cherché ici à ce que l'ordi-

nateur ne traduise que la sagesse intuitive, sans préciser à l'avance le nombre de parties à obtenir, et cela à partir des cas les plus divers. C'est ainsi qu'est née la méthode IPHIGÉNIE, en 1972. Un travail récent sur cette méthode de typologie, effectué à l'Ecole Centrale de Paris, vient de mettre en évidence une propriété surprenante du point de vue de la richesse de l'information (cf. 4 et 5) : IPHIGÉNIE aboutit en moyenne à une partition très proche de la « partition médiane », équidistante des 2 partitions extrêmes que forment respectivement le regroupement complet et la séparation totale. C'est ce que nous nous proposons de présenter ici en commençant par un rappel du procédé lui-même.

2. Le procédé

Soit E un ensemble contenant n éléments $\{x_1, x_2, ..., x_n\}$ que nous souhaitons classer. Nous connaissons sur ces n éléments des informations diverses qui nous permettent d'établir un indice d'écart entre 2 éléments quelconques de E.

Rappelons ici qu'un indice d'écart d[1] est une application de $(E \times E)$ dans $\mathbb{R}^+$ vérifiant les 2 propriétés suivantes (3) :

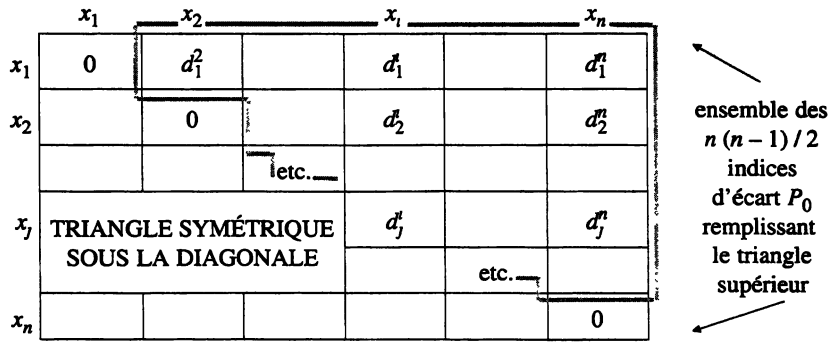
1) $(\forall x \in E)\,\big(d(x, x) = 0\big)$.

2) $(\forall x \in E)\,(\forall y \in E)\,\big(d(x, y) = d(y, x)\big)$.

Remarque

Un indice d'écart est plus souple qu'une distance. En effet, il peut prendre des valeurs nulles pour des couples d'éléments distincts de E (non-réciprocité du premier axiome). D'autre part, l'inégalité triangulaire n'a pas à être respectée (pas de troisième axiome).

Voici le tableau correspondant :



1. la lettre d a été choisie pour symboliser un indice d'écart à cause de l'idée de distance vulgarisée.

Dans le triangle supérieur, apparaissent $n\ (n\ 1)\ /\ 2$ valeurs formant l'ensemble P_0. Le bon sens nous dicte les 4 opérations suivantes :

1) On cherche le minimum de P_0.

2) On regroupe les éléments de E correspondants.

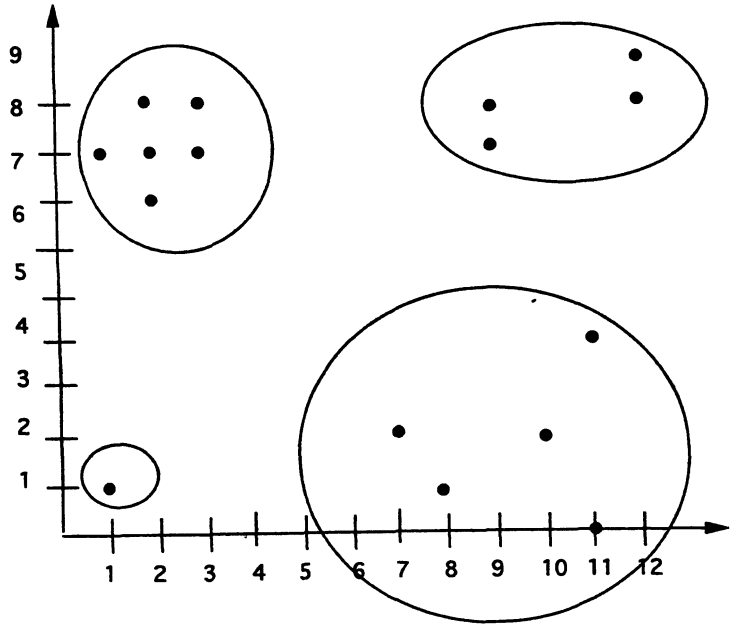
3) On cherche le maximum de P_0.

4) On sépare les éléments correspondants de E.

On élimine alors ces 2 valeurs extrêmes. Soit P_1 le nouvel ensemble d'indices d'écart obtenu : il contient $\left(n\ (n-1)/2\right) - 2$ éléments. Le bon sens nous dicte encore d'effectuer dans P_1 les 4 opérations indiquées ci-dessus dans P_0. Nous éliminerons à nouveau les 2 valeurs extrêmes et obtiendrons un nouvel ensemble P_2 d'indices de distance contenant $\left(n\ (n-1)/2\right) - 4$ éléments.

Et ainsi de suite, nous cherchons le maximum et le minimum de ce nouvel ensemble pour les éliminer, etc. Nous aboutirons ainsi nécessairement à une incohérence au bout d'un certain nombre d'itérations. Il ne nous reste plus qu'à examiner tous les regroupements effectués : LE CLASSEMENT IPHIGÉNIE APPARAIT ALORS.

Voici un exemple pour un ensemble de 16 points pris dans un espace à 2 dimensions, la distance euclidienne ayant été choisie (indice d'écart a fortiori) :



Voici un autre exemple très simple où chaque étape de notre itération est précisément décomposée. Il s'agit d'un ensemble de 6 éléments (a, b, c, d, e et f) entre lesquels nous avons l'indice de distance suivant :

	a	*b*	*c*	*d*	*e*	*f*
a	0	1	8	14	9	15
b	1	0	12	17	2	16
c	8	12	0	3	13	18
d	14	17	3	0	5	11
e	9	2	13	5	0	19
f	15	16	18	11	19	0

Nous décomperons nos itérations en remplissant pas à pas le tableau suivant :

Itération	*Indice minimal*	*Indice maximal*	*Partition résultat*
I	$1 = d(a, b)$	$19 = d(e, f)$	(*ab*) (*c*) (*d*) (*e*) ⟷ (*f*)
II	$2 = d(b, e)$	$18 = d(c, f)$	(*abe*) (*c*) ↘ (*d*) ↗ (*f*)
III	$3 = d(c, d)$	$17 = d(b, d)$	(*abe*) (*cd*) ↕ (*f*) ↕
IV	$5 = d(d, e)$	$16 = d(b, f)$	impossible

(Les séparations obligatoires sont indiquées par des flèches)

Lors de la 4e itération, une incohérence arrive : on ne peut mettre ensemble *d* et *e* puisque *e* est déjà réuni avec *b* et que *b* doit être séparé de *d*.

La partition-résultat comporte donc 3 parties :

$$(a, b, e), (c, d), (f).$$

Remarque

Il se peut qu'un même indice d'écart se retrouve plusieurs fois. Ainsi, il peut arriver qu'une certaine itération indique plusieurs regroupements ou plusieurs séparations.

3. Intérêts particuliers d'IPHIGÉNIE

1. Le premier but de la méthode Iphigénie, est d'être le reflet du **bon-sens**, et de pouvoir ainsi servir de référence au milieu des très nombreuses méthodes existantes pour créer des typologies.

Il en découle tout naturellement une grande **simplicité** de principe que l'on peut aisément expliquer au non-spécialiste. Ceci est particulièrement intéressant dans le cadre d'une activité de conseil. En effet, nombreux sont les clients, intelligents et curieux des méthodes que l'on propose pour analyser leurs données et les regrouper, selon leurs propres termes *« en mettant ensemble ce qui se ressemble et en séparant ce qui ne se ressemble pas »*. Ce sera pour ces clients une grande satisfaction que de pouvoir comprendre le procédé de typologie utilisé.

2. Le deuxième intérêt d'Iphigénie, est sa **flexibilité.** Un simple indice d'écart sur un ensemble E, suffit en effet pour appliquer notre méthode. Or il est très aisé de définir un tel indice, que les données sur E soient quantitatives (avec ou sans pondérations), qualitatives, d'ordre logiques, ordinales ou pré-ordinales. On peut aussi appliquer Iphigénie, non pas sur l'ensemble E lui-même, mais sur l'ensemble des critères qui s'y appliquent. Rien n'empêche enfin d'appliquer plusieurs fois consécutives la méthode Iphigénie, afin de créer des hiérarchies, dont nous pouvons d'ailleurs constater des ressemblances avec des hiérarchies obtenues par d'autres méthodes, comme l'ultramétique inférieure maximale, etc.

On trouvera en annexe quelques idées d'indices d'écart selon le type de données rencontré.

3. Le troisème intérêt d'Iphigénie sera qualifié de **richesse informative.** Nous nous référons ici à la base de la théorie de l'information (cf. 3). L'entropie d'une expérience qui consiste à séparer les n éléments d'un ensemble à n éléments est $\log n$. Celle que consiste à les regrouper tous est 0. Chacune de ces 2 partitions extrèmes est inintéressante au sens de la recherche typologique et nous chercherons donc à nous en éloigner au maximum en créant une moyenne de p groupes dont l'entropie $\log p$, soit aussi éloignée que possible de chacune des 2 précédentes. Il nous faudra donc respecter l'équation :

$$\log p - 0 = \log n - \log p$$

c'est-à-dire :

$$2 \log p = \log n$$

ou :

$$p = \sqrt{n}$$

Une typologie optimale, au sens de la théorie de l'information, comportera donc $\sqrt{n}$ groupes de $\sqrt{n}$ éléments, en moyenne. Or, voici les statistiques obtenues après 600 applications de n[1] points au hasard placés dans un carré de 20 cm de côté, pour la distance euclidienne :

1. Les essais ont été faits sur 8 valeurs de n différentes, ce qui fait en tout 4 800 essais.

n	Nombre moyen d'éléments par groupe e_n	$(e_n - \sqrt{n})/\sqrt{n}$	Nombre moyen de groupe g_n	$(g_n - \sqrt{n})/\sqrt{n}$	Rappel : $\sqrt{n}$
5	2,29	1,79 %	2,25	0,45 %	2,24
10	3,29	1,90 %	3,23	2,22 %	3,16
15	4,11	6,20 %	3,94	1,81 %	3,87
20	4,75	6,26 %	4,59	2,68 %	4,47
25	5,36	7,20 %	5,09	1,80 %	5,00
30	5,79	5,66 %	5,66	3,28 %	5,48
35	6,37	7,60 %	6,05	2,20 %	5,92
40	6,81	7,75 %	6,46	2,22 %	6,32

Le rapprochement est surprenant. D'autre part, le fait de trouver à la fois un nombre d'éléments par groupe et un nombre de groupes moyen supérieurs à $\sqrt{n}$, qui semble presque choquant a priori, est en réalité bien normal. Il suffit pour nous en persuader d'observer le cas extrêmement simple d'un ensemble de 5 éléments avec ses 2 partitions extrêmes : 5 groupes de 1 élément et 1 groupe de 5 éléments. Cela donne en moyenne 3 groupes de 3 éléments dont le produit est évidemment supérieur à 5.[1]

Il semble ainsi statistiquement évident, au vu des pourcentages d'erreur calculés, qu'Iphigénie possède une richesse informative particulière. Mais attention, ce résultat n'est pas encore démontré, comme il serait très intéressant de la faire, par des déductions successives relevant de l'algèbre combinatoire. Il s'agit là d'une question ouverte...

4. Conclusion

Une typologie ne peut être présentée sans la donnée du critère optimisé correspondant. Nous utilisons pour cela la notion de rang généralisé r (cf. 2) par rapport à l'indice d'écart d, dans l'ensemble F des couples distincts de E :

$$F = \{x_i , x_j / j \neq i\}$$

La relation d'équivalence obtenue par la méthode Iphigénie, $\mathcal{R}^*$, sera celle qui maximise le critère C avec :

1. Il serait très facile de généraliser une telle propriété.

$$C(\mathcal{R}^*) = \underset{z \in \mathbb{R}^*}{\text{Max}}\ z/\left[\left(r(x_i, x_j) \leqslant z \Rightarrow x_i \not\mathcal{R} x_j\right) \wedge \left(\left(r(x_i, x_j) \geqslant \left(n(n-1)/2\right) - z\right) \Rightarrow x_i \mathcal{R} x_j\right)\right]$$

Ce critère est donc bien lourd à manipuler... La méthode Iphigénie au contraire est particulièrement « manipulable ». Rappelons qu'elle ne demande pas, comme le font d'autres méthodes, une décision a priori sur le nombre de groupes à obtenir. Elle se veut seulement l'expression de la sagesse intuitive, le bon sens en quelque sorte, applicable à de très nombreux cas, tout en donnant une information intéressante. Pour cela, elle pourrait représenter une certaine **référence** au milieu de toutes les autres méthodes existantes. Et nous conclurons en citant notre très regretté ami Guy Der Megreditchian, ancien directeur de la statistique mathématique à la météorologie nationale ; qui écrivait en 1973 : « La méthode Iphigénie est si claire et pleine de bon-sens, que nous pouvons seulement nous demander pourquoi personne n'y a pensé depuis de nombreuses années ».

BIBLIOGRAPHIE

(1) BERNARD G. et BERRONDO-BESSON M. "Douze méthodes d'analyse multicritère", R.I.R.O., vol. 3, 1971, 19-66 (article).

(2) BERRONDO-BESSON Marie "Rang généralisé et agrégation de classement", R.A.I.R.O., vol. 1, 1975, 37-59 (article).

(3) CHOQUET *Topologie,* Masson, 1964.

(4) PACTEAU François-Xavier et ROUY Stéphane *Iphigénie, un procédé de typologie original,* Projet personnalisé à l'École Centrale de Paris, juin 1992.

(5) YAGLOM A.M. et YAGLOM I.M. *Probabilités et Information,* Dunod, 1959.

ANNEXE

Nous présentons ici quelques indices d'écart possibles parmi bien d'autres :

Types de données	*Dans l'ensemble des n éléments de E*	*Dans l'ensemble des m critères*
Quantitatives	$\sqrt{\sum_{k=1}^{m} p_k \left(\frac{y_k^i - y_k^j}{\sigma_k}\right)^2}$ (ou distances euclidiennes dans les cas les plus simples	$1 - \lvert r_{k,\,l} \rvert$ ↓ coefficient de corrélation entre les séries correspondant aux crières respectifs k et l
Qualitatives	$\sum_{k=1}^{m} \frac{p_k X_{i,\,j}^k}{C_k - 1}$ ↗ $\begin{cases} 1 \text{ si} \\ y_k^i \neq y_k^j \\ 0 \text{ sinon} \end{cases}$ ↓ nombre de possibilités correspondant au critère k	Nombre de couples de E possédant la même qualité selon les critères k et l ↑ $1 - (U_{k,\,l} / V_{k,\,l})$ ↓ nombre de couples de E possédant la même qualité selon l'un au moins des critères k et l
Logiques	$\sum_{k=1}^{m} p_k X_{i,\,j}^k$ ↗ $\begin{cases} 1 \text{ si} \\ y_k^i \neq y_k^j \\ 0 \text{ sinon} \end{cases}$	Voir indice ci-dessus
Ordinales ou préordinales	$\sqrt{\sum_{k=1}^{m} p_k (r_k^i - r_k^j)^2}$ ↙ rang généralisé de x_i selon le critère k (bibliographie 2)	Nombre de couples (x_i, x_j) de E tels que : $x_i > x_j$ selon les critères k et l ↑ $1 - (U_{k,\,l} / V_{k,\,l})$ ↙ nombre de couples (x_i,x_j) de E tels que : $x_i > x_j$ selon l'un au moins des 2 critères k et l