La typologie

Michel Calciu

Introduction

Segmentation et typologie

En marketing il y a un grand intérêt de développer des modalités utiles pour classifier des objets. Très souvent les objets à classifier sont les clients [1] . L'objectif est de grouper les clients potentiels dans des groupes homogènes assez grands pour être profitablement gérés.

En marketing on distingue deux modalités de classification: la segmentation et la typologie [2] (Bon, J. et Gregory, P., 1986).

La segmentation essaye de "fractionner l'ensemble de la population en utilisant des critères choisis en fonction de leur pertinence pour le sujet donné".

La typologie essaye de rassembler les individus qui se ressemblent du point de vue des variables qui les caractérisent, elle cherche à regrouper les individus selon les ressemblances et proximités.

La typologie [3] a pour objet la description d'un ensemble d'individus ou d'objets caractérisés par un ensemble d'attributs, à l'aide de leur regroupement en classes [4] . Ces classes sont établies de telle sorte que les objets appartenant à la même classe soient les plus semblables possibles et que les objets appartenant à deux classes différentes soient les plus dissemblables possibles.

La répartition des individus entre les classes peut être une finalité de la technique, mais également le repérage des attributs qui permettent de mieux différencier les groupes.

Exemples:

- L'étude de différentes marques d'un produit selon leurs caractéristiques perçues par les consommateurs, de façon à établir le positionnement respectif de ces différentes marques. Les types seront constitués des marques qui bénéficient de perceptions similaires sur l'ensemble des caractéristiques

- L'étude d'une population en termes d'activités, d'intérêts et d'opinions. Les individus appartenant au même type manifestent le même genre d'activités, d'intérêts ou d'opinions. Cette étude permet d'aboutir à une typologie de styles de vie.

Comparaison

La typologie rappelle, par ses objectifs, d'autres techniques de traitement de données, et plus particulièrement la segmentation, l'analyse discriminante et l'analyse en composantes principales. Mais elle en diffère par les méthodes qu'elle utilise et par la présentation de ses résultats.

La définition des proximités

a) Les données

En typologie, chaque individu est caractérisé par les valeurs prises sur un ensemble de variables. Ces variables peuvent être de nature très différente.

l - On distinguera, en premier lieu, les variables selon le rôle qu'elles jouent dans la démarche : on distinguera les variables actives et les variables passives de la typologie. Les variables actives servent à constituer les groupes alors que les variables passives vont servir, dans un deuxième temps, à expliquer ces groupes.

2 - En second lieu, les variables diffèrent selon la nature de la mesure utilisée. Les variables les plus fréquemment utilisées seront soit de nature quantitative soit de nature qualitative.

Les données quantitatives apparaissent les plus simples à traiter. Elles soulèvent cependant certaines difficultés lorsque les unités de mesure utilisées ne sont pas les mêmes d'une variable à une autre. Dans une telle situation, il conviendra de standardiser les variables afin de ne pas privilégier telle ou telle variable.

En fonction de la nature des variables : quantitatif, qualitatif, binair on distingue plusieurs mesures de distances

Distance euclidienne

Dans un espace euclidien (orthogonal) en deux dimensions, comme celui de la figure 1, la distance dij entre deux objets i et j est calculé selon le théorème de Pythagore :

Figure 1: Evaluation de la proximité entre individus

Les différentes distances qui viennent d'être évoquées reposent sur une conception spécifique de la proximité : deux objets sont proches quand il y a peu d'écart entre eux sur l'ensemble des dimensions concernées. D'autres indicateurs de distance reposent sur la notion de structure : on considère alors que deux individus sont proches dès que leurs structures d'attributs sont similaires.

La distance du chi2

La distance du pour les tableaux de dénombrement, ainsi que le coefficient de corrélation entre les caractéristiques des individus pour les autres données quantitatives sont alors utilisables.

Les données du tableau 1. permettent d'illustrer ces différentes conceptions de la proximité. En termes de distance euclidienne, a et b paraissent plus proches que a et c d'une part et b et c d'autre part :

Tableau 1- Variable \ observations

Les distances euclidiennes sont : dab = 20 dac = 37 dbc= l09

Si l'on emploie la distance du , les proximités sont différentes puisque b et c sont jugés identiques:

dab = 0,257 dac = 0,318 dbc = 0

Indices de distance pour données qualitative binaires

2 - Quand il s'agit de données qualitatives la proximité entre deux individus sera jugée d'autant plus forte que ces deux individus présentent des caractéristiques communes. On distinguera ainsi :

- les coïncidences positives (P) : nombre de fois où, pour les deux individus, une caractéristique donnée sera présente;

- les coïncidences négatives (N) : nombre de fois où, pour les deux individus une caractéristique donnée sera absente;

- les non-coïncidences (U): nombre de fois où la caractéristique sera présente pour un des deux individus et absente pour l'autre.

A partir de ces coïncidences ou non-coïncidences plusieurs coefficients d'association peuvent être définis. Par exemple:

P/(P + N + U)

pourcentage de coïncidences positives par rapport au nombre total de coïncidences positives possibles [6] .

Les indicateurs de proximité entre groupes

Problèmes d'agrégation de groupes

Lorsqu'il s'agit d'évaluer la proximité entre deux groupes d'individus, plusieurs procédures sont utilisables.

Imaginons le cas de cinq objets analysés sur deux variables présenté dans le tableau 2.

Tableau 2 - Variables\Individus

Le plus proche voisin

l - Dans la méthode du plus proche voisin, la distance entre deux groupes I et II est assimilée à la distance entre les deux éléments les plus proches, l'un appartenant au groupe I et l'autre au groupe II.

Sur la figure 2, le groupe I contient les éléments a et c; le groupe II les éléments b et d. L'application de la méthode du plus proche voisin amène à évaluer la distance entre I et II par la distance dbe.

d[(b),(ea)] = Min[dbe, dba] = dbe

L'utilisation de cette méthode entraîne le risque de faire apparaître des groupes très hétérogènes, puisqu'on ne se préoccupe pas des éléments les plus extrêmes des groupes. Dans l'exemple considéré, le groupe [b] apparaît plus proche du groupe [ea] que du groupe [dc], alors que a est très éloigné de b.

Figure 2: Evaluation de la proximité entre groupes

Le voisin le plus éloigne

2 - Dans la méthode du voisin le plus éloigne, la distance entre deux groupes I et II est assimilée à la distance entre les deux éléments les plus éloignés, l'un appartenant au groupe I et l'autre au groupe II. Sur la figure 2., la distance entre I et II serait alors mesurée par dba

d[(b),(ea)] = Max[dbe, dba] = dba

Dans l'exemple de la figure 2, le groupe I serait associé au groupe II et non au groupe I.

Le chaînage moyen

La constitution des groupes

Le tableau de distances

Les données initiales sont constituées par un tableau individus-variables où chacun des n individus est caractérisé sur un ensemble de dimensions. A partir de ce tableau initial, une matrice de proximité (n.n) peut être établie : chaque entrée ij de la matrice donne la distance dij qui sépare l'individu i et l'individu j.

Une fois connues les n(n - 1)/2 proximités entre individus, la constitution de groupes peut être envisagée. Elle se réalise selon deux grandes catégories de procédures qui recourent respectivement à des méthodes hiérarchiques ou non hiérarchiques.

a) Les méthodes hiérarchiques

Dans cette première catégorie de méthodes, le nombre de groupes auquel on désire parvenir n'est pas fixé a priori; il sera déterminé au vu des résultats.

I - La méthode hiérarchique ascendante ou agrégative utilise la procédure suivante :

I) parmi l'ensemble des n individus on sélectionne les deux plus proches par lecture des distances dans la matrice de proximité : un premier regroupement est opéré entre ces deux individus;

2) une nouvelle matrice de distances est établie qui ne comprend que n - l lignes correspondant à n - 2 individus isolés et I groupe comprenant 2 éléments. A nouveau les deux objets les plus proches sont repérés et regroupés. Ces deux objets sont soit deux individus isolés ou un individu et un groupe;

3) la phase 2 est renouvelée jusqu'à ce que tous les individus soient regroupés.

Cette méthode est hiérarchique puisque une succession de regroupements sont opérés, de celui qui met en jeu les proximités les plus fortes vers celui qui correspond aux proximités les plus faibles. Elle est dite ascendante puisque l'analyse remonte de l'individu isolé vers le groupe [7] .

A première vue, cette procédure apparaît assez contradictoire avec l'objectif fixe, puisque, en fin de parcours, tous les individus se retrouvent dans une

seule classe. En fait, l'analyste, à partir de l'étude des différents niveaux de regroupements décidera du nombre de groupes qui lui paraît le plus judicieux à l'aide d'un niveau de distance-seuil au-delà de laquelle les regroupements seront jugés trop hétérogènes. Cette étude est facilitée par l'emploi d'une représentation graphique du processus de regroupement, le dendogramme.

Application de la méthode

2 - Le tableau 3, fournit un exemple d'application de la procédure qui vient d'être présentée. Les données utilisées ont déjè été présentées au tableau 2, où cinq individus [a,b,c,d,e] sont repérés sur deux variables X1, et X2 (voir aussi figure 2). A partir de ces données la matrice de proximités (tableau 3.a) est construite. Ce sont ici des distances euclidiennes qui sont employées : par exemple la distance dab entre les individus a et b est obtenue par :

La distance la plus faible est celle qui sépare c et d: ils forment le premier groupe. Une nouvelle matrice de proximités (tableau 3b) est établie pour les individus isolés a, b, e et pour le groupe [c,d]. La mét^hode du plus proche voisin a été employée pour calculer la distance entre le groupe [c,d] et les individus isolés.

Pour la deuxième étape de regroupement, la plus forte proximité apparaît entre les individus a et e qui forment alors un groupe [a,e]. La nouvelle matrice (tableau 3c) qui découle de ce regroupement permet de voir que b doit être rattaché à ce groupe, d'où la création d'un ensemble [a,e,b]. Le tableau 3d montre qu'une distance de 6 sépare le groupe [c,d] du groupe [a,e,b].

Procédure de constitution de groupes

a) Matrice des distances:

b) 1re étape: regroupement de (c + d); nouvelle matrice des distances:

c) 2e étape: regroupement de (a + e); nouvelle matrice de distances:

d) 3e étape: regroupement de (a + b); nouvelle matrice de distances:

Les résultats sont visualisés par le dendogramme qui apparaît dans la figure 3. Un distance-seuil de 3 amènerait ainsi à choisir une typologie en trois classes : [c,d], [a,e] et b. Une distance-seuil de 4,5 ne ferait apparaître que deux classes : [c,d] et [a,e,b]; mais le groupe II serait alors moins homogène. Un arbitrage doit être effectué entre l'homogénéité des classes et leur nombre.

b) Les méthodes non hiérarchiques

Les méthodes hiérarchiques deviennent très lourdes à utiliser si le nombre d'objets à classer est important : pour n individus, il faut passer n - I étapes de regroupement qui se traduisent par la construction de n - 1 matrices.

Les méthodes non hiérarchiques ou encore nodales demandent souvent moins de calculs et sont donc capables de traiter de plus grands nombres d'individus. Dans cette seconde catégorie de méthodes, le nombre k de groupes est fixé a priori.

Plusieurs procédures sont concevables. La plus simple consiste à sélectionner dans la population étudiée k individus-type autour desquels seront agglomérés, un par un, les autres n - k individus en fonction de leurs distances respectives par rapport aux individus de référence (méthode dite des K-means).

Les k individus-type sont choisis en fonction de certaines de leurs caractéristiques ou plus simplement à l'issue d'un processus de sélection aléatoire. Dans ce dernier cas, il sera nécessaire de procéder à

plusieurs simulations avant d'aboutir à une configuration acceptable, c'est-à-dire qui permette de faire apparaître des groupes suffisamment homogènes et différents.

Ce schéma de base est susceptible de recevoir de nombreuses variantes. Par exemple, c'est la méthode des nuées dynamiques' à la fin d'une première étape d'agrégation, les centroïdes de classes sont calculés et servent de bases de référence pour une nouvelle étape de calcul. Des individus jusqu'ici assignés à une classe peuvent alors recevoir une nouvelle affectation.

Explication et validation des groupes

a) Explication

Une fois les groupes formés, il s'agit de les décrire, c'est-à-dire repérer les variables qui ont joué le plus grand rôle dans leur formation et par la suite caractériser les types.

I - La description est tout d'abord réalisée à l'aide des variables actives. Pour chaque variable, une moyenne est calculée par groupe. Si les différences de moyenne qui apparaissent entre les groupes sont faibles pour une variable donnée, celle-ci aura peu contribué à la formation des classes. Au contraire,

des différences significatives seront le signe d'un pouvoir discriminant important.

Le caractère discriminant d'une variable donnée pourra être apprécié à l'aide de procédures graphiques permettant de visualiser le profil des différents groupes. On pourra également avoir recours à des tests statistiques, l'analyse de la variance par exemple, si l'on a affaire à des données quantitatives.

2 - L'utilisation de variables passives permet d'introduire dans la description des groupes d'autres variables que celles qui ont permis la constitution des groupes. Ce seront, par exemple des variables d'identification telles que l'âge, la PCS, le sexe. Le fait de pouvoir associer un type de comportement avec ces variables d'identification peut être de nature à orienter une stratégie de segmentation.

b) Validation

Les groupes formés sont-ils réellement différents les uns des autres ? Plusieurs procédures de validation sont concevables :

- recours au test d'analyse de la variance sur chaque variable active;

- application d'une analyse discriminante où la variable à expliquer est le type de groupe auquel appartiennent les individus;

- refaire une nouvelle typologie sur un ensemble d'individus-témoin qui n'auront pas été utilisés dans la première analyse.

Application

Données

Nuage de points

Tableau de classification

Dendrogramme (la méthode du lien simple)

[1] Gilbert A. Churchill, "Marketing Research, Methodological Foundations", 5e Ed., Dryden Press, 1991.

[2] Bon, Jean et Pierre Gregory, "Techniques marketing", Vuibert, Paris, 1986.

[3] Vedrine J.-P. "Le traitement des données en marketing", Ed. Organisation, Paris, 1991.

[4] On parle aussi de méthode de classification ou de taxinomie numérique.

[5] De façon plus généraie on peut utiliser la distance de Minkoswky: dij = [Erreur! Signet non défini.(Xik - Xjk)n]1/n

[6] Quand on affaire à des données de nature différente, certaines qualitatives et d'autres quantitatives, il est possible de construire des indices de proximité mixtes.