Agrégation optimale sous contrainte de contiguïté : aspects théoriques et mise en œuvre avec applications à des cas pratiques.

Marc CHRISTINE[1] et Michel ISNARD[2]

Cet article reprend et enrichit une problématique déjà présentée aux JMS 2000[3]. Il s'agit de constituer, au sein d'une population de référence, des classes présentant des conditions d'homogénéité ou d'hétérogénéité maximales vis-à-vis de certaines caractéristiques quantitatives.

L'originalité du problème réside dans le fait qu'on associe également aux unités statistiques de la population un référencement géographique fournissant des informations sur le « positionnement » relatif des unités et sur la topologie des classes à constituer. Plus précisément, le positionnement sera appréhendé par le concept de *contiguïté*, qui constituera une relation binaire entre les unités de la population. On astreindra alors les classes constituées à être *connexes* vis-à vis de la relation de contiguïté.

L'objectif est donc de construire une méthode pour constituer de manière automatique une partition de la population de référence en classes de nombre donné, astreintes à des seuils de taille fixés, connexes, et réalisant un optimum (minimum ou maximum) d'une fonction objectif définie comme la somme d' une caractéristique de chaque classe, baptisée *inertie*.

L'exemple où la notion d'inertie est la plus simple à interpréter est celui où l'on cherche à minimiser (si l'on veut que les classes formées soient homogènes) ou maximiser (si l'on veut qu'elles soient hétérogènes) la variance intra-classes d'une variable quantitative. Dans ce cas, l'inertie d'une classe est sa variance. Cette inertie peut être calculée à partir d'une distance euclidienne définie entre les éléments de la population de référence.

Il est possible en fait de généraliser cette notion d'inertie en introduisant une pseudo-distance d entre les unités de la population de référence, non nécessairement euclidienne, et non nécessairement nulle lorsqu'on on mesure la distance d'une unité à elle-même. Sous réserve de donner des poids α_i à chaque unité i de la population, l'inertie d'une classe K sera donnée par :

$$I(K) = \frac{1}{2} \frac{\sum_{i \in K} \sum_{j \in K} \alpha_i \alpha_j d_{i,j}^2}{\sum_{i \in K} \alpha_i}.$$

[1] Insee, DCSRI

[2] Insee, DCSRI

[3] « Un algorithme de regroupement d'unités statistiques selon certains critères de similitude », Marc CHRISTINE et Michel ISNARD, VII[èmes] Journées de Méthodologie statistique, 4-5 décembre 2000.

Cette formule redonne l'expression de la variance de la classe lorsque la distance considérée $d_{i,j}$ est euclidienne.

Le problème de partitionnement de la population, astreignant les classes aux contraintes indiquées ci-dessus et réalisant l'optimum de la somme des inerties des classes, possède nécessairement des solutions, s'agissant de populations finies, mais celles-ci sont inaccessibles. Seules des solutions algorithmiques conduisant à des optima locaux sont envisageables. Elles sont fondées d'une part sur des outils de classification ascendante hiérarchique (CAH), en se limitant à chaque étape de l'agrégation à la constitution de classes connexes, d'autre part à des procédures d'échanges d'unités entre classes afin d'améliorer la valeur de la fonction objectif à partir d'une solution initiale. Toutes les procédures informatiques résultantes ont été écrites spécifiquement en langage SAS.

Les méthodes exposées seront mises en œuvre et illustrées sur des cas concrets. Un premier groupe d'exemples permettra de montrer comment construire une partition des communes au sein d'une région ou d'un département en un nombre fixé de classes connexes, en s'appuyant sur des critères de minimisation ou de maximisation de la variance intra-classe pour des variables quantitatives standards : ceci illustre le cas de la distance euclidienne et d'une inertie s'interprétant en termes de dispersion.

Dans le second groupe d'exemples, on cherchera à appliquer ces techniques à des cas non euclidiens, par exemple la construction de zones d'attraction de l'emploi ou de « bassins de vie » obtenus en comparant département de naissance et département de résidence.