Les difficultés liées aux dettes au Québec

www.endettement.inrs.ca

Classification des endettés selon l'origine de la dette

Cette fiche synthèse et les données qu'elle mobilise sont issues de l'Enquête sur l'endettement des ménages québécois (EEMQ), réalisée en janvier et février 2022 auprès de 4816 adultes résidant au Québec dans le cadre du projet de recherche « Le surendettement parmi les ménages québécois ». Visitez le site web du projet pour obtenir la documentation complète, les détails méthodologiques et des fiches synthèses additionnelles .

> ***Citer cette fiche :*** *Les difficultés liées aux dettes au Québec. 2023. «Classification des endettés selon l'origine de la dette». Date d'accès jour/mois/année. https://endettement.inrs.ca/annexe méthodologique/*

Il existe plusieurs méthodes pour établir des profils-types. En statistique descriptive multivariée, la classification des observations (individus) est couramment utilisée pour établir des profils-types. Une stratégie classique consiste à réaliser une analyse des composantes principales (ACP) ou analyse des correspondantes multiples (ACM) des données puis à appliquer une méthode de classification sur les scores des individus (mesurés sur les composantes principales obtenues). Ici, pour établir le profil-type des endettés selon l'origine de la dette, nous optons pour une démarche en deux étapes. Il s'agit de faire la classification des variables suivie de la classification des observations (individus).

L'objectif de la classification de variables est de construire des classes de variables fortement liées entre elles, c'est-à-dire celles sur lesquelles il y a un lien dans la façon dont les individus ont répondu aux questions. On fait recours à l'algorithme de ClustOfVar développé par Chavent et al. (2011, 2012) pour réaliser la classification des variables. L'avantage de cet algorithme est qu'il est adapté aux données quantitatives, aux données qualitatives et aux données mixtes.

L'approche ClustOfVar est une méthode de classification hiérarchique ascendante de variable (Chavent et al. 2012). Elle fournit simultanément des groupes de variables ainsi que les variables synthétiques associées aux classes de variables. L'approche ClustOfVar maximise un critère d'homogénéité basé sur la notion de corrélation pour les variables quantitatives et de rapport de corrélation pour les variables qualitatives. L'homogénéité $H(C_k)$ de la classe est une mesure d'adéquation entre les variables de la classe C_k et la variable synthétique quantitative de la classe notée y_k. Elle est définie par :

$$H(C_k) = \sum_{x_j \in C_k} r^2_{x_j, y_k} + \sum_{z_j \in C_k} \eta^2_{y_k|z_j}$$

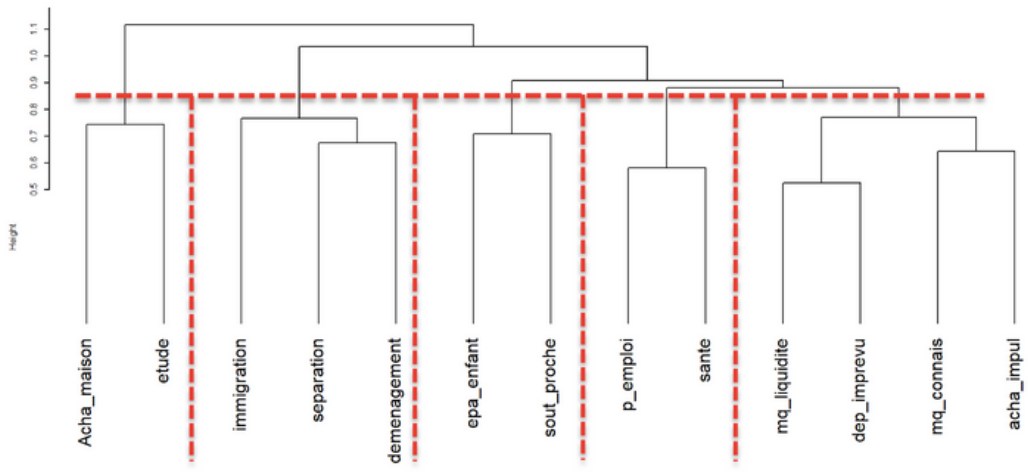
Où r^2 désigne la corrélation de Pearson au carré entre y_k et la variable quantitative x_j et n^2 désigne le rapport de corrélation entre y_k et la variable qualitative z_j . La variable synthétique y_k est la variable "la plus liée" aux variables de la classe au sens du critère H qu'elle maximise.

La deuxième étape consiste à faire la classification des individus à partir des variables synthétiques obtenues de l'algorithme ClustOfVar. L'objectif de la classification d'observations est de réunir des individus qui se ressemblent. Le mécanisme consiste à séparer n individus en K groupes d'individus. Il existe plusieurs méthodes et algorithmes associés permettant de faire la classification des individus. L'une des méthodes les plus utilisées est la classification hiérarchique ascendante (CAH). Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un arbre de classification. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée. Ici, nous allons effectuer une CAH basée sur le critère de Ward.

Classification des sources de l'endettement

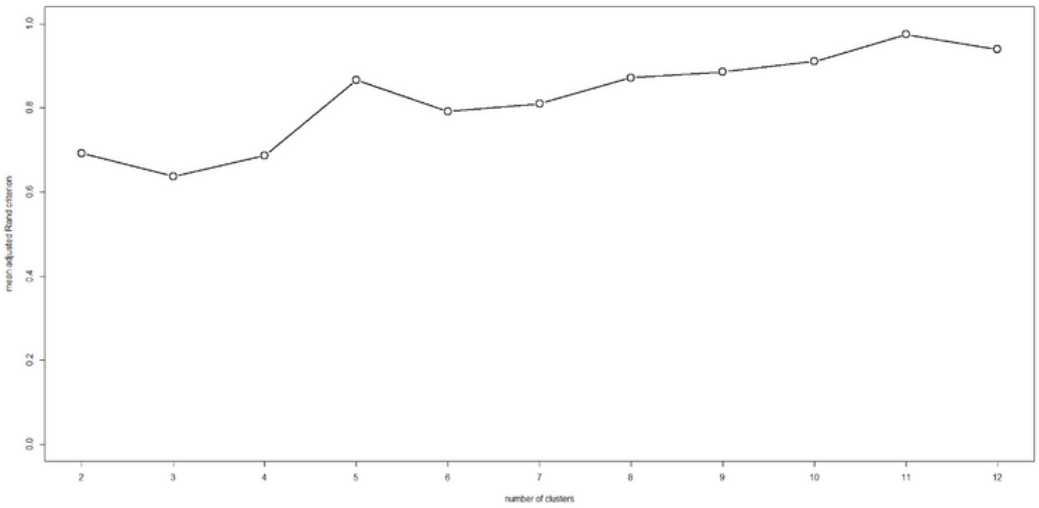
L'algorithme de ClustOfVar nous a permis d'établir le lien entre les sources de l'endettement. L'analyse révèle qu'il existe un lien fort entre « achat de maison/véhicule » et « étude ». Sur le dendrogramme, on voit une forte association entre « Manque de connaissance » et « Achat impulsif ». Le dendrogramme montre également les différentes possibilités de partition des sources de l'endettement. En regardant la partie supérieure du dendrogramme, on voit clairement qu'on peut faire une partition en deux classes de variables, « Achat maison/véhicule » et « Études » dans une classe et les variables autres variables dans une autre classe. On voit par ailleurs qu'on peut faire une partition en trois classes, en quatre classes, en cinq classes ou plus. Le dendrogramme permet d'avoir une idée sur le nombre de découpages possibles, mais il ne permet pas de savoir quelle est la meilleure partition en termes de stabilité dans le jeu de données.

Figure 1. Dendogramme



Le diagramme de stabilité donne une idée sur la meilleure partition en termes de stabilité dans le jeu de données. À partir de nos données, il ressort qu'une partition des sources de l'endettement en deux classes est plus stable qu'une partition en trois classes (figure 2). On voit également qu'avec une partition en cinq classes, on a plus de stabilité qu'avec une partition en trois classes. La partition en onze classes est visiblement la partition la plus stable. Notre objectif étant de réduire le mieux possible le nombre de variables, nous avons donc opté pour une partition des sources de l'endettement en cinq classes.

Figure 2. Diagramme de stabilité de la partition



Le tableau 1 résume la composition des cinq classes de variables. Le rapport de corrélation entre chaque variable qualitative et la variable synthétique quantitative (VS) de la classe (indiqué entre parenthèses) montre que plus le nombre de variables de la classe est faible, plus les variables qui composent la classe sont fortement reliées à la VS de la classe. On peut avancer comme explication que, pour les plus grandes classes, certaines valeurs sont plus faibles, car elles regroupent des variables de thématiques plus diversifiées.

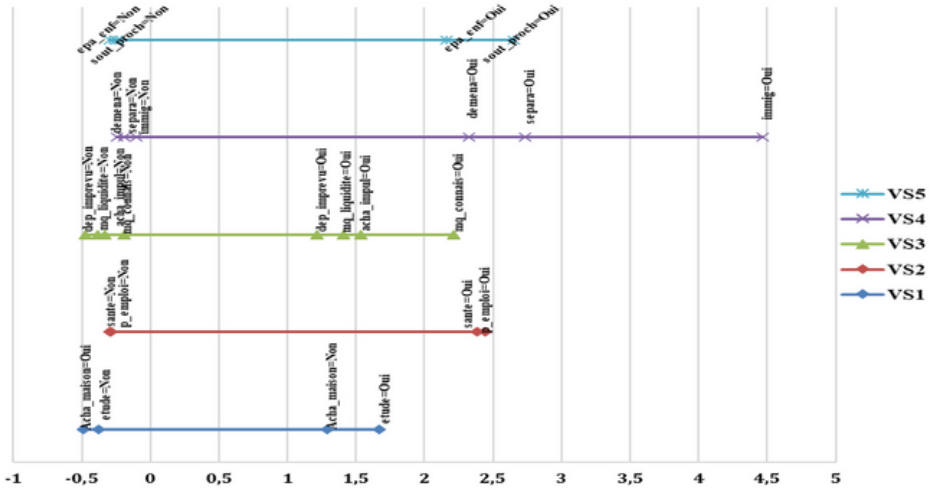
Tableau 1. Composition des classes de variables

Classe	1	2	3	4	5
Variable (rapport de corrélation)	acha_mais (0,63) etude (0,63)	sante (0,71) p_emploi (0,71)	dep_imprevu (0,58) mq_liquid (0,55) acha_impul (0,51) mq_connais (0,42)	demena (0,75) separ (0,53) immig (0,46)	epa_enf (0,65) sout_proch (0,65)
Pourcentage d'inertie expliquée	79,68	70,47	48,54	64,23	77,46

La figure ci-dessous permet de caractériser les variables synthétiques, elle montre la valeur moyenne des VS par modalités sur le gradient des VS. Les valeurs moyennes de VS1 des modalités « acha_mains=Non » et les « etude=Oui » sont proches et positives. En ce qui concerne les modalités « acha_mais=Oui » et « etude=Non », les valeurs de VS1 sont proches et négatives.

Ce qui sous-entend que si un individu répond « oui » pour achat de maison, il est plus probable qu'il réponde non pour étude. Pour les autres, on remarque que si un individu répond « oui » pour une variable de la classe, il est plus probable qu'il réponde « oui » pour les autres variables de la classe.

Figure 3. Gradient des variables synthétiques



Profils-types des endettés selon les sources de l'endettement

La CAH nous a permis de dégager cinq profils des endettés selon la source de l'endettement. Environ 51% des endettés sont dans la classe 1.

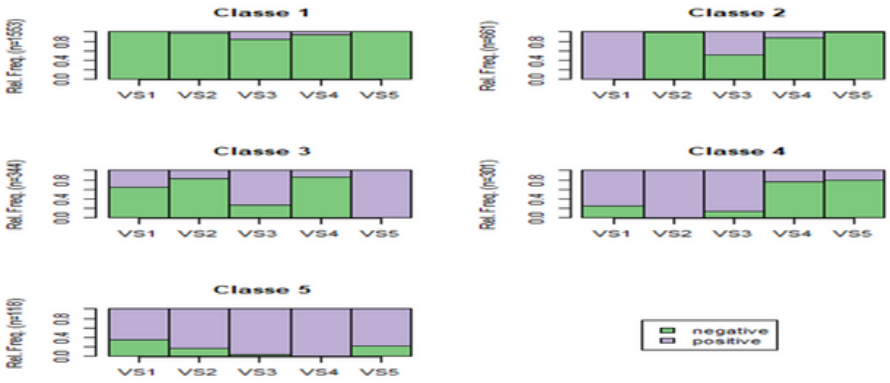
Dans les classes 3 et 4 on a respectivement 12 et 10 % des endettés. Seulement 4 % des endettées sont dans la classe 5. La classe 2 représente 22% des endettés

Tableau 2. Composition des classes d'individus

Classe	1	2	3	4	5
Effectif	1553	661	344	301	118
Pourcentage pondéré	51	22	12	11	4

La figure 4 résume la caractérisation des différentes classes. On voit clairement, la classe 1 est caractérisée par des individus qui ont essentiellement des valeurs négatives pour l'ensemble des variables synthétiques. Cela signifie que la raison principale de l'endettement des individus de cette classe est l'achat d'une maison/véhicule (voir figure 3). La classe 2 est caractérisée principalement par des individus qui ont des valeurs positives pour VS1, cela veut dire que la principale raison de l'endettement dans cette classe est l'étude. Dans la classe 3, la proportion des valeurs positives est plus importante pour la VS5, alors l'endettement des ndividus de cette classe est lié essentiellement à l'épanouissement des enfants et le soutien apporté à un proche. L'endettement des individus de la classe 4 est lié essentiellement à la perte d'emploi et au problème de santé. La classe 5 est constituée essentiellement par des individus qui ont contracté des dettes pour des raisons de déménagement, de séparation, d'immigration, de dépenses imprévues, de manque de connaissances et d'achat impulsif.

Figure 4. Caractérisation des classes d'individus à partir des VS de ClustOfVar



Références

Kuentz-Simonet V., Lyser S., Candau J., Deuffic P., Chavent M., Saracco J., 2013, « Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement », Journal de la Société Française de Statistique, 154(2), p. 37-63.

Chavent M., Kuentz V., Liquet B., Saracco J., 2011, « Classification de variables: le package clustofvar », 43èmes Journées de Statistique (SFdS), p. 6-p.

Chavent M., Kuentz-Simonet V., Liquet B., Saracco J., 2012, « ClustOfVar: An R Package for the Clustering of Variables », Journal of Statistical Software, 50, p. 1-16. doi:10.18637/jss.v050.i13

Gilbert N., Mewis R. E., Sutcliffe O. B., 2020, « Classification of fentanyl analogues through principal component analysis (PCA) and hierarchical clustering of GC–MS data », Forensic Chemistry, 21, p. 100287. doi:10.1016/j.forc.2020.100287

Harkat M.-F., 2003, Détection et localisation de défauts par analyse en composantes principales, phdthesis, Institut National Polytechnique de Lorraine - INPL.

Kuentz V., Lyser S., Candau J., Deuffic P., 2015, « ClustOfVar-based approach for unsupervised learning: Reading of synthetic variables with sociological data », Electronic Journal of Applied Statistical Analysis, 8(2), p. 170-197.
Palm R., 1998, « L'analyse en composantes principales: principes et applications », Notes de Statistique et d'Informatique, 2.

Schaeffer Y., Kuentz-Simonet V., Rambonilaza T., 2022, « Approche par clustering de variables de la qualité de vie à l'échelle des territoires–la méthode ClustOfVar », Revue dEconomie Regionale Urbaine, p. 5zj-534.

Institut national de la recherche scientifique

Cette publication a été préparée par l'équipe de Les difficultés liées aux dettes au Québec. Vous pouvez envoyer vos commentaires relatifs à cette publication à l'adresse suivante : https://endettement.inrs.ca/contact/