METHODOLOGIE DE CREATION D'UN INDICE DE DEFAVEUR CONTEXTUELLE – UN OUTIL PERMETTANT L'ANALYSE DES INEGALITES SOCIALES DE SANTE

Benoît Lalloué[1,2], Jean-Marie Monnez[2], Cindy Padilla[1], Denis Zmirou-Navier[1] & Séverine Deguen[1]

[1] *EHESP, Avenue du Professeur Léon-Bernard 35000 RENNES ;*
[2] *IECN, Faculté des Sciences, Université de Lorraine, 54506 VANDOEUVRE-LES-NANCY ;*
benoit.lalloue@ehesp.fr ; jean-marie.monnez@univ-lorraine.fr ; cindy.padilla@ehesp.fr ;
denis.zmirou@ehesp.fr ; severine.deguen@ehesp.fr

Résumé. Afin d'étudier l'existence d'inégalités sociales de santé, les données contextuelles constituent une bonne alternative aux caractéristiques socioéconomiques individuelles souvent difficilement accessibles. Des indices peuvent être utilisés pour synthétiser les multiples dimensions du statut socioéconomique. Dans ce contexte, les objectifs principaux de ce travail sont de développer une procédure statistique afin de créer un indice socioéconomique contextuel puis d'effectuer une discrétisation de cet indice. Ce développement a été réalisé à partir de données collectées à l'échelle de l'IRIS sur 3 grandes agglomérations françaises, Lille, Lyon et Marseille. La méthode utilise plusieurs ACP successives pour sélectionner les variables et créer l'indice. Les catégories de défaveur socio-économique sont constituées par une CAH sur facteurs d'ACP. Vingt variables socio-économiques, couvrant les multiples dimensions de la défaveur, sont retenues dans la constitution de l'indice socioéconomique. Une partition optimale obtenue par CAH est en trois classes qui s'ordonnent le long du premier axe de l'ACP.

Mots-clés. Indice, socioéconomique, inégalités sociales de santé, analyse en composantes principales, classification ascendante hiérarchique.

Abstract. To study social health inequalities, contextual data may constitute an appropriate alternative to individual socioeconomic characteristics, often hard to retrieve. Indices can be used to summarize the multiple dimensions of the neighborhood socioeconomic status. The main objectives of this work are to develop a statistical procedure to create a neighborhood socioeconomic index and to investigate the influence of the clustering method of the deprivation index on the measure of health inequalities. Our study setting is composed of three major French metropolitan areas: Lille, Lyon and Marseille. The statistical unit is the French census block. The methodology uses several successive principal components analyses to select variables and create the index. Deprivation categories are drawn with hierarchical clustering on principal components. Twenty socioeconomic variables are selected in the neighborhood deprivation index. An optimal clustering obtained by HC is in three classes and the clusters are ordered along the first axis of the PCA.

Keywords. Index, socioeconomic, social health inequalities, principal component analysis, hierarchical clustering.

1 Introduction

Les inégalités sociales de santé sont bien documentées dans la littérature épidémiologique : globalement, les populations plus défavorisées ont un risque plus élevé de maladie que les populations plus favorisées pour de nombreux événements de santé (issues de grossesse et mortalité infantile, maladies cardiovasculaires et respiratoires, santé mentale, …) (Marmot 2005). Une majorité d'études explorant ces inégalités utilisent des caractéristiques socioéconomiques individuelles, mais celles-ci sont souvent difficiles à recueillir. Les données agrégées peuvent alors constituer une alternative appropriée pour la recherche en santé publique. De plus, de récentes

études ont montré que, même après ajustement sur les caractéristiques socioéconomiques individuelles, la défaveur mesurée sur le lieu de résidence expliquait malgré tout une part significative des inégalités de santé (Chaix et al. 2007; Zeka et al. 2008).

Le statut socioéconomique contextuel est une notion complexe qui combine diverses dimensions comme l'emploi, le revenu, l'éducation, le logement, ou les liens sociaux (Braveman et al. 2005; Galobardes et al. 2006; Krieger et al. 1997; Morris & Carstairs 1991). Ces multiples dimensions peuvent être résumées au sein d'un indice de défaveur contextuelle. Cependant, parmi ces indices, certains utilisent un faible nombre de variables et/ou sont construits par simple somme, et peuvent ne pas couvrir totalement le concept de défaveur. Le développement d'une méthodologie rigoureuse apparaît pertinent afin de s'assurer que l'indice est statistiquement justifié et constitue une bonne approximation du concept de défaveur.

Dans ce contexte, l'objectif majeur de cette étude est d'exposer une méthodologie basée sur des critères statistiques pour sélectionner des variables socioéconomiques en vue de créer un indice socioéconomique contextuel. Afin d'étudier sa généralisation et d'explorer comment la méthode de classification peut influer sur la mesure de l'association entre les inégalités sociales et la santé, elle a été appliquée à trois agglomération françaises aux profils contrastés.

2 Matériel et méthodes

Zone d'étude et échelle

Notre zone d'étude est composée de trois agglomérations françaises majeures : Lille Métropole, le Grand Lyon et l'unité urbaine d'Aix-Marseille. L'unité statistique est l'IRIS (Îlots Regroupés pour l'Information Statistique), une unité infra-communale définie par l'INSEE. Il s'agit de la plus petite unité administrative pour laquelle les données socioéconomiques et démographiques sont disponibles en France. L'IRIS compte en moyenne 2000 habitants et cette unité géographique est construite pour être aussi homogène que possible en termes de caractéristiques sociodémographiques. Les contours des IRIS ont été définis en prenant en compte le paysage urbain et les obstacles qui peuvent se présenter (grandes artères de trafic, espaces verts, plans d'eau et rivières). Les IRIS sont divisés en trois catégories : habitat, activité et divers. Les éléments de ces deux dernières catégories d'IRIS, aux profils particuliers, ont été systématiquement considérés comme des individus illustratifs dans les analyses statistiques.

Données socioéconomiques

Les données socioéconomiques proviennent du recensement national de 1999 et fournissent des dénombrements de la population, des ménages et des logements à l'échelle de l'IRIS dans tous les aspects sociaux, économiques et démographiques. En utilisant ces données brutes, 48 indicateurs ont été construits, à l'échelle de l'IRIS suivant les définitions de l'INSEE. Nous avons choisi ces indicateurs pour être représentatifs du concept de défaveur et similaires aux indicateurs les plus fréquemment utilisés dans la littérature. Ces 48 indicateurs concernent la famille, le foyer, le statut migratoire, la mobilité, l'emploi et revenu, l'éducation ou encore le logement. Certaines variables redondantes (fortement corrélées et représentant la même notion) ont été volontairement introduites (7 variables de chômage ; 3 variables sur les actifs) afin de déterminer celle qui représente le mieux la notion concernée.

Création de l'indice socioéconomique

Nous avons effectué une analyse factorielle multiple (AFM) en partitionnant l'ensemble des variables en plusieurs sous-ensembles, mais le choix du nombre de sous-ensembles et de leur composition peut prêter à discussion et nous n'avons finalement pas retenu cette analyse. Nous avons alors créé un indice socioéconomique en trois étapes.

Premièrement, une attention particulière a été accordée aux deux groupes de variables fortement corrélées. Nous avons appliqué une ACP sur chacun de ces groupes, le premier facteur est un facteur de taille et pour limiter le nombre de variables introduites dans la définition de

l'indice nous avons décidé de ne garder que la variable la plus corrélée avec ce facteur.

Deuxièmement, nous avons utilisé une ACP pour ne conserver dans l'indice que les variables les plus corrélées au premier facteur. Contrairement à la majorité des études développant une nouvelle méthode pour construire un indice socioéconomique contextuel, nous avons basé la sélection de variables sur l'utilisation d'un critère statistique plutôt qu'en se référant à la littérature.

Troisièmement, une dernière ACP a été réalisée intégrant les variables sélectionnées à l'étape précédente pour créer l'indice. Sous réserve que le premier axe de cette ACP puisse être interprété comme un « axe de défaveur » (ce qui est attendu étant donné que les variables sélectionnées sont des indicateurs de la défaveur), nous avons alors défini l'indice de défaveur comme étant la première composante, centrée (par construction) et réduite.

Nous avons choisi d'appliquer cette procédure sur chaque agglomération séparément (obtenant différents indices spécifiques à ces zones) et ensuite de construire un indice global sur les trois agglomérations rassemblées. Finalement, nous obtenons quatre indices différents construits avec la même méthodologie que nous souhaitons comparer.

Classification ascendante hiérarchique et seuils optimaux

Certaines applications de l'indice (étude d'associations non linéaires, représentations cartographiques, …) le nécessitant, nous nous sommes intéressés à la construction de catégories de défaveur.

Les quantiles sont fréquemment utilisés. Cependant, les classes créées ne vérifient aucun critère d'optimalité. Nous avons alors utilisé une classification ascendante hiérarchique (CAH) à partir des facteurs de l'ACP. Cependant, on utilise en général plusieurs facteurs de l'ACP pour effectuer la CAH. Or, dans ce travail, l'objectif est de créer des classes à partir d'un indice unidimensionnel construit à partir du premier facteur.

Suivant le nombre de classes souhaité, deux cas peuvent se présenter :

- Soit les classes construites par CAH se distribuent suivant plus d'une dimension (le second axe de l'ACP (ou plus) influençant par exemple la partition) ; alors, il n'est pas adéquat de construire une partition uniquement basée sur notre indice et nous devons réduire le nombre de classes.
- Soit les classes se distribuent suivant le premier axe de l'ACP (notre indice) seulement ; alors, nous déterminons des seuils (avec un algorithme itératif simple en faisant varier les seuils suivant les valeurs de l'indice et en conservant une partition qui donne le meilleur taux de concordance avec celle obtenue par CAH) permettant de créer des classes basées uniquement sur notre indice.

Comparaison entre indices et entre partitions

Afin de comparer statistiquement les différents indices quantitatifs obtenus, nous avons utilisé le coefficient de corrélation linéaire. Les partitions ont quant à elles été comparées deux à deux grâce au pourcentage de concordance entre partitions, c'est-à-dire le pourcentage d'IRIS dans la même classe pour les deux partitions (i.e. la diagonale de la matrice de confusion). La création des indices, des partitions, la détermination des seuils et les comparaisons ont été effectuées avec le logiciel R (R Development Core Team 2011) et le package FactoMineR (Lê et al. 2008).

3 Résultats

Création d'indices, sélections de variables et contributions

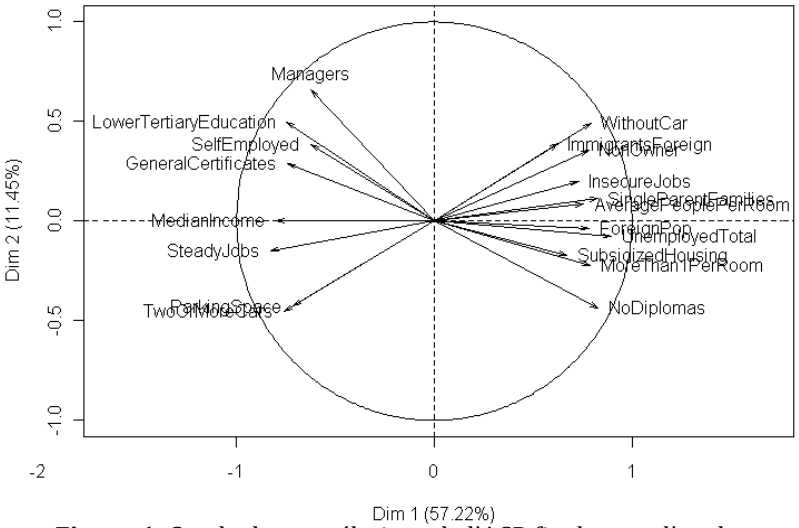

Figure 1. Cercle des corrélations de l'ACP finale pour l'analyse globale.

Table 1. Variance expliquée par les deux premiers axes des ACP finales.

	Lille	**Lyon**	**Marseille**	**Globale**
1er axe	60.73%	57.79%	57.29%	57.22%
2e axe	12.13%	16.71%	14.66%	11.45%

Pour chacun des quatre indices créés (un par agglomération et un global), le premier axe de l'ACP a pu être interprété comme un axe « de défaveur » (Figure 1) opposant des variables de défaveurs sociale et matérielle (chômage, familles monoparentales, surpeuplement, …) à des variables dites de « faveur » (revenu, emplois stables, haut niveau d'éducation, …). Le premier facteur des 4 ACP explique une part importante de la variance totale (Table 1).

Le nombre de variables sélectionnées, non défini *a priori*, est d'environ 20 pour chaque agglomération et recouvre les différents domaines connus du concept de défaveur. Ce nombre important, comparé à la plupart des autres indices, facilite une interprétation spatiale fine, très utile pour déterminer les cibles clés des actions locales des décideurs en santé publique.

Parmi les variables sélectionnées, 15 sont communes à tous les indices et contribuent pour plus de 77% au premier facteur. Ce résultat montre la stabilité de notre méthode et révèle qu'en dépit des différences entre les agglomérations, les mêmes variables expliqueraient une large part de la variabilité socioéconomique et les déterminants communs de la défaveur à l'échelle de l'IRIS. Cependant, l'ordre des contributions des variables est différent entre les agglomérations, bien qu'aucune ne soit très éloignée de la contribution moyenne. Ceci, ainsi que les variables propres à chaque agglomération, nous donne une indication sur la spécificité de la défaveur dans ces agglomérations. La méthodologie que nous proposons permet donc de construire un indice pouvant être appliqué localement.

Comparaison entre les indices

Nous avons comparé d'une part les indices de chaque agglomération avec l'indice global restreint à chacune respectivement et d'autre part notre indice avec ceux de Carstairs (égal à la somme des proportions de chômeurs, de foyers sans voiture, de logements surpeuplés et d'ouvriers) et Townsend, qui sont parmi les indices les plus utilisés dans la littérature, construits sur les mêmes zones.

Table 2. Coefficients de corrélation entre les indices

	Lille	Lyon	Marseille	Global
Global [a]	0.99	1	0.99	/
Carstairs	0.92	0.96	0.91	0.94
Townsend	0.98	0.94	0.96	0.96

[a] Lorsque l'on compare l'indice global avec l'indice d'une agglomération, l'indice global est restreint aux IRIS de l'agglomération.

En général, les corrélations entre agglomérations et indice global sont toutes supérieures à 0,9 (Table 2) et une association linéaire est clairement visible. De même, la corrélation entre notre indice et ceux de Carstairs et Townsend est toujours supérieure à 0.9, ce qui suggère que la notion mesurée par notre indice est proche du concept de défaveur mesuré par ces indices.

Comparaisons entre les classes

Un nombre de classes fréquemment utilisé en épidémiologie spatiale est cinq. Pour les 4 indices, la partition obtenue par CAH prend en compte les deux premiers axes de l'ACP pour construire les 5 classes de défaveur. Par conséquent, il semble que cinq soit un nombre de classes trop important si l'on souhaite n'utiliser que le premier facteur dans l'indice. Cela montre également que la partition en quintiles n'est pas optimale. En outre, la partition optimale apparaît être en trois classes (voir figure 2) et ces trois classes sont ordonnées suivant le premier axe de l'ACP. Nous avons alors déterminé des seuils optimaux (voir plus haut) permettant de créer des classes basées uniquement sur notre indice.

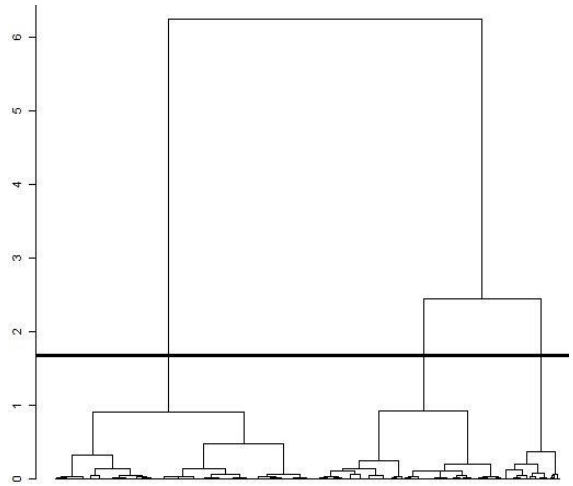

Figure 2. Dendrogramme de la CAH

Nous avons trouvé une très faible concordance entre les partitions en 5 classes par CAH et les partitions par quintiles (Table 4), ce qui est une conséquence du fait que la CAH n'utilise pas qu'un seul axe de l'ACP finale pour créer la partition. Les taux de concordance entre les partitions en trois classes par CAH et par tierciles sont entre 69 et 78%. Les taux de concordance entre les partitions en 3 classes créées par CAH et celles obtenues par seuils optimaux sont quant à eux tous supérieurs à 90%, ce qui confirme le fait que la partition en 3 classes est faite le long du premier axe.

Table 4. Taux de concordance entre les différentes techniques de classification et indices.

					Carstairs		Townsend	
	CAH (5) vs. quintiles	**CAH (3) vs. tierciles**	**CAH (3) vs. seuils**	**Seuils vs. tierciles**	**CAH (5) vs. quintiles**	**Seuils vs. tierciles**	**CAH (5) vs. quintiles**	**Seuils vs. tierciles**
Lille Métropole	41%	78%	98%	79%	38%	70%	42%	78%
Grand Lyon	48%	74%	93%	78%	47%	77%	40%	75%
Aix-Marseille	48%	69%	97%	67%	51%	67%	50%	69%
Globale	63%	71%	97%	72%	57%	70%	55%	71%

4 Conclusion

Dans cette étude, nous avons développé une procédure visant à créer des indices socioéconomiques, prolongeant un travail précédemment publié par l'équipe (Havard et al. 2008). Nous avons appliqué et validé cette méthode sur trois agglomérations françaises afin de démontrer sa généralisation possible sur le territoire français. Cette méthode permet de produire un indice socioéconomique statistiquement justifié et de fines possibilités d'interprétation. L'ensemble des variables sélectionnées montre un grand nombre de déterminants communs de la défaveur et identifie également quelques particularités de chaque zone. La comparaison des méthodes de classification montre également qu'une attention doit être portée sur cette partie afin d'obtenir des classes plus homogènes et de consolider les résultats.

Bibliographie

[1] Marmot, M. (2005) Social determinants of health inequalities. *Lancet*, 365(9464), 1099-1104.

[2] Chaix, B., Rosvall, M. & Merlo, J. (2007) Recent increase of neighborhood socioeconomic effects on ischemic heart disease mortality: a multilevel survival analysis of two large Swedish cohorts. *American Journal of Epidemiology*, 165(1), 22-26.

[3] Zeka, A., Melly, S.J. & Schwartz, J. (2008) The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in Eastern Massachusetts. *Environmental Health: A Global Access Science Source*, 7, 60.

[4] Braveman, P.A. et al. (2005) Socioeconomic status in health research: one size does not fit all. *JAMA: The Journal of the American Medical Association*, 294(22), 2879-2888.

[5] Galobardes, B. et al. (2006) Indicators of socioeconomic position (part 2). *Journal of Epidemiology and Community Health*, 60(2), 95-101.

[6] Krieger, N., Williams, D.R. & Moss, N.E. (1997) Measuring social class in US public health research: concepts, methodologies, and guidelines. *Annual Review of Public Health*, 18, 341-378.

[7] Morris, R. & Carstairs, V., 1991. Which deprivation? A comparison of selected deprivation indexes. *Journal of Public Health Medicine*, 13(4), 318-326.

[8] R Development Core Team (2011) R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria*. Available at: http://www.R-project.org

[9] Lê, S., Josse, J. & Husson, F. (2008) FactoMineR: An R package for multivariate analysis. *Journal of statistical software*, 25(1), 1–18.

[10] Havard, S. et al. (2008) A small-area index of socioeconomic deprivation to capture health inequalities in France. *Social Science & Medicine (1982)*, 67(12), 2007-2016.