Une approche par classification de variables à l'aide de la méthode 'ClustOfVar' pour analyser la qualité de vie à l'échelle communale

A. Labenne[(a)], V. Kuentz-Simonet[(a)], M. Rambonilaza[(a)], M. Chavent[(b)], J. Saracco[(b)]

L'analyse de la **qualité de vie** repose sur des domaines bien définis. Nous disposons d'un total de 56 variables, décrivant 303 communes de la façade atlantique, qui peuvent être associées à ces différents domaines. Le but ici est de **réduire la dimension** des données afin de redéfinir les domaines importants de la qualité de vie en réduisant leur nombre et en évitant les redondances d'information. Pour cela, nous utilisons une approche par classification de variables. En effet, en réorganisant les variables en classes homogènes, nous allons construire des **variables synthétiques** sans imposer de contraintes d'orthogonalité. Par la suite, nous effectuerons une **typologie des communes** à l'aide des variables synthétiques obtenues.

L'approche par classification de variables

- Elle 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 C_k est une mesure d'adéquation entre les variables de la classe et la **variable synthétique quantitative** de la classe, notée $\boldsymbol{y_k}$. Elle est définie par : $H(C_k) = \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{x}_j,\mathbf{y}_k} + \sum_{\mathbf{z}_j \in C_k} \eta^2_{\mathbf{y}_k|\mathbf{z}_j}$

Où r² désigne la corrélation de Pearson au carré entre $\boldsymbol{y_k}$ et la variable quantitative $\boldsymbol{x_j}$ et η² désigne le rapport de corrélation entre $\boldsymbol{y_k}$ et la variable qualitative $\boldsymbol{z_j}$.

- La variable synthétique quantitative $\boldsymbol{y_k}$ est la variable « la plus liée » aux variables de la classe au sens du critère H qu'elle maximise.

ClustOfVar sur les données de qualité de vie

Les 12 domaines de la qualité de vie dans le Système Européen des Indicateurs Sociaux	Thèmes disponibles Bases de données à l'échelle communale France (INSEE)	Nombre de variables : Variables quantitatives / Variables qualitatives
Population	Population	1
Conditions familiales	Population	5
Conditions de logement et accès aux services	Conditions de vie - société	17 + 9
Education	Enseignement - Education	2
Conditions d'emploi	Travail - Emploi	15
Revenu et niveau de vie	Revenus - Salaires	1
Sécurité sociale (accès aux soins)	Santé	2
Environnement	Occupation du sol[(a)]	4
Santé / Transport / Loisirs / Participation politique et liens sociaux	NON DISPONIBLE	NON DISPONIBLE

[(a)] : Provient de la base Corine Land Cover

ClustOfVar

A l'aide du dendrogramme, on décide de retenir la partition en 4 clusters de variables

Variable synthétique n°1			Variable synthétique n°2		
Variables	Squared loadings	r(va.synth,va.quanti) η²(va.synth,va.quali)	Variables	Squared loadings	r(va.synth,va.quanti) η²(va.synth,va.quali)
Res_Princip_Maison	0,87	-0,94	Non_Diplomés	0,71	0,85
Res_Princip_Appart	0,86	0,93	Profession_Interm	0,64	-0,8
Territoires_Artificialisés	0,78	0,88	Diplomés_Bac_ou_plus	0,58	-0,76
Densite_Population	0,68	0,83	Logmt_construc_75_89	0,52	-0,72
Res_Princip_HLM	0,59	0,77	Cadres_et_Intellectuels	0,51	-0,71
Creche	0,55	0,55	Agriculteurs_Exploitants	0,47	0,69
HalteGarderie	0,55	0,55	Territoires_Agricoles	0,43	0,66
Logmt_construc_49_74	0,55	0,74	Logmt_construc_ap_90	0,35	-0,59
Variable synthétique n°3			**Variable synthétique n°4**		
Variables	Squared loadings	r(va.synth,va.quanti) η²(va.synth,va.quali)	Variables	Squared loadings	r(va.synth,va.quanti) η²(va.synth,va.quali)
Menage_couple_avec_enfants	0,68	-0,82	Pharmacie	0,77	0,77
Emploi_dans_Dpt	0,68	-0,82	Medecin_Gené	0,74	0,74
Retraités	0,67	0,82	Boulangerie	0,73	0,73
Emploi_Commune_residence	0,61	0,78	CafBoissons	0,61	0,61
Emploi_Commune_même_zone	0,38	-0,62	Restaurant	0,59	0,59
Menage_couple_sans_enfants	0,32	0,57	Poste	0,59	0,59
Emploi_Commune_même_unité_urb	0,3	-0,55	Carburant	0,56	0,56
Ouvriers_Employés	0,27	-0,52	BanqueCE	0,53	0,53

Interprétation des variables synthétiques

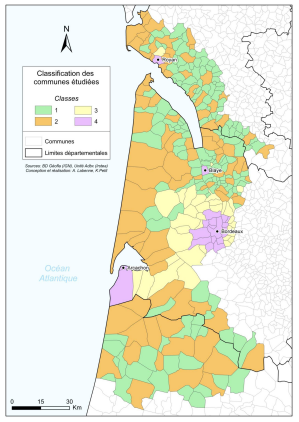
- Le **cluster 1** représente les conditions de logement des individus.
- Le **cluster 2** représente les niveaux de diplôme et les catégories socio-professionnelles (conditions de revenus).
- Le **cluster 3** représente les différentes situations familiales ainsi que les conditions d'emploi.
- Le **cluster 4** représente l'accès aux services publics et aux autres services.

Typologie des communes à l'aide des variables synthétiques

- La typologie des communes a été réalisée à l'aide d'une classification ascendante hiérarchique (CAH).
- Au vu du dendrogramme, il a été décidé de retenir la partition en 4 classes d'individus.
- Voici la caractérisation des classes par les variables synthétiques issues de ClustOfVAR (par construction les variables synthétiques sont centrées, leur moyenne totale est donc nulle) :

CLASSES	Variables	V.test	Distribution intra classe		Distribution totale
			Moyenne	Ecart-type	Ecart type total
CLASSE 1	Var.synth.1	9,63	1,55	0,47	2,98
	Var.synth.2	8,76	1,01	1,52	2,14
	Var.synth.4	-15,25	-2,21	1,45	2,69
CLASSE 2	Var.synth.4	10,37	2,51	0,95	2,69
	Var.synth.3	7,45	1,46	1,51	2,18
CLASSE 3	Var.synth.4	4,85	2,11	1,33	2,69
	Var.synth.2	-9,57	-3,32	1,66	2,14
	Var.synth.3	-10,36	-3,65	1,39	2,18
CLASSE 4	Var.synth.4	5,61	3,36	0,27	2,69
	Var.synth.3	-2,9	-1,4	2,2	2,18
	Var.synth.2	-4,72	-2,25	1,36	2,14
	Var.synth.1	-14,04	-9,31	3,04	2,98

Interprétation des classes de communes



- **La classe 1** représente des communes de résidence pour l'accès aux maisons individuelles, des territoires agricoles habités par une population peu diplômée et d'agriculteurs, ces communes n'ont que très peu accès à l'ensemble des services (principalement des communes de l'estuaire).
- **La classe 2** concerne des communes habitées par des retraités, les emplois sont souvent à l'échelle de la commune et l'accès aux services est important (le plus souvent, ce sont des communes de la façade littorale).
- **La classe 3** contient des communes peu denses avec des accès aux services très importants (ce sont des communes périphériques des centres urbains).
- **La classe 4** représente des communes denses et avec un très bon accès aux services (communes des centres urbains).

Références bibliographiques

- Chavent M., Kuentz V., Liquet B., Saracco J., ClustOfVar: An R Package for the Clustering of Variables, *Journal of Statistical Software.* **Vol. 50**, pp. 1-16.
- Heinz-Herbert Noll, Towards a European System of Social Indicators: Theoretical Framework and System Architecture, Social Indicators Research, 2002, **vol. 58**, issue 1, pages 47-87

[(a)] **Irstea, UR ADBX Aménités et dynamiques des espaces ruraux,** 50 avenue de Verdun Gazinet Cestas, F-33612, France.

[(b)] **Inria Bordeaux Sud Ouest, Équipe CQFD Contrôle de qualité et fiabilité des données,** 351 cours de la Libération, 33405 Talence cedex, France.

www.irstea.fr

Workshop StatLearn – 8 et 9 avril 2013 - Bordeaux