Analyse en composante principales d'un tableau de distributions macroéconomiques.

Sun Makosso-Kallyth

MDC Pain Centre, McMaster University 1280 Main Street West
Hamilton, Ontario L8S 4K1 Canada.

makossok@macmaster.ca

Résumé. Cet article présente l'application de deux extensions de l'analyse en composantes principales à un tableau de distributions macroéconomiques. Les méthodes présentées s'inscrivent dans le cadre de l'analyse des données symbo-liques. Elles étendent l'analyse en composantes principales aux variables sym-boliques de type histogramme. Dans la première méthode on détermine les moyen-nes des histogrammes, on effectue ensuite une ACP classique du tableau des moyennes et on projette en éléments supplémentaires les hypercubes obtenus à partir de la transformation des histogrammes en intervalles. En outre, pour amé-liorer le codage des modalités des variables de type histogramme, on présente un nouveau codage basé sur les scores de Ridit. Dans la seconde méthode en revanche, on détermine les quantiles, on définit par la suite une mesure de cor-rélation d'histogrammes à partir des quantiles qui se correspondent. On procède ensuite par la détermination des vecteurs propres de la matrice de corrélation et la projection des quantiles en éléments supplémentaires. Nous appliquons les deux méthodes à des données macroéconomiques et montrons leur intérêt en analyse des données exploratoire.

Mots clés : variable symbolique de type histogramme, analyse en composantes principales, quantiles, scores de Ridit.

1 Introduction

L'objet de cet article est celui de comparer l'application de deux extensions de l'analyse en composantes principales (ACP) à un tableau de distributions macroéconomiques. L'objet de l'ACP d'un tableau de distributions dans un tel contexte est celui d'expliquer les relations entre des variables qui décrivent de grands agrégats économiques. On peut par exemple s'intéresser à la description du profil démographique et environemental (niveaux d'émission de Gaz à effet de serre, nombre d'enfants par femme, etc.) de certaines zones géographiques telles que l'Afrique, l'Amérique du Nord et l'Europe. Dans une telle optique, si on souhaite comprendre les disparités socioéconomiques entre ces zones géographiques, on peut recourir à l'analyse de données symboliques. En effet, les récentes contributions en analyse des données symboliques ont permis d'énormes avancées en analyse des données exploratoires. Suite aux travaux de (Diday (1988) et Diday (1989)), l'analyse des données symboliques (ADS) a connu d'importants bouleversements au point de devenir une discipline à part entière en apprentissage

Analyse en composante principales d'un tableau de distributions macroéconomiques.

TABLE 1 – *Exemple de variable symbolique de type histogramme*

Region	PIB par habitant			Taux mortalité	
Modalité ξ_j	$\leq$ 1 k\$	$]1, 20]$	> 20	$\leq$ 0.10	> 0.10
Afrique	0.340	0.660	0.000	0.245	0.755
Alena	0.000	0.333	0.667	1.000	0.000
AsieOrientale	0.067	0.801	0.133	1.000	0.000
Europe	0.000	0.322	0.677	0.742	0.258

statistique (Cf Verde et Diday 2014). Les méthodes développées dans le cadre de l'ADS (voir Bock et Diday (2000), Billard et Diday (2006), Diday et Noirhomme-Fraiture (2008)) sont complémentaires à l'analyse de données classiques et elles se prêtent bien aux données complexes ayant plusieurs niveaux d'analyse. Dans le cas où l'on disposerait par exemple d'un tableau de données sur la situation macroéconomique de pays (unité de base), si on s'intéresse à la situation macroéconomique des espaces géographiques dont ces pays font partie (Union Européenne , Union Afrique, ALENA, etc.), en définissant comme nouvelle unité d'analyse la zone géographique, il est par exemple possible de définir (pour chaque variable) la distribution empirique de chaque zone géographique et d'appliquer par la suite une analyse multivariée de ces distributions. L'on peut par exemple recourir à l'analyse en composantes principales de variables symboliques de type histogramme. Plusieurs approches ont d'ailleurs été proposées à cet égard (Nagabhushan et Kumar (2007), Rodriguez et al. (2000), Makosso-Kallyth et Diday (2012), Verde et al. (2015), Ichino (2011), Diday (2013), Makosso-Kallyth (2015)). Dans cet article nous appliquons et comparons deux extensions ou adaptations de l'ACP aux variables symboliques de type histogramme. Pour ce faire, nous présentons premièrement la théorie inhérente à ces approches. Nous appliquons ensuite ces deux approches à un tableau de distributions macroéconomiques et comparons en dernier lieu ces deux méthodes.

2 Approache basée sur l'ACP des barycentres

Dans cette section, nous présentons l'approche proposée par Makosso-Kallyth et Diday (2012). Elle considère la moyenne (moment d'ordre 1) comme le principal élément caractéristique d'une distribution. Cette approche effectue premièrement le codage des modalités des variables. Elle determine ensuite les moyennes des variables de type histogramme et effectue une ACP de ces moyennes. Pour la représentation de la dispersion des variables compte tenu de leur nature symbolique, l'approche proposée par Makosso-Kallyth et Diday (2012) transforme les histogrammes en intervalles via l'inegalité de Tchebychev. Ensuite, elle projette en éléments supplémentaires les hypercubes induits par les intervalles sur les axes principaux de l'ACP des moyennes.

2.1 Notations

Soient n le nombre d'individus, p celui des variables, et m_j celui des modalités d'une variable symbolique de type histogramme Y_j ; Y_j est telle que $\mathsf{Y}_j = \{\xi_j, \mathsf{H}_{ij}\}$. On suppose également que : pour $i = 1, \ldots, n$; $j = 1, \ldots, p$ et $k = 1, \ldots, m_j$; n représente le nombre d'individus ; p celui des variables et m_j celui des modalités d'une variable de type histogramme Y_j ; Pour une valeur donnée de j, une variable symbolique de type histogramme Y_j de façon générale est telle que $\mathsf{Y}_j = \{\xi_j; \mathsf{H}_{.j}\}$ où $\xi_j = \left(\xi_j^{(1)}, \ldots, \xi_j^{(m_j)}\right)$ est le vecteur contenant les modalités de la variable Y_j, $\mathsf{H}_{.j}$ est le vecteur contenant les fréquences relatives H_{ij}. Les $\mathsf{H}_{\mathrm{ij}}{}^{(k)}$ vérifient la relation $\sum_{k=1}^{m_j} \mathsf{H}_{ij}^{(m_j)} = 1$. Dans la table 1, on a par exemple Y_1 (la variable PIB) qui est telle que $\mathsf{Y}_1 = \{\xi_1; \mathsf{H}_{.1}\}$ avec $\xi_1 = (]-\infty, 1];]1, 20];]20; +\infty])$ et

$$\mathsf{H}_{.1} = \begin{pmatrix} \mathsf{H}_{11} \\ \mathsf{H}_{21} \\ \mathsf{H}_{31} \\ \mathsf{H}_{41} \end{pmatrix}$$

avec $\mathsf{H}_{11} = (0.340, 0.660, 0)$, $\mathsf{H}_{21} = (0, 0.333, 0.667)$, $\mathsf{H}_{31} = (0.0067, 0.801, 0.133)$, $\mathsf{H}_{41} = (0, 0.322, 0.677)$.

2.2 Codage des modalité des variables de type histogramme.

L'objet du codage des modalités des variables est celui d'assigner des valeurs numériques appelées scores aux modalités des variables. Pour ce faire, Makosso-Kallyth et Diday (2012) proposent deux types de codage intitulés *codage paramétrique* et *codage non paramétrique*. Nous présentons également un codage basé sur *les scores de Ridit* bruts, standardisés et normalisés.

2.2.1 Codage paramétrique des modalités des variables.

Soit $\mathcal{D}_j = (\beta_{1_j}, \beta_{2_j})$ le domaine contenant l'ensemble des valeurs possibles prises par les modalités de la variable Y_j. Dans la table 1 la plus petite valeur possible prise par la modalité PIB (Y_1) est $\beta_{1_1} = 0$. En revanche, $\beta_{2_1} = +\infty$ car β_{2_1} la valeur maximale d'une modalité du PIB n'est majorée par aucune valeur. Dans la table 1 on a par exemple $Y_{11} = \{\xi_1, \mathsf{H}_{11}\}$ avec $\xi_1 = \{]-\infty, 1],]1, 20],]20, +\infty[\}$; $\mathsf{H}_{11} = (0.340; 0.660; 0.000)$. Ensuite, Makosso-Kallyth et Diday (2012) procède comme suit (voir aussi Makosso-Kallyth et Diday (2010)) :

1. Pour tout j fixé, $\delta_j = \inf_{k_j=1,\ldots,m_j} L_{k_j}$, L_{k_j} étant la longueur des intervalles des modalités $\xi_j^{(k_j)}$. Si certaines modalités $\xi_j^{(k_j)}$ sont de longueur infinie, i.e., si elles sont de la forme $I =]-\infty, a_j]$ ou de la forme $J =]b_j, +\infty[$,
alors on remplace I par $I' =]e_j, a_j]$ où

$$e_j = \begin{cases} \beta_{1_j} & \textit{si } a_j - \delta_j < \beta_{1_j} \\ a_j - \delta_j & \textit{sinon} \end{cases} .$$

De même on remplace J par $J' =]b_j, h_j]$ avec

$$h = \begin{cases} \beta_{2_j} & \textit{si } b_j + \delta_j > \beta_{2_j} \\ b_j + \delta_j & \textit{sinon} \end{cases} .$$

Dans la table 1 par exemple, la modalité $\xi_1^{(2)} =]1, 20]$ a la plus petite longueur $L_{2_1} = 19$. Par conséquent on remplace $\xi_1^{(1)}$ par $\xi_1^{'(1)} =]\max(1 - 19, 0), 1] =]0, 1]$ et $\xi_1^{(3)}$ par $\xi_1^{'(3)} =]20, \min(20 + 19, +\infty)] =]20, 39]$.

2. Si les modalités des différentes variables de type histogramme en jeu n'ont pas la même unité de mesure, on remplace chaque intervalle $]a'_j, b'_j]$ par un intervalle ajusté de la forme $]a'_j/(b'_j - a'_j); b'_j/(b'_j - a'_j)]$.

Au niveau de l'affectation des scores des modalités, le codage paramétrique assigne à une modalité un vecteur de scores $s_j = (s_j^{(1)}, \ldots, s_j^{(m)})$ où $s_j^{(kj)}$ est égal au centre des intervalles ajustées pour $k_j = 1, \ldots, m_j$.

Le codage paramétrique a cependant quelques limites. Premièrement, son application requiert une connaissance parfaite du domaine macroéconomique. Ensuite, le choix des centres des classes des intervalles de longueur extrême ou infinie ne devrait se faire de façon arbitraire. Au vu de ces limites, il est préférable d'appliquer d'autres codages alternatifs tel que le codage non paramétrique.

2.2.2 Codage non paramétrique des modalités des variables.

Le codage non paramétrique utilise comme scores des modalités, le rang qui leur est associé. Si on se refère à la table 1 par exemple, les scores des modalités des classes seront

$$s_j^{(1)} = 1, s_j^{(2)} = 2, \ldots, s_j^{(mj)} = m_j.$$

Dans le cas la variable PIB, on a comme scores : $s_1^{(1)} = 1,\ s_1^{(2)} = 2,\ s_1^{(3)} = 3$. Pour la variable taux de mortalité, on a par contre $s_2^{(1)} = 1,\ s_2^{(2)} = 2$.

Le codage non paramétrique est d'un usage simple. Il est adapté à des modalités ordonnées équidistantes. Cependant, si pour une variable Y_j les longueurs des modalités $\xi_j^{(k)}$ ne sont pas égales, ou si les écarts entre deux modalités consécutives sont distincts, le codage non paramétrique pourrait ne pas s'avérer réaliste. Dans le cas d'une variable telle que le statut social d'une personne seule, si on suppose que les modalités de cette variable sont $\xi_j^{(1)}$=classe pauvre (salaire mensuel net en euros =< 729), $\xi_j^{(2)}$= classe populaire (salaire mensuel $\in]729, 1183]$), $\xi_j^{(3)}$=classe moyenne (salaire mensuel $\in]1183, 2177]$), $\xi_j^{(4)}$= classe aisée (salaire mensuel $\in [2177, 2917[$, $\xi_j^{(5)}$= Riche (salaire mensuel >=2917), [1] l'écart entre une personne riche

[1] cf. http ://www.lefigaro.fr/social/2014/04/16/09010-20140416ARTFIG00110-tes-vous-riche-pauvre-ou-appartenez-vous-a-la-classe-moyenne.php

et une personne de la classe aisée n'est pas forcément similaire à l'écart entre une personne pauvre et une personne de la classe populaire. L'utilisation du codage non paramétrique est dans ce cas inapproprié. On présente dans la section suivante un codage basé sur l'utilisation des scores de Ridit.

2.2.3 Codage à partir des scores de Ridit.

Les scores de Ridit ont été introduits par Bross (1958). Ils ont à la base une interprétation probabiliste (probabilité qu'une variable aléatoire soit inférieure à une valeur de référence). En analyse de données qualitative (ou 'catégorique'), ils sont également utilisés comme scores des modalités de variables ordinales (cf. Agresti (2002), Mantel (1979), Donaldson (1998)). Si on considère par exemple un vecteur de fréquences relatives $\mathsf{H}_{ij} = (\mathsf{H}_{ij}^{(1)}, \mathsf{H}_{ij}^{(2)}, \ldots, \mathsf{H}_{ij}^{(m)})$, les socres de Ridit r_k sont tels que :

$$r_k = 0.5\mathsf{H}_{ij}^{(k)} + \sum_{m<k} \mathsf{H}_{ij}^{(m)}. \tag{1}$$

Pour adapter l'application des scores de Ridit aux modalités des variables de type histogramme, on propose premièrement la détermination du vecteur moyen $\overline{\mathsf{H}}_{.j} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{H}_{ij}$. Ensuite, les scores de Ridit bruts des variables de type histogrammes sont définis de la manière suivante :

$$s_j^{(k)} = 0.5\overline{\mathsf{H}}_{.j}^{(k)} + \sum_{m_j<k} \overline{\mathsf{H}}_{.j}^{(m_j)}. \tag{2}$$

Dans le cas de la table 1 par exemple, les vecteurs moyens $\overline{\mathsf{H}}_{.1}$ et $\overline{\mathsf{H}}_{.2}$ associé a cette table sont respectivement $\overline{\mathsf{H}}_{.1} = (0.102, 0.529, 0.369)$ et $\overline{\mathsf{H}}_{.2} = (0.747, 0.253)$. Les scores de Ridit brutes quant à eux sont $s_1^{(1)} = 0.051,\ s_1^{(2)} = 0.102 + \frac{0.529}{2} = 0.366,\ s_1^{(3)} = 0.102 + 0.529 + \frac{0.369}{2} = 0.816$ pour la variable PIB, et $s_2^{(1)} = 0.373,\ s_2^{(2)} = 0.873$ pour le taux de mortalité. Pour tenir compte des différences entre variables, nous définissons également les scores de Ridit standardisés

$$s_j'^{(k)} = \frac{s_j^{(k)} - \mu_{s_j}}{\sigma_{s_j}}, \tag{3}$$

où μ_{s_j} et σ_{s_j} représentent respectivement les moyennes et variances empiriques du vecteur des scores bruts $s_j = (s_j^{(1)}, \ldots, s_j^{(m_j)})$. Les scores de Ridit standardisés sont $s_1'^{(1)} = -0.937,\ s_1'^{(2)} = -0.116,\ s_1'^{(3)} = 1.053$ pour la variable PIB. De même, pour le taux de mortalité on a $s_2'^{(1)} = -0.707$ et $s_2'^{(2)} = 0.707$. La standardisation a néanmoins tendance à assigner des scores élevés en valeurs absolues aux modalités extrêmes et des valeurs proche de zéros aux modalités intermédiaires. Il est également possible de définir des scores de Ridit normalisés de sorte que :

$$s_j''^{(k)} = \frac{s_j^{(k)}}{\sum_{k=1}^{m_j} s_j^{(k)}}. \tag{4}$$

Analyse en composante principales d'un tableau de distributions macroéconomiques.

2.3 ACP des centres et representation des individus.

Après le codage des modalités des variables, Makosso-Kallyth et Diday (2012) déterminent les moyennes ou barycentres g_{ij} de chaque histogramme Y_{ij}. Ces valeurs moyennes sont telles que :

$$g_{ij} = \sum_{k_j=1}^{m_j} s_j^{(kj)} \mathsf{H}_{ij}^{(k_j)}. \tag{5}$$

Une ACP classique du tableau des moyennes $g = (g_{ij})_{i=1,\ldots,n\ ;\ j=1,\ldots,p}$ est ensuite appliquée. Soient $u_\alpha,\ \alpha = 1,\ldots,p$ les p axes principaux de l'ACP du tableau des moyennes g. Pour représenter les individus compte tenu de leur nature symbolique, Makosso-Kallyth et Diday (2012) transforment les histogrammes en intervalles via l'inégalité de Tchebytchev. Ainsi, pour toute variable aleatoire X_j de moyenne empirique $\overline{X}_j$ et d'écart type empirique σ_{Xj} et pour tout nombre $t \geq 0$, la proportion d'information comprise entre l'intervalle $[\overline{X_j} - t\sigma(X_j), \overline{X}_j + t\sigma(X_j)]$ est supérieure ou égale à $1 - \frac{1}{t^2}$. Autrement dit, si P est une mesure de probabilité on a :

$$\mathsf{P}\left(X_j \in \left[\overline{X_j} - t\sigma_{X_j}, \overline{X_j} + t\sigma_{X_j}\right]\right) \geq 1 - \frac{1}{t^2} \tag{6}$$

Ainsi, pour une valeur donnée de t, Makosso-Kallyth et Diday (2012) transforment chaque fréquence H_{ij} en intervalle $[c_{ij}, d_{ij}]$ via l'inegalité de Tchebychev. Ensuite, Makosso-Kallyth et Diday (2012) construisent les hypercubes associés aux concepts ou individus symboliques. Si on suppose que le nombre de variables est égale à p, chaque hypercube a dans ce cas 2^p sommets. Soit $\mathcal{M}_i$ l' hypercube associée au $i$ème concept. Pour $p = 2$ par exemple, on associe au concept i l'intervalle $([a_{i1}, b_{i1}], [a_{i2}, b_{i2}])$. L'hypercube $\mathcal{M}_i$ est dans ce cas

$$\mathcal{M}_i = \begin{bmatrix} a_{i1} & a_{i2} \\ a_{i1} & b_{i2} \\ b_{i1} & a_{i2} \\ b_{i1} & b_{i2} \end{bmatrix}.$$

Makosso-Kallyth et Diday (2012) projettent chaque hypercube $\mathcal{M}_i$ (matrice d'ordre $2^p \times p$) sur u_α le αème axe factoriel et representent ainsi en 2D dimension les individus sous la forme de rectangle (à partir des $\min$ et des $\max$ des projection des hypercubes).

L'approche proposée par Makosso-Kallyth et Diday (2012) se focalise essentiellement sur la moyenne des histogrammes. Dans le cas de distributions asymétriques, il est préférable de recourir à d'autres éléments caractéristiques des distributions tels que les quantiles par exemple. Par ailleurs, le choix du codage a une incidence sur les résultats finaux. C'est dans cette optique que nous présentons dans la section suivante une approche basée sur la corrélation moyenne des quantiles qui se correspondent.

3 Approche basée sur corrélation moyenne des quantiles qui se correspondent.

L'approche basée sur la corrélation des quantiles qui se correspondent (voir Makosso-Kallyth (2015)) nécessite premièrement le choix de m le nombre commun de quantiles de chaque variable de type histogramme. C'est ainsi que l'on obtient Q la table d'ordre $(n \times m) \times p$ de sorte que $\mathsf{Q} = (\mathsf{Q}_1, \ldots, \mathsf{Q}_p)$, $\mathsf{Q}_j = \left(\mathsf{Q}_j^{(1)}, \ldots, \mathsf{Q}_j^{(m)}\right)$ et $\mathsf{Q}_j^{(k)}$ est la $n \times p$ table de quantiles. La correlation entre deux variables symboliques de type histogramme est déterminée par la correlation moyenne des quantiles qui se correspondent via la relation ci-dessous

$$R^*_{Y_1,Y_2,\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}} = \frac{exp(2\overline{Z_{Y_1,Y_2,\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}}) - 1}{exp(2\overline{Z_{Y_1,Y_2,\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}}) + 1}. \tag{7}$$

où

$$\overline{Z_{Y_1,Y_2,\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}} = \frac{1}{m}\sum_{k=1}^{m} Z_{Y_1^{(k)},Y_2^{(k)},\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}. \tag{8}$$

$$\text{et } Z_{Y_1^{(k)},Y_2^{(k)},\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}} = arctanh(R_{Y_1^{(k)},Y_2^{(k)},\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}). \tag{9}$$

où $R_{Y_1^{(k)},Y_2^{(k)},\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}$ représente la corrélation de Pearson des quantiles qui se correspondent. Pour tout $R_{Y_1^{(k)},Y_2^{(k)},\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}} = 1$ (resp. -1), on supposera que $R^*_{Y_1,Y_2,\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}} = 1$ (resp. -1) . La relation (7) induit une matrice de corrélation $R^*_{\mathsf{Q}_j^{(1)},\ldots,\mathsf{Q}_j^{(m)}}$. Les vecteurs propres u_α de cette matrice de corrélation font office d'axes principaux. Pour la représentation des concepts sur les axes factoriels, on projette les quantiles $\mathsf{Q}^{(1)} \ldots \mathsf{Q}^{(m)}$ sur les axes factoriels u_α et on représente les rectangles ou les enveloppes convexes des quantiles projettés. De même, pour les cartes de corrélation, on détermine la corrélation au sens de (7) entre les $\mathsf{Q}^{(1)} \ldots \mathsf{Q}^{(m)}$ et leur projection sur les axes factoriels.

4 Application.

On applique les deux méthodes précédemment décrites à un jeu de données de la Banque Mondiale. Ces données portent sur 10 variables (voir annexe). Il s'agit du *PIB* par habitant, du niveau de la sous alimentation (*Sous*), du niveau de la consommation d'électricité en Kwh par personne (*Elec*), des émissions de gaz dans l'atmosphère (*gaz*), du niveau de la population (*Population*), du niveau d'investissement (*Invest*), de l'Indice de développement humain *IDH*, du nombre moyen d'enfants par femme (*Fecondite*), du nombre de téléphones portables par personne (*Telephone*), de la croissance démographique (*CROIS*), du taux de mortalité (*Mortalite*), des dépenses en matière de santé *Sante* (cf. table 2 et les tables en annexe). Les informations sont disponibles par continent sous la forme d'histogrammes à cinq modalités. Les continents représentent les individus symboliques.

Analyse en composante principales d'un tableau de distributions macroéconomiques.

4.1 Application de l'approche basée sur l'ACP des barycentres.

4.1.1 Utilisation du codage non paramétrique.

Nous appliquons premièrement la méthode de Makosso-Kallyth et Diday (2012) sans recourir à la transformation angulaire dans le prétraitement des variables. Nous utilisons le codage non paramétrique. Les figures 1 et 2 contiennent le plan des projections obtenus à partir des hypercubes et la carte des corrélations au sens de Pearson entre les moyennes des variables $g_{.j},\ j = 1, \ldots, 10$ et leurs composantes principales. Dans la figure 1, les individus sont représentés sous la forme de rectangles. Cela permet de mettre en évidence leur dispersion. La dispersion dont il est question ici correspond à la variabilité de chaque concept par rapport à la valeur moyenne de chaque variable. La variabilité des deux premiers axes factoriels est respectivement de 66.79% and 16.76%. Le premier axe du plan de projection met en évidence l'opposition entre d'une part les régions développées (Europe et Aléna) et l'autre les régions sous-développées (Afrique). L'axe 1 met également en évidence une sorte de troisième groupe constitué par l'Asie Orientale, les états de l'ex-URSS, l'amérique du sud et du centre et le proche et moyen Orient. Quant au deuxième axe factoriel, il oppose d'une part les pays de l'ex union sovietique et de l'autre ceux proche orient.

TABLE 2 – *Variables sélectionnées.*

Variables symboliques	Catégorie
PIB par habitant$	$<$ 1 k$, [1k$;5k$] ; [5k$;10k$] ; [10k$,20k$] ; $> 20k$$
% de la sous alimentation	$< 3\%$; [3%, 10 %] ; [10 % ;25 %] ; [25% ;35 %] ; $> 35\%$
Croissance de la Population$	$< 0.0\%$; [0 ;1] ; [1 ;2] ;[2 ; 4] ; $>= 4$
IDH (indice de développement Humain)	< 0.5 ; [0.5 ; 0.6] ; [0.6 ;0.7] ; [0.7 ;0.8] ; $>= 0.8$
Taux de mortalité	$< 0.5\%$; [0.5% ; 1%] ; [1% ;1.5%] ; [1.5% ;2.0%] ; $>= 2.0\%$
Population en Millions	$<$ 1 M ; [1 ; 5] M ; [5 ;10] M ; [10 ;100] M ; $>=$ 100 M
Electricité kwh/pers	< 500 ; [500 ; 1000] ; [1000 ;5000] ; [5000 ;15000] ; $>= 15000$
GAZ kt /hab	< 1 ; [1 ; 2] ; [2 ;5] ; [5 ;10] ; $>= 10$
Fecondité	<2 ; [2,3[; [3 ; 5[; [5 ;6[; ≥ 6 ;
Depense Santé en % du PIB	$< 3\%$; [3% ; 4%] ; [4% ;6%] ; [6% ;8%] ; $>= 8\%$

TABLE 3 – *Corrélation entre les moyennes des variables et les 5 premières composantes prin-cipales.*

Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
PIB	-0.73	0.55	-0.20	-0.30	-0.13
SousAl	0.96	-0.19	0.18	-0.06	-0.04
Electricite	-0.88	0.07	-0.45	0.05	-0.02
GAZEmission	-0.89	-0.14	-0.16	-0.37	-0.04
Population	0.23	-0.20	-0.17	0.70	-0.60
Fecondite	0.83	-0.33	-0.34	-0.19	-0.14
CroissPopulat	0.10	-0.93	-0.33	-0.01	0.07
IDH	0.19	0.06	0.92	-0.29	0.18
TauMortalite	0.88	0.35	-0.30	0.01	0.13
DepSante	-0.82	-0.37	0.01	0.05	0.36

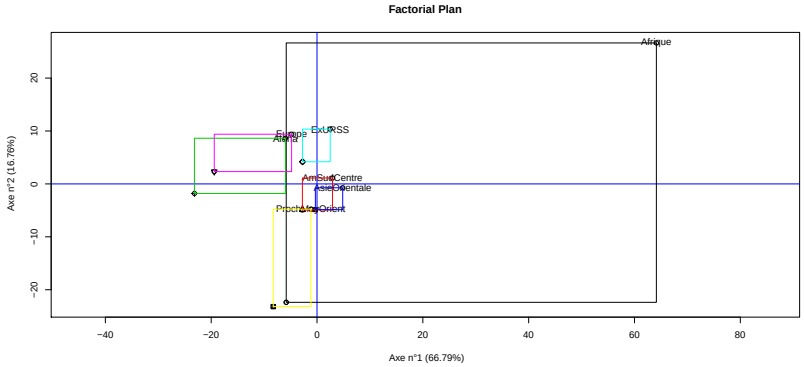

FIGURE 1 – *Plan des projections de l'ACP des moyennes des variables.*

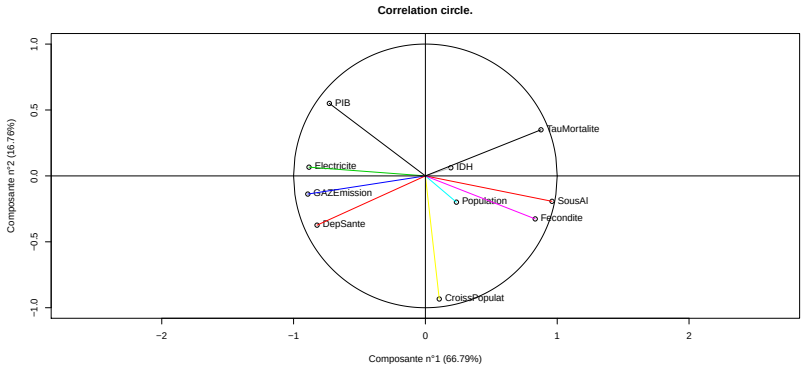

FIG. 2 – *Carte des Corrélations de l'ACP des moyennes des variables.*

La figure 2 représente la carte des corrélations entre les moyennes des variables et leurs composantes principales. La table 3 donne quelques détails à cet égard. Les valeurs moyennes des variables Sous alimentation, Taux de mortalité, Fécondité sont positivement corrélées à la première composante principale. Les valeurs moyennes de PIB, Électricité, Émission de GAZ et dépense santé sont en revanche négativement corrélées à la première composante principale. Les figures 1 et 2 mettent en évidence le fait qu'en Afrique, une minorité de pays ont un taux de mortalité, une croissance démographie, un niveau de la sous alimentation faibles. Pour le taux de mortalité, on note par exemple le fait que 38.8% de pays de la zone Afrique ont un taux de mortalité $>= 1.5\%$. Dans la zone Alena, 0% de pays ont un taux de mortalité supérieur à 1.5%.
De même, on est aussi forcé de constater qu'en Europe et dans la zone Alena, la majorité des pays ont un niveau élevé des dépenses en matière de santé, PIB, IDH, émission de GAZ, consommation d'électricité. L'Europe et l'ALENA sont par exemple les deux zones géographiques dans lesquelles plus de 66% de pays ont un PIB par habitant supérieur à 20000k$. On constate aussi que dans les états de l'ex-URSS, la plupart des états ont une croissance démographique faible. En effet, 60% des pays de l'ex-URSS ont une croissance de la population négative. Enfin, la visualisation des régions sous la forme de rectangles permet de constater par exemple qu'en Afrique, les différences entre pays en termes de variables macroéconomique sont très importantes. Si on considère le taux de mortalité en zone Afrique par exemple,

Analyse en composante principales d'un tableau de distributions macroéconomiques.

6.1% de pays ont un taux de mortalité inférieur à 0.5%, 18.4% de pays ont un taux de mortalité compris entre 0.5% et 0.6%, 36.7% de pays ont un taux de mortalité compris entre 0.6% et 0.7%, 30.6% de pays ont un taux de mortalité compris entre 0.7% et 0.8% et 8.2% de pays ont un taux de mortalité supérieur à 0.8%. Enfin, le codage non paramétrique utilisé pour obtenir ces résultats suppose que les écarts entre deux modalités consécutives sont constants. Compte tenu de cette limite, nous proposons dans la section suivante l'utilisation des scores de Ridit brutes et standardisés.

4.1.2 Utilisation des scores de Ridit.

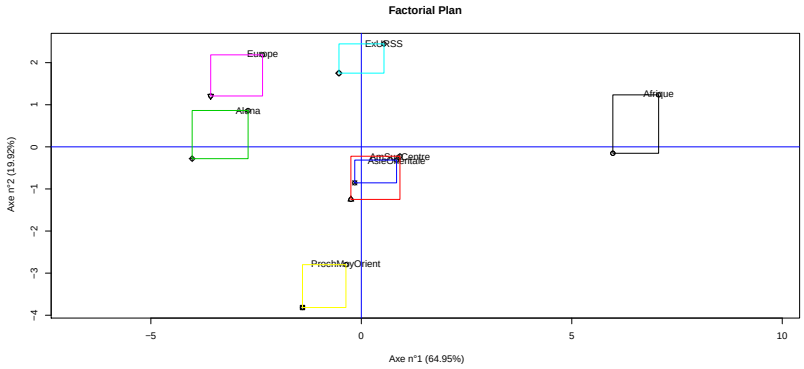

FIGURE 3 – *Plan des projections obtenu par les scores de Ridit bruts.*

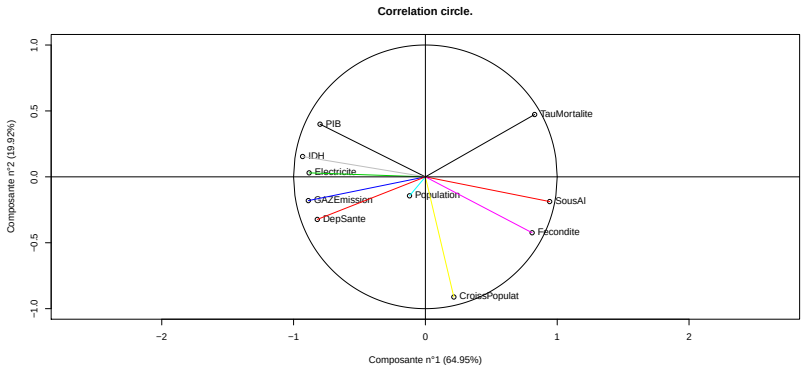

FIG. 4 – *Carte des Corrélations obtenu par les scores de Ridit bruts.*

Nous utilisons à présent les scores de Ridit bruts dans l'ACP basée sur les barycentres. Le premier et deuxième axe factoriel ont respectivement un pourcentage de variabilités de 64.95% et 19.92%. (cf. figures 3 et 4). Le positionnement des zones géographiques sur les axes factoriels (cf. figure 3) est similaire de celui obtenu par le codage non paramétrique. Toutefois, dans le premier axe de la figure 3, l'opposition entre Europe et l'Alena d'une part et l'Afrique est beaucoup plus manifeste. Au niveau des variables, la variable IDH est mieux représentée dans la figure 4. En effet, lorsqu'on utilise les scores de Ridit bruts, la corrélation de Pearson entre la variable IDH et la première composante principale est -0.93 (voir tableau 3). Cependant, cette corrélation est de l'ordre de 0.19 lorsqu'on utilise le codage non paramétrique.

Analyse en composante principales d'un tableau de distributions macroéconomiques.

TABLE 4 – *Corrélation entre les moyennes (par les scores de ridit brutes) et leur composantes princiaples.*

Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
PIB	-0.80	0.40	-0.22	0.18	0.35
SousAl	0.94	-0.19	0.27	0.02	-0.00
Electricite	-0.88	0.03	-0.45	0.04	-0.13
GAZEmission	-0.89	-0.18	-0.26	0.05	0.11
Population	-0.12	-0.14	-0.05	0.87	-0.43
Fecondite	0.81	-0.42	-0.32	0.12	0.22
CroissPopulat	0.22	-0.91	-0.30	-0.11	-0.15
IDH	-0.93	0.15	0.20	-0.10	0.23
TauMortalite	0.83	0.47	-0.28	-0.07	-0.05
DepSante	-0.82	-0.32	-0.09	-0.43	0.09

TABLE 5 – *Corrélation entre les moyennes (par les scores de ridit standardisés) et leur compo-santes princiaples.*

Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
PIB	-0.96	0.17	-0.07	-0.03	0.20
SousAl	0.99	-0.11	-0.08	-0.03	-0.02
Electricite	-0.97	-0.04	0.11	0.20	-0.08
GAZEmission	-0.93	-0.15	0.14	0.23	-0.19
Population	-0.34	-0.10	-0.87	0.33	0.02
Fecondite	0.72	-0.49	0.14	0.29	0.36
CroissPopulat	0.28	-0.93	0.18	0.14	-0.05
IDH	-0.97	0.04	0.02	-0.19	0.11
TauMortalite	0.62	0.61	0.24	0.44	0.03
DepSante	-0.93	-0.14	0.25	0.11	0.13

4.1.3 Utilisation des scores de Ridit standardisés.

Dans le but d'uniformiser les valeurs des scores de Ridit des variables à analyser, il est également possible de normaliser ou standardiser ces scores. Nous appliquons donc l'approche de Makosso-Kallyth et Diday (2012) en utilisant des scores standardisés. Les résultats obtenus dans les figures 5 et 6 sont semblables à ceux obtenus dans les figures 3 et 4. Les tables 3 et 4 contenant les corrélations entre les moyennes des variables et leurs composantes principales sont quasi similaires. Toutefois, la figure 5 met plus en évidence l'hétérogénéité des zones géographiques notamment l'Afrique. En outre, la proximité en termes de caractéristiques macroéconomique de l'Afrique, l'Amérique du Sud et du centre ainsi que l'Asie orientale est plus perceptible.

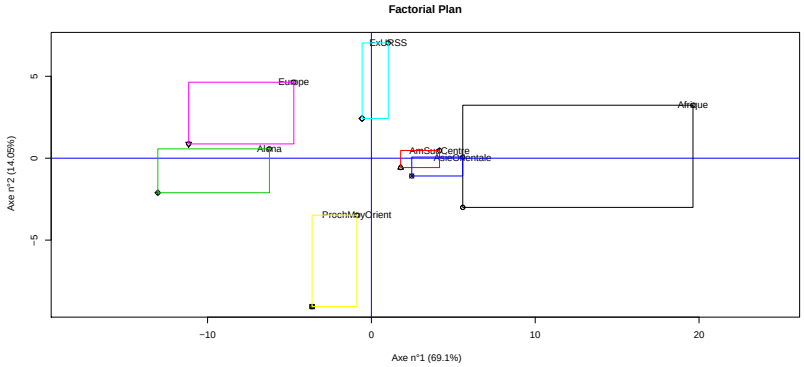

FIGURE 5 – *plan de projections obtenu par les scores de Ridit standardisés.*

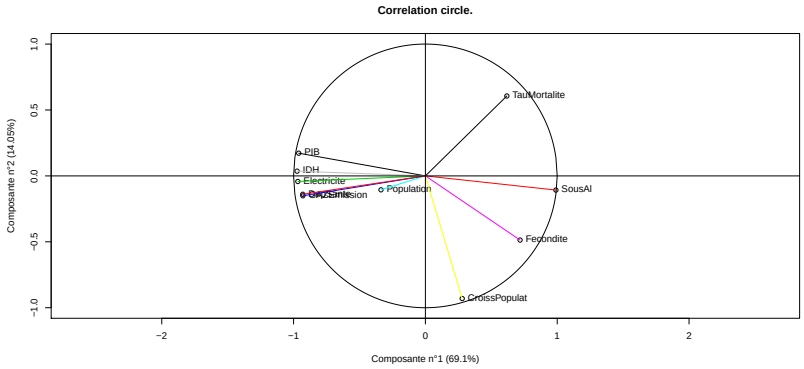

FIG. 6 – *Carte des Corrélations obtenu par les scores de Ridit standardisés.*

L'ACP Makosso-Kallyth et Diday (2012) se focalise essentiellement sur la moyenne des distributions dans la détermination des axes principaux. Elle nécessite également un codage des modalités des variables ainsi que le choix de la valeur de t dans la règle de Tchebytchev. Dans la section suivante nous appliquons la seconde approche décrite dans le cadre de cet article.

Les figures 7 et 8 représentent respectivement les plans de projections des individus. Dans la figure 7, les individus sont représentés sous la forme de rectangles. Plus la surface des rectangles est grande, plus la dispersion des concepts est importante. La dispersion des concepts ici se rapporte à la dispersion des quantiles et non à la dispersion des moyennes. Au regard de la figure 7, on est forcé de noter que L'Alena (Usa, Canada, Mexique) semble être la zone géographique la plus homogène.

Dans la figure 8 nous représentons en 2D les enveloppes convexes de la projection des quantiles (voir aussi Irpino et al. (2003)). Par rapport aux rectangles, les enveloppes convexes en deux dimensions permettent de mieux visualiser la dispersion des concepts. La figure 8 conforte la constatation précédente selon laquelle l'Alena serait la zone géographique la plus homogène.

Analyse en composante principales d'un tableau de distributions macroéconomiques.

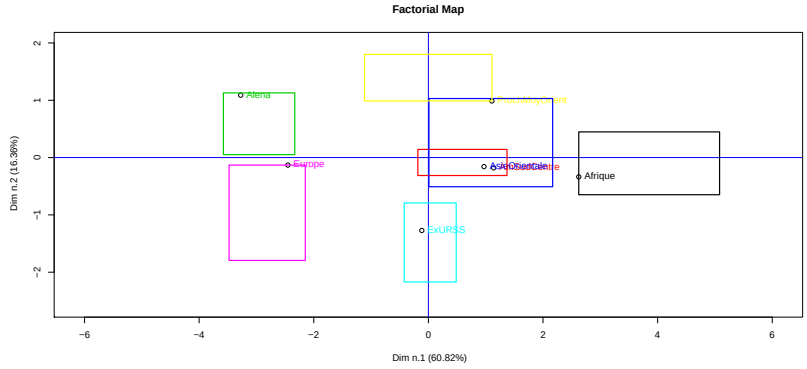

FIGURE 7 – *Plan de projections avec représentation des rectangles.*

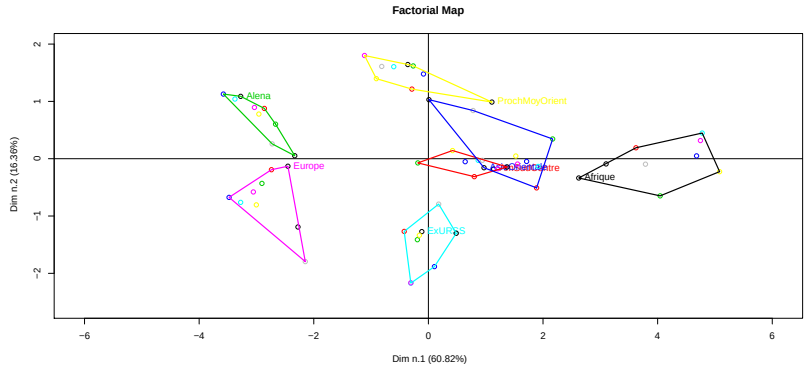

FIGURE 8 – *Plan de projections avec représentation des enveloppes convexes.*

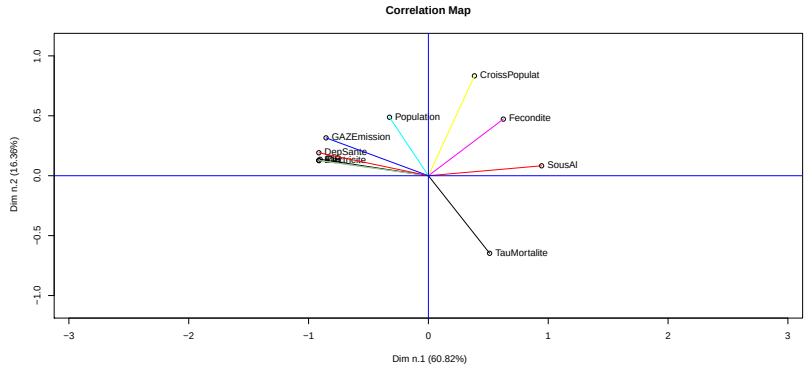

FIGURE 9 – *Carte des corrélations.*

TABLE 6 – *Corrélation entre les quantiles des variables et les 5 premières composantes prin-cipales.*

Variable	Comp.1	Compo.2	Comp.3	Comp.4	Comp.5
PIB	-0.90	0.14	-0.03	0.62	-0.28
SousAl	0.94	0.08	0.22	-0.29	0.13
Electricite	-0.92	0.13	-0.41	0.43	-0.02
GAZEmission	-0.85	0.32	-0.48	0.50	0.38
Population	-0.32	0.49	0.65	0.32	0.02
Fecondite	0.62	0.47	-0.19	0.21	0.08
CroissPopulat	0.38	0.83	-0.16	0.19	0.42
IDH	-0.91	0.13	-0.27	0.06	-0.17
TauMortalite	0.51	-0.65	0.19	0.14	0.02
DepSante	-0.91	0.19	-0.32	0.66	-0.20

La figure 9 ci-dessus permet d'interpréter les figure 7 et 8. Les résultats obtenus dans ces graphiques sont assez similaires de ceux obtenus avec l'approche proposée par Makosso-Kallyth et Diday (2012). On note par exemple que sur l'axe 1, l'Afrique est opposé à l'Europe et l'Alena. Les raisons de cette constatations sont essentiellement dues au fait que des variables telles que la croissance de la population et le taux de mortalité sont élevés en Afrique mais faible en Europe et en Amérique du nord (Alena). Enfin, contrairement à la méthode des barycentres avec codage non paramétrique, la variable IDH est bien représentée sur l'axe 1 (voir la table 6).

Analyse en composante principales d'un tableau de distributions macroéconomiques.

5 Discussion.

Cet article présente l'application de deux extensions de l'ACP à un tableau de distributions macroéconomiques. Les méthodes décrites déterminent les axes principaux des moyennes ou des quantiles. La méthode basée sur la détermination des barycentres requiert au préalable un codage des modalités. Plusieurs approches sont envisageable à cet effet. Le codage non paramétrique par exemple est d'un usage simple et est moins contraignant que le codage paramétrique. Cependant, une des limites du codage non paramétrique est le fait de considérer de manière implicite que les écarts entre centre deux classes consécutives sont constants. Pour venir à bout de cette limite, on propose également un codage par les scores de Ridit. Dans les analyses effectuées, ces scores (brutes et standardisés) ont permis de mieux mettre en évidence du point de vue graphique (voir les figures 5 et 3) l'opposition entre pays riche (Alena, Europe) et pays pauvre ou en voie de développement (Afrique). Dans la détermination des axes principaux, l'ACP proposée par Makosso-Kallyth et Diday (2012) est essentiellement classique. Cependant, son aspect symbolique découle du fait qu'elle détermine des hypercubes d'intervalles et qu'elle projette ces hypercubes en éléments supplémentaires. Bien qu'il soit possible d'utiliser en éléments actifs les hypercubes ou les intervalles issus de la transformation des histogrammes dans le cadre d'une ACP symbolique, une telle démarche équivaudrait à l'analyse d'un tableau d'intervalles et non à l'analyse directe d'un tableau de distribution. De plus, la transformation du tableau d'histogrammes ou de distribution en tableau d'intervalle n'est pas bijective. Pour un tableau donné de distributions on peut faire correspondre plusieurs tableau d'intervalles. Ainsi, assimiler le tableaux d'intervalles transformés à un tableau actifs conduirait de facto à de multiples système d'axes principaux. Dans l'ACP de Makosso-Kallyth et Diday (2012) par contre, les axes principaux sont uniques. Toutefois, la visualisation des individus varie en fonction des intervalles construit par l'inégalité de Tchebychev.

L'approche proposé par Makosso-Kallyth (2015) utilise plusieurs point caractéristiques des distributions (les quantiles). Elle ne nécessite aucun codage de modalités. Elle exige toutefois la spécification du nombre de quantiles ainsi que leur localisation. Contrairement aux moyennes, les quantiles se prêtent bien à toutes les formes de distributions (symétriques, non symétriques, etc.).

6 Conclusion

Cet article compare deux adaptations ou extensions de l'ACP à un tableau de distributions. Ces approches se prêtent bien à l'analyse exploratoire de données à deux niveaux de généralité (pays et zones géogra-phiques, par exemple). Elles sont des compléments de l'ACP classique et de peuvent renforcer l'analyse des données exploratoires de données complexes. Elles permettent par exemple à partir d'un simple coup d'oeil l'identification d'agrégats économiques similaires. Elles permettent aussi de s'imprégner de la dis-persion de ces agrégats économiques. L'approche proposée par Makosso-Kallyth et Diday (2012), est plus ou moins tributaire du choix du codage des modalités des variables. Du fait de l'impact du codage, on peut par exemple utiliser dans une première mesure le codage par les scores de Ridit standardisés et recourir à des codages alternatifs dans le cadre d'une analyse de sensibilité. L'approche basée sur les quantiles par contre, nécessite un choix a priori du nombre de quantiles ainsi que leur localisation. Elle ne nécessite cependant aucun codage. Toutefois, lorsque le nombre de variables p devient très grand, les méthodes présentées, notamment l'ACP de Makosso-Kallyth et Diday (2012), deviennent fastidieuses et les problèmes liés à la malédiction de la dimension (voir Bellman (1961)) peuvent resurgir. Dans de telles circonstances le recours à des versions régularisées de l'ACP pourrait améliorer les approches présentées.

Références

Agresti, A. (2002). *Categorical data analysis*. Wiley series in probability and statistics. Hoboken (N.J.) : J. Wiley.

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

Billard, L. et E. Diday (2006). *Symbolic Data Analysis : conceptual statistics and data Mining*. Berlin : Wiley series in computational statistics.

Bock, H.-H. et E. Diday (2000). *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Informations from Complex Data*. Berlin : Springer.

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics 14*(1), pp. 18–38.

Diday, E. (1988). Introduction à lanalyse des données symboliques : objets symboliques modaux et implicites. In *Deuxis Journées Symbolique-Numérique, Université dOrsay.*, Orsay, France, pp. 127–139.

Diday, E. (1989). Introduction approche symbolique en analyse des donn. *RAIRO - Operations Research - Recherche Opérationnelle 23*(2), 193–236.

Diday, E. (2013). Principal component analysis for bar charts and metabins tables. *Statistical analysis and data mining 6*(5), 403–430.

Diday, E. et M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Chichester : Wiley Interscience.

Donaldson, G. W. (1998). Ridit scores for analysis and interpretation of ordinal pain data. *European Journal of Pain 2*(3), 221–227.

Ichino, M. (2011). The quantile method for symbolic principal component analysis. *Stat. Anal. Data Min. 4*(2), 225–233.

Irpino, A., C. Lauro, et R. Verde (2003). Visualizing symbolic data by closed shapes. In *Between Data Science and Applied Data Analysis*, pp. 244–251. Springer Berlin Heidelberg.

Makosso-Kallyth, S. (2015). Principal axes analysis of symbolic histogram variables. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, n/a–n/a.

Makosso-Kallyth, S. et E. Diday (2010). Analyse en axes principaux de variables symboliques de type histogramme. In *42s Journées de Statistique*, Marseille, France.

Makosso-Kallyth, S. et E. Diday (2012). Adaptation of interval pca to symbolic histogram variables. *Advances in Data Analysis and Classification 6*, 147–159.

Mantel, N. (1979). Ridit analysis and related ranking procedures–use at your own risk. *American Journal Epidemiol.* (109), 25–29.

Nagabhushan, P. et R. P. Kumar (2007). Histogram pca. In *ISNN (2)*, pp. 1012–1021.

Rodriguez, O., E. Diday, et S. Winberg (2000). Generalization of the principal components analysis to histogram data. In *4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases*, Lyon, France.

Verde, R., A. Irpino, et A. Balzanella (2015). Dimension reduction techniques for distributional symbolic data. *Cybernetics, IEEE Transactions on PP*(99), 1–1.

Analyse en composante principales d'un tableau de distributions macroéconomiques.

Annexes

Les tableaux de distributions utilisées sont données ci-dessous :

TABLE 7 – *Variable PIB*

	Region	PIB1	PIB2	PIB3	PIB4	PIB5
1	Afrique	0.34	0.51	0.06	0.08	0.00
2	AmSudCentre	0.00	0.27	0.54	0.18	0.00
3	Alena	0.00	0.00	0.00	0.33	0.67
4	AsieOrientale	0.07	0.67	0.07	0.07	0.13
5	ExURSS	0.00	0.43	0.28	0.28	0.00
6	Europe	0.00	0.00	0.16	0.16	0.68
7	ProchMoyOrient	0.00	0.50	0.25	0.00	0.25

TABLE 8 – *Variable Sous alimentation*

	Region	SousAlim1	SousAlim2	SousAlim3	SousAlim4	SousAlim5
1	Afrique	0.06	0.13	0.28	0.19	0.34
2	AmSudCentre	0.04	0.36	0.46	0.09	0.04
3	Alena	0.67	0.33	0.00	0.00	0.00
4	AsieOrientale	0.13	0.13	0.53	0.20	0.00
5	ExURSS	0.21	0.50	0.14	0.07	0.07
6	Europe	0.74	0.26	0.00	0.00	0.00
7	ProchMoyOrient	0.12	0.62	0.12	0.00	0.12

TABLE 9 – *Variable Electricité*

	Region	Electricte1	Electricte2	Electricte3	Electricte4	Electricte5
1	Afrique	0.58	0.19	0.23	0.00	0.00
2	AmSudCentre	0.14	0.33	0.48	0.05	0.00
3	Alena	0.00	0.00	0.33	0.33	0.33
4	AsieOrientale	0.33	0.27	0.20	0.20	0.00
5	ExURSS	0.00	0.00	0.87	0.13	0.00
6	Europe	0.00	0.00	0.35	0.48	0.16
7	ProchMoyOrient	0.15	0.00	0.38	0.31	0.15

TABLE 10 – *Variable Emission de Gaz*

	Region	EmisGaz1	EmisGaz2	EmisGaz3	EmisGaz4	EmisGaz5
1	Afrique	0.61	0.15	0.12	0.08	0.04
2	AmSudCentre	0.29	0.33	0.29	0.05	0.05
3	Alena	0.00	0.00	0.33	0.00	0.67
4	AsieOrientale	0.33	0.20	0.20	0.20	0.07
5	ExURSS	0.13	0.20	0.20	0.27	0.20
6	Europe	0.00	0.03	0.10	0.71	0.16
7	ProchMoyOrient	0.08	0.08	0.23	0.08	0.54

TABLE 11 – *Variable Population*

	Region	Population1	Population2	Population3	Population4	Population5
1	Afrique	0.08	0.24	0.14	0.51	0.02
2	AmSudCentre	0.05	0.24	0.33	0.33	0.05
3	Alena	0.00	0.00	0.00	0.33	0.67
4	AsieOrientale	0.06	0.11	0.06	0.50	0.28
5	ExURSS	0.00	0.47	0.27	0.20	0.07
6	Europe	0.10	0.23	0.23	0.45	0.00
7	ProchMoyOrient	0.09	0.27	0.18	0.36	0.09

TABLE 12 – *Variable Fecondite*

	Region	Fecondite1	Fecondite2	Fecondite3	Fecondite4	Fecondite5
1	Afrique	0.09	0.27	0.36	0.27	0.00
2	AmSudCentre	0.00	0.71	0.29	0.00	0.00
3	Alena	0.33	0.67	0.00	0.00	0.00
4	AsieOrientale	0.62	0.25	0.12	0.00	0.00
5	ExURSS	0.60	0.40	0.00	0.00	0.00
6	Europe	0.93	0.07	0.00	0.00	0.00
7	ProchMoyOrient	0.00	0.67	0.33	0.00	0.00

TABLE 13 – *Variable Croissance de la population.*

	Region	CroisPopul1	CroisPopul2	CroisPopul3	CroisPopul4	CroisPopul5
1	Afrique	0.00	0.10	0.26	0.63	0.00
2	AmSudCentre	0.00	0.30	0.65	0.04	0.00
3	Alena	0.00	0.33	0.67	0.00	0.00
4	AsieOrientale	0.00	0.28	0.50	0.22	0.00
5	ExURSS	0.60	0.13	0.27	0.00	0.00
6	Europe	0.00	0.81	0.12	0.08	0.00
7	ProchMoyOrient	0.00	0.00	0.25	0.50	0.25

Analyse en composante principales d'un tableau de distributions macroéconomiques.

TABLE 14 – *Variable IDH*

	Region	IDH1	IDH2	IDH3	IDH4	IDH5
1	Afrique	0.45	0.31	0.12	0.08	0.04
2	AmSudCentre	0.00	0.04	0.09	0.52	0.35
3	Alena	0.00	0.00	0.00	0.00	1.00
4	AsieOrientale	0.00	0.28	0.11	0.39	0.22
5	ExURSS	0.00	0.00	0.13	0.53	0.33
6	Europe	0.00	0.00	0.00	0.03	0.97
7	ProchMoyOrient	0.00	0.15	0.00	0.31	0.54

TABLE 15 – *Variable Taux de mortalité*

	Region	TauxMort1	TauxMort2	TauxMort3	TauxMort4	TauxMort5
1	Afrique	0.06	0.18	0.37	0.31	0.08
2	AmSudCentre	0.09	0.91	0.00	0.00	0.00
3	Alena	0.33	0.67	0.00	0.00	0.00
4	AsieOrientale	0.17	0.83	0.00	0.00	0.00
5	ExURSS	0.00	0.40	0.47	0.13	0.00
6	Europe	0.00	0.74	0.26	0.00	0.00
7	ProchMoyOrient	0.61	0.38	0.00	0.00	0.00

TABLE 16 – *Variable dépense santé*

	Regions	DEP1	DEP2	DEP3	DEP4	DEP5
1	Afrique	0.70	0.08	0.22	0.00	0.00
2	AmSudCentre	0.04	0.22	0.70	0.04	0.00
3	AmNord	0.00	0.00	0.33	0.00	0.67
4	AsOrient	0.56	0.22	0.06	0.11	0.06
5	ExURSS	0.20	0.20	0.53	0.07	0.00
6	Europe	0.00	0.00	0.23	0.11	0.66
7	PrMoyOr	0.15	0.08	0.38	0.23	0.15