International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 11 No. 2 Nov. 2014, pp. 371-378

http://www.ijisr.issr-journals.org/

Analyse des composantes principales : cas d'un échantillon des prestataires logistiques de la région du grand Casablanca

[Principal component analysis : case of a sample of logistics service providers of the great Casablanca region]

Moulay El Mehdi Falloul

Doctorant en économie et finance appliquée,
Université Hassan II Mohammedia,
Mohammedia, Maroc

ABSTRACT: The principal component analysis, introduced by Hotelling in 1933, is a descriptive method which is aimed at the analysis of the data tables which does not have a particular structure, in other words, comments at first glance with no distinction between variables, or between individuals. The PCA aims to summarize information contained in an array consisting of large number of rows and columns, a few graphs in two dimensions, and more a number of digital features. We will use this method to analyze a sample of logistics service providers in the region of the great Casablanca in Morocco.

KEYWORDS: Principal component analysis, random sample, logistics service providers, explained variance, component diagram.

RESUME: L'analyse en composante principale, introduite par Hotelling en 1933, est une méthode descriptive qui a pour but l'analyse des tableaux des données qui ne présente pas une structure particulière, autrement dit, des observations ne comportant à priori aucune distinction, ni entre variables, ni entre individus. L'objectif de l'ACP est de résumer l'information contenue dans un tableau constitué de nombre élevé de lignes et de colonnes, en quelques représentations graphiques à deux dimensions, plus un certain nombre de caractéristiques numériques. Nous allons utiliser cette méthode pour analyser un échantillon des prestataires logistiques de la région du grand Casablanca au Maroc.

MOTS-CLEFS: Analyse en composante principale, échantillon aléatoire, prestataires logistiques, variance expliquée, diagrammes des composantes.

1 INTRODUCTION

Le contexte dans lequel on applique cette technique est le suivant: on observe sur N individus P caractères ou variables quantitatives présentant des relations multiples qu'on veut analyser.

Décrire à l'aide d'un graphique le lien entre deux variables quantitatives est simple: il suffit de porter sur deux axes orthogonaux les valeurs des variables en question pour observer à l'œil le lien entre les deux variables, quitte à effectuer ensuite une analyse ou des tests statistiques plus précis. Pour 3 variables, une démarche analogue mène à un graphique à 3 dimensions ou en perspective. Pour 4 variables et plus, il n'est plus possible de procéder de la même manière, et travailler par couple de variables ou par triple de variables risque de masquer des interactions complexes [1]. D'où l'idée de mettre au

point une technique permettant de résumer l'information apportée par ces P caractéristiques en la détruisant le moins possible. Cette technique utilise des combinaisons linéaires des variables, elle est donc mieux adaptée aux relations linéaires.

2 LES BASES MATHÉMATIQUES DE L'ACP

2.1 LA PROCÉDURE POUR OBTENIR LES COMPOSANTS

Première composante principale ≡*combinaison linéaire* $a_1^t Z$ *qui maximise*

$Var\ (a^t Z)$ *Sous réserve de* $a^t a = 1$ *et*

$$\Leftrightarrow \quad Var\ (a_1^t Z) \geq Var\ (b^t Z) \quad pour\ tout\ b^t b = 1$$

Deuxième composante principale ≡combinaison linéaire $a_2^t Z$ qui maximise $Var\ (a^t Z)$ sous réserve de $a^t a = 1$, $a_2^t a_2 = 1$.

Et $Cov\ (a_1^t Z, a_2^t Z) = 0$. $\Leftrightarrow$ $a_2^t Z$ Maximise $Var(a^t Z)$ et n'est pas également corrélées avec la première composante principale.

À l'ième étape, **la *i*^{ème} *composante principale*** ≡combinaison linéaire $a_i^t Z$ qui maximise $Var\ (a^t Z)$ sous réserve de $a^t a = 1$, $a_i^t a_i = 1$.

et $Cov\ (a_i^t Z, a_k^t Z) = 0, k < i$. $\Leftrightarrow$ $a_i^t Z$ maximise $Var(a^t Z)$ et n'est pas également corrélées avec la première composante principal (i-1)t.

Intuitivement, ces composantes principales avec grande variance contiennent des informations « importantes ». En revanche, ces composantes principales avec petite variance peuvent être « redondantes » [2]. Par exemple, supposons que nous ayons 4 variables Z_1, Z_2, Z_3, et Z_4 .

Soit $Var\ (Z_1) = 4, Var\ (Z_2) = 3, Var\ (Z_3) = 2$

Et $Z_3 = Z_4$.

En outre, supposons que Z_1, Z_2, Z_3 sont mutuellement non corrélées. Ainsi, parmi ces quatre variables, seulement 3 d'entre eux sont nécessaires puisque deux d'entre eux sont les mêmes. En utilisant la procédure pour obtenir les principales composantes ci-dessus, alors la première composante principale est

$$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} = Z_1,$$

La deuxième composante principale est

$$\begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} = Z_2,$$

La troisième composante principale est $\begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}Z_1\\Z_2\\Z_3\\Z_4\end{bmatrix} = Z_3$

Et la quatrième composante principale est

$$\begin{bmatrix}0 & 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}Z_1\\Z_2\\Z_3\\Z_4\end{bmatrix}=\frac{1}{\sqrt{2}}(Z_3-Z_4)=0$$

La quatrième composante principale est donc redondante. Autrement dit, seulement 3 morceaux « important » des informations cachées dans Z_1, Z_2, Z_3 et Z_4 .

Théorème :

$a_1, a_2, \ldots, a_p$ sont les vecteurs propres de Σ valeurs propres correspondants

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ [3]. En outre, l'écart des principales composantes sont les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$. C'est. $Var\ (Y_i) = Var\ (a_i^t Z) = \lambda_i$

Justification

Puisque Σ est symétrique et inversible, $\Sigma = P \Lambda P^t$, où *P* est une matrice orthonormée, Λ est une matrice diagonale avec des éléments diagonaux $\lambda_1, \lambda_2, \ldots, \lambda_p$, la i[ème] colonne de *P* est le vecteur orthonormé a_i $(a_i^t a_j = a_j^t a_i = 0, i \neq j, a_i^t a_i = 1)$ et λ_i est la valeur propre de Σ correspondant à a_i [4].

Ainsi,

$\Sigma = \lambda_1 a_1 a_1^t + \lambda_2 a_2 a_2^t + \ldots + \lambda_p a_p a_p^t$.

Pour tout vecteur unitaire $b = c_1 a_1 + c_2 a_2 + \ldots + c_p a_p$ ($a_1, a_2, \ldots, a_p$ est une base de R^P),

$c_1, c_2, \ldots, c_p \in R$, $\sum_{i=1}^{p} c_i^2 = 1$,

$$Var\ (b^t Z) = b^t \Sigma b = b^t (\lambda_1 a_1 a_1^t + \lambda_2 a_2 a_2^t + \cdots + \lambda_p a_p a_p^t) b$$
$$= c_1^2 \lambda_1 + c_2^2 \lambda_2 + \cdots + c_p^2 \lambda_p \leq \lambda_1 \quad ,$$

Et

$Var\,(a_1^t Z) = a_1^t \Sigma a_1 = a_1^t (\lambda_1 a_1 a_1^t + \lambda_2 a_2 a_2^t + \cdots + \lambda_p a_p a_p^t) a_1 = \lambda_1$.

Ainsi, $a_1^t Z$ ***est la première composante principale et*** $Var(a_1^t Z) = \lambda_1$.

De même, pour tout vecteur *c* satisfaisant $Cov\,(c^t Z, a_1^t Z) = 0$,

Puis $$c = d_2 a_2 + \ldots + d_p a_p,$$

Où $$d_2, d_3, \ldots, d_p \in R$$

Et $$\sum_{i=2}^{p} d_i^2 = 1 .$$

Puis,

$$Var(c^t Z) = c^t \Sigma c = c^t (\lambda_1 a_1 a_1^t + \lambda_2 a_2 a_2^t + \cdots + \lambda_p a_p a_p^t) c$$
$$= d_2^2 \lambda_2 + \cdots + d_p^2 \lambda_p \le \lambda_2$$

Et

$$Var\,(a_2^t Z) = a_2^t \Sigma a_2 = a_2^t (\lambda_1 a_1 a_1^t + \lambda_2 a_2 a_2^t + \cdots + \lambda_p a_p a_p^t) a_2 = \lambda_2$$

Ainsi, $a_2^t Z$ ***est la deuxième composante principale et*** $Var(a_2^t Z) = \lambda_2$.

Les autres composantes principales peuvent être justifiées de la même façon.

2.2 ESTIMATION

Les principales composantes ci-dessus sont les principales composantes théoriques. Pour trouver les composantes principales « estimées », nous estimons que la matrice de variance-covariance théorique Σ de l'échantillon variance-covariance $\hat{\Sigma}$ [5],

$$\hat{\Sigma} = \begin{bmatrix} \hat{V}(Z_1) & \hat{C}(Z_1, Z_2) & \cdots & \hat{C}(Z_1, Z_p) \\ \hat{C}(Z_2, Z_1) & \hat{V}(Z_2) & \cdots & \hat{C}(Z_2, Z_p) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{C}(Z_p, Z_1) & \hat{C}(Z_p, Z_2) & \cdots & \hat{V}(Z_p) \end{bmatrix},$$

Où

$$\hat{V}(Z_j) = \frac{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2}{n-1},\ \hat{C}(Z_j, Z_k) = \frac{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{n-1},\ j, k = 1, \ldots, p.$$

Où les $\overline{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}$. *Ensuite, supposons que* $e_1, e_2, \ldots, e_p$ *sont des vecteurs propres orthonormés de* $\hat{\Sigma}$ *correspondant aux valeurs propres* $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_p$ *. Ainsi, l'*$i^{ème}$ *composantes principales estimée est* $\hat{Y}_i = e_i^t Z, i = 1, \ldots, p.$ *et la variance estimée de la composante principale estimée d'IE est* $\hat{V}(\hat{Y}_i) = \hat{\lambda}_i$.

3 CONSTRUCTION DE LA BASE DE DONNEES ET RESULTATS

3.1 CONSTITUTION DE L'ÉCHANTILLON

Nous avons constitué notre base des données sur la base d'un échantillon de 17 sociétés de prestations du service logistique choisies selon la technique du tirage aléatoire, ces entreprises opèrent essentiellement dans la région du grand Casablanca. Notre échantillon se compose sept principales caractéristiques métriques ; le montant du capital social en million de dirhams, la surface des entrepôts en mettre carrés, le nombre de véhicules et équipements logistiques, le nombre de palettes de capacité de stockage par entreprise et le nombre de collaborateurs de chaque entreprise.

3.2 RÉSULTATS DE L'ÉTUDE

3.2.1 STATISTIQUES DESCRIPTIVES DES VARIABLES

Tableau 1. Statistiques descriptives

	Moyenne	Ecart-type	n analyse	etype/moyenne
capital	1397647,06	2130344,84	17	1,524236623
surface	13447,65	31129,584	17	2,314871669
nmbrvehicule	38,94	46,623	17	1,197303544
nmbrepalettes	1128,82	2922,321	17	2,588828157
nmbrecollaborateur	64,29	78,801	17	1,225711619

D'après Le coefficient de variation, on peut conclure que toutes les variables sont très dispersées, ce qui indique que les entreprises ont des profils très hétérogènes.

3.2.2 MATRICE DE CORRÉLATION DES VARIABLES INITIALES

Tableau 2. Statistiques descriptives

		capital	surface	Nombre de véhicules	Nombre de palettes	Nombre de collaborateur
Corrélation	capital	1,000	,940	,292	,311	,627
	surface	,940	1,000	,328	,224	,659
	nmbrvehicule	,292	,328	1,000	,573	,680
	nmbrepalettes	,311	,224	,573	1,000	,256
	nmbrecollaborateur	,627	,659	,680	,256	1,000

Dans l'ensemble les variables sont moyennement et faiblement corrélées, néanmoins on note une assez forte corrélation entre le montant du capital de l'entreprise et la surface de de stockage (0.940).

3.2.3 CHOIX DES COMPOSANTES PRINCIPALES

Tableau 3. Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Som mes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,009	60,174	60,174	3,009	60,174	60,174
2	1,151	23,028	83,201	1,151	23,028	83,201
3	,641	12,824	96,025			
4	,150	3,008	99,033			
5	,048	,967	100,000			

La première a une valeur propre de 3.009 qui représente 60.174% de la variance totale des variables initiales. Les 2 premières composantes contribuent, tous les deux, à 83.201% de la variance initiale.

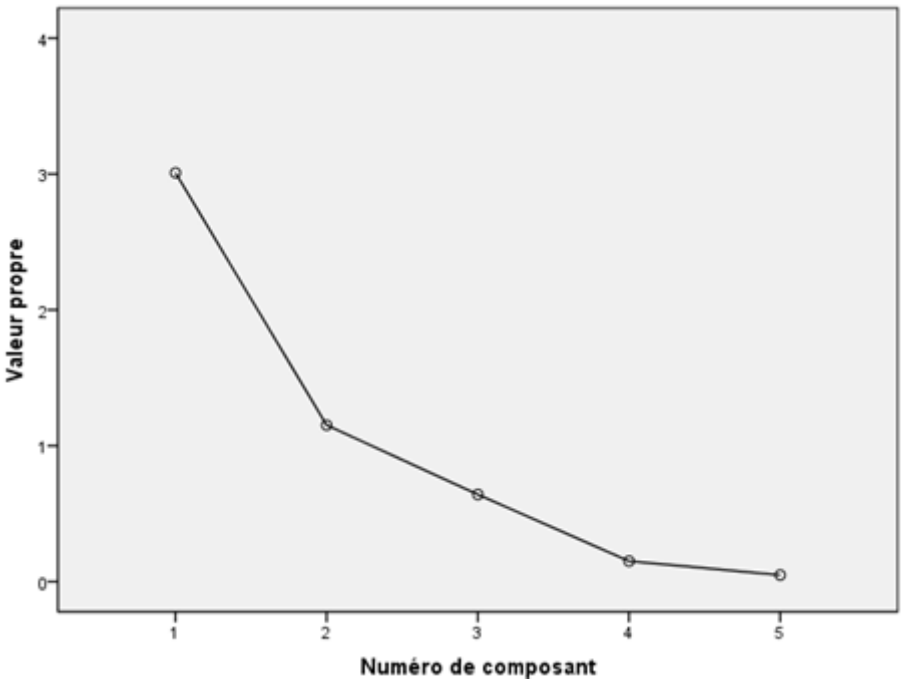

Fig. 1. Graphique de valeurs propres

D'après le graphe des valeurs propres, on peut retenir deux composantes principales, la différence entre la deuxième et le troisième est très important.

Tableau 4. *Qualité de représentation*

	Initial	Extraction
capital	1,000	,915
surface	1,000	,947
nmbrvehicule	1,000	,848
nmbrepalettes	1,000	,712
nmbrecollaborateur	1,000	,738

Les deux composantes contribuent à 91.5% de la variance du montant du capital.

D'après le tableau de la qualité de représentation, les deux composantes sont suffisantes pour synthétiser les variances de la majorité des variables.

Tableau 5. *Graphique de valeurs propres*[a]

	Composante	
	1	2
surface	,860	-,456
nmbrecollaborateur	,859	-,013
capital	,857	-,424
nmbrvehicule	,710	,586
nmbrepalettes	,541	,648

a. 2 composantes extraites.

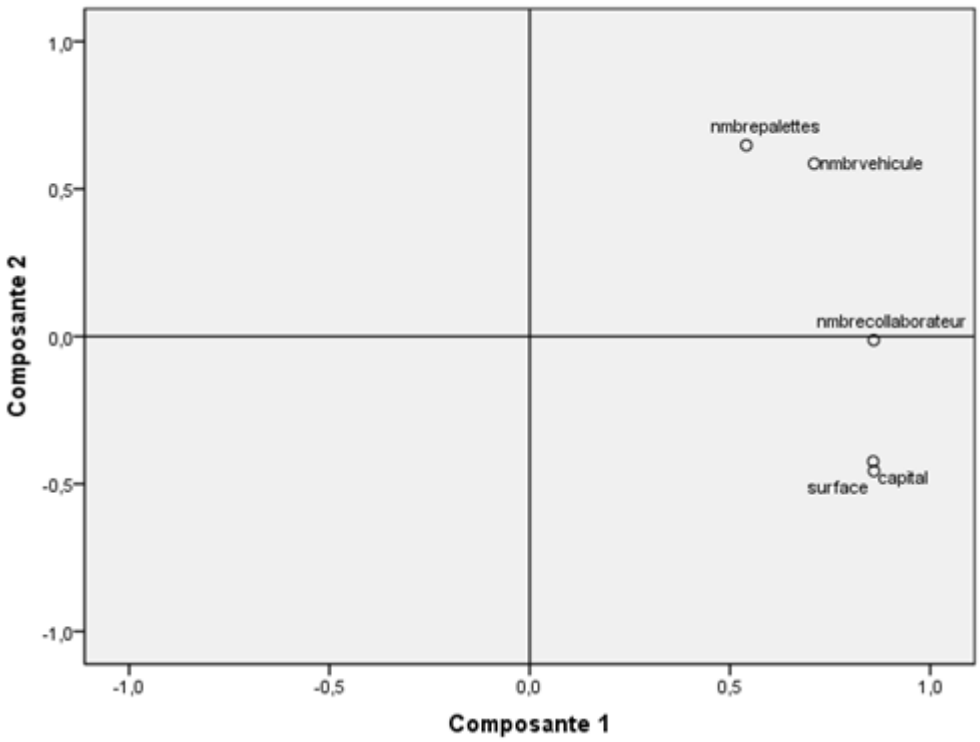

Fig. 2. *Diagramme des composantes*

La première composante est fortement corrélée avec la surface de stockage, le capital investi et le nombre de collaborateurs

La deuxième composante est assez corrélée avec le nombre de véhicules et le nombre de palettes. Elle est négativement corrélée avec la surface de stockage, le capital et le nombre de collaborateur. On peut conclure que la deuxième composante met en opposition deux catégories d'entreprises, une catégorie qui détient un très important capital fixe et une deuxième catégorie qui détient un capital variable assez important.

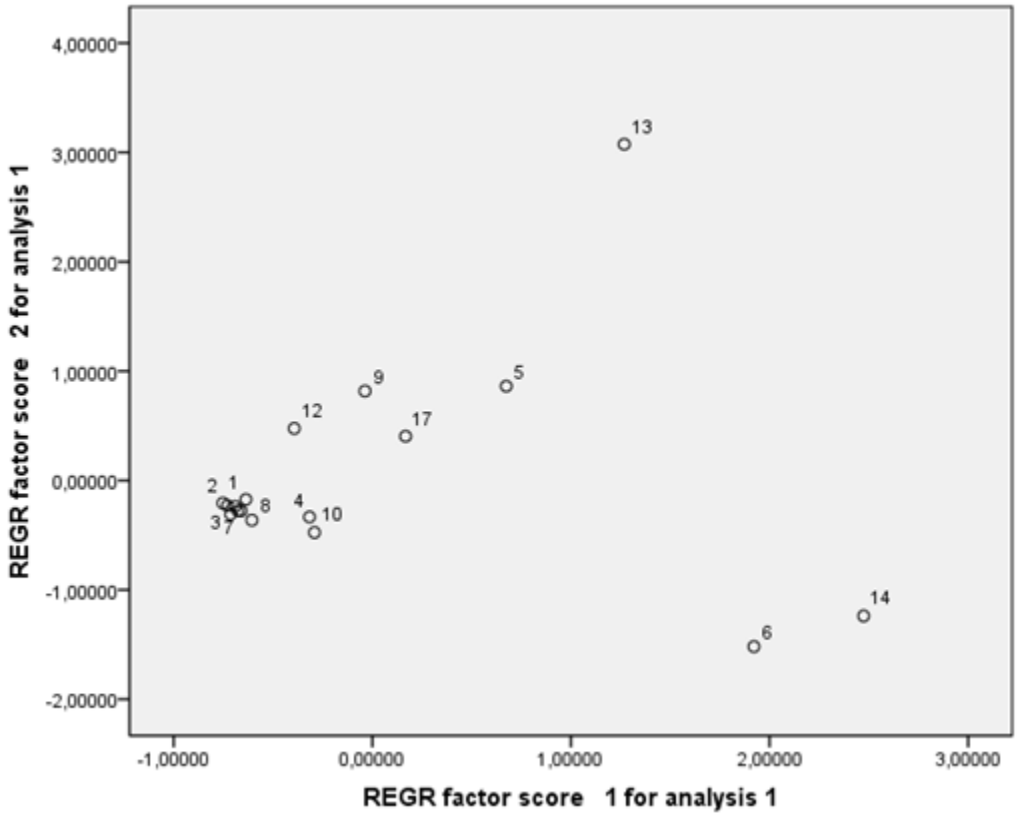

Fig. 3. Représentation des individus

Le graphe des individus montre que les entreprise 6 et 14 représentent les plus grandes entreprises en terme de capital fixe et, l'entreprise 13 représente la plus grande entreprise en terme de capital fixe, les autres entreprises ont un comportement qui n'est pas bien défini.

4 CONCLUSION

Le principe d'une Analyse en composante principale est donc de remplacer les variables initiales, généralement corrélées, par des variables non corrélées de variances progressivement décroissantes, les premières pouvant faire l'objet d'une interprétation particulière et les dernières pouvant être négligées.

REFERENCES

[1] M. M. Tatsuoka: Multivariate analysis, Wiley, 1st edition, 1988.
[2] J.F. Hair et Al: Multivariate data analysis with readings, MacMillan, 7th edition, 1995.
[3] R. Tomassone : Régression nouveaux regard sur une ancienne méthode statistique,2nd édition, Masson, 1992.
[4] B. Escofier et J.Pages : Analyse factorielles simples et multiples. Objectifs, méthodes et interprétation, Dunod, 2nd edition, 1990.
[5] S. Chatterjee and B. Price: Regression analysis by example, 2nd edition, Wiley, 1991.