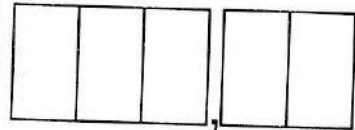


ZOU  
PhilippeL2+CP12  
2013

## DEVOIR ECRIT ANALYSE DE DONNEES - juin 2014

Groupe B

*Sans calculatrice, ni document. Vous composerez sur le sujet. Indiquez votre groupe de TD.***Exercice 1 : Démonstration de cours**

On rappelle les notations utilisées en ACM :

- $n$  : nombre d'individus ou d'unités statistiques
- $q$  : nombre de variables  $Q_k$
- $m_k$  : nombre de modalités de la variable  $Q_k$
- $m$  : nombre total de modalités

- $X_j$  : variable indicatrice  $X_j(i)$  : valeur de la variable indicatrice  $X_j$  pour l'individu  $i$ .
- $n_j$  : effectif de la modalité  $j$
- $p_j$  : fréquence de la modalité  $j$  par rapport à l'effectif total (égal à  $n_j$ )

On rappelle également l'expression de la matrice à diagonaliser en AFC, avec les notations usuelles :

$$T^* = {}^tX^* X^* = \left[ \frac{1}{\sqrt{k_{\cdot i} \times k_{\cdot j}}} \sum_{l=1}^n \frac{k_{li} \times k_{lj}}{k_{l\cdot}} \right]_{\substack{1 \leq i \leq p \\ 1 \leq j \leq p}}$$

1. Rappeler l'expression de l'inertie totale  $R$  dans l'AFC du tableau disjonctif complet.

$$R = \frac{\chi^2}{N} \quad \left( R = \sum d_j^2 = 1 - \text{Tr}(T^*) = 1 - \sum_{j=1}^p T_{jj}^* = 1 \right)$$

2. Démontrer cette formule.

$$T_{jj}^* = {}^tX^* X^* = \left[ \frac{1}{\sqrt{k_{\cdot j} k_{\cdot j}}} \sum_{k=1}^q \frac{k_{kj} \times k_{kj}}{k_{k\cdot}} \right]_{\substack{1 \leq j \leq p \\ 1 \leq j \leq p}}$$

$$= \frac{1}{k_{\cdot j}} \sum_{i=1}^n \frac{k_{ij}^2}{k_{i\cdot}} = \sum_{i=1}^n \frac{k_{ij}^2}{k_{\cdot j} k_{i\cdot}}$$

$$R = \sum_{j=1}^p \sum_{i=1}^n \frac{k_{ij}^2}{k_{i\cdot} k_{\cdot j}} - 1 = \sum_{i=1}^n \sum_{j=1}^p \frac{k_{ij}^2}{k_{i\cdot} k_{\cdot j}} - 1$$

$$\chi^2 = N \sum_{i=1}^n \sum_{j=1}^p \frac{(k_{ij} - k_{i\cdot} k_{\cdot j})^2}{k_{i\cdot} k_{\cdot j}}$$

$$\frac{\chi^2}{N} = \sum_{i=1}^n \sum_{j=1}^p \frac{k_{ij}^2}{k_{i\cdot} k_{\cdot j}} - 2 \sum_{i=1}^n \sum_{j=1}^p \frac{k_{ij}}{k_{i\cdot} k_{\cdot j}} + \sum_{i=1}^n \sum_{j=1}^p \frac{k_{i\cdot} k_{\cdot j}}{k_{i\cdot} k_{\cdot j}}$$

$$\sum_{i=1}^n \sum_{j=1}^p \frac{k_{ij}^2}{k_{i\cdot} k_{\cdot j}} = \sum_{i=1}^n k_{i\cdot} \sum_{j=1}^p \frac{k_{ij}}{k_{\cdot j}} = \sum_{i=1}^n k_{i\cdot} = 1$$

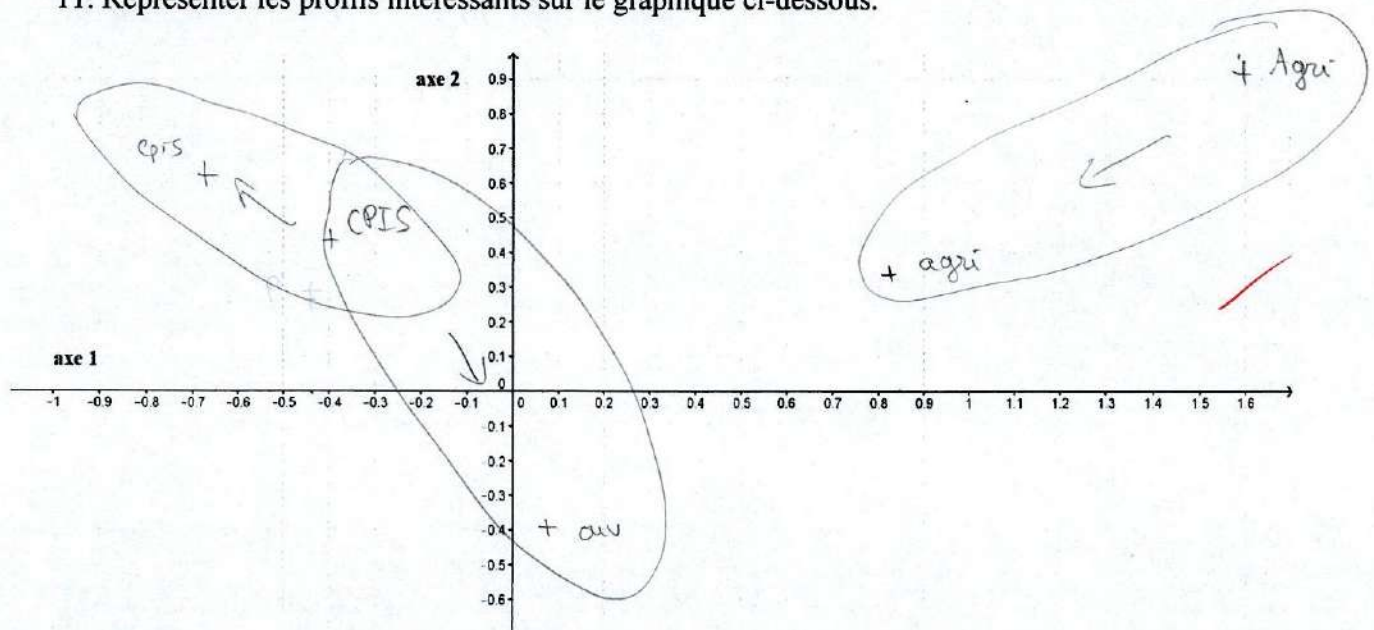
$$\frac{\chi^2}{N} = \sum_{i=1}^n \sum_{j=1}^p \frac{k_{ij}^2}{k_{i\cdot} k_{\cdot j}} - 1 \quad \text{soit } R.$$



## Composantes principales des colonnes

|      | axe 1    |       |       | axe 2    |       |       |
|------|----------|-------|-------|----------|-------|-------|
|      | coord    | qlt   | cr    | coord    | qlt   | cr    |
| AGRI | - 1,602  | 0,786 | 0,642 | 0,836    | 0,214 | 0,294 |
| ACCE | -0,186   | 0,081 | 0,020 | 0,114    | 0,031 | 0,012 |
| CPIS | - -0,486 | 0,499 | 0,227 | - 0,461  | 0,448 | 0,344 |
| PI   | -0,154   | 0,433 | 0,028 | -0,094   | 0,160 | 0,017 |
| EMP  | -0,024   | 0,021 | 0,000 | -0,121   | 0,538 | 0,012 |
| OUV  | 0,228    | 0,296 | 0,083 | - -0,346 | 0,679 | 0,321 |

11. Représenter les profils intéressants sur le graphique ci-dessous.



12. Finir l'interprétation de ces données.

les personnes dont les pères sont agriculteurs ont une proportion plus importante de suivre cette voie que les autres, ce qui peut facilement s'expliquer par la localisation de cette profession.

Cette tendance est également visible chez les autres professions comme chez les cadres.

Cependant, une chose est étonnante. On peut observer de nombreux jeunes devenus ouvriers alors que leurs parents avaient/ont une situation confortable (cadres, chef d'entreprise ...)

On peut également observer une certaine disposition des professions  
sur le graphique.

A droite nous avons les professions les plus "rurales" (Agriculteurs), à gauche, les plus  
"urbaines" (cadres) et au centre, le reste.



**Exercice n°2 :**

Dans le cadre d'une étude portant sur les 28 pays de l'Union Européenne, on considère les variables suivantes :

- Population, en millions d'habitants.
- Superficie, en centaines de milliers de km<sup>2</sup>.
- PIB, en milliards d'euros.
- Population de moins de 20 ans, en millions.
- Population de plus de 60 ans, en millions d'habitants.
- Nombre de chômeurs, en millions.
- Dépenses annuelles d'éducation, en milliards d'euros.
- Dépenses annuelles de protection sociale, en milliards d'euros.
- Nombres de personnes en dessous du seuil de pauvreté, en millions.

1. Quel type d'analyse peut-on réaliser sur ces données ? Pourquoi ?

On peut réaliser un ACP car on est en présence de données quantitatives.

et ! x

2. Quelle problématique peut-on envisager ?

On peut se demander si la géographie des pays (population ou superficie par exemple) est indépendante de sa richesse.

+ (≈)

3. Lors de l'analyse, on met en évidence un axe factoriel dont la contribution à l'inertie totale est d'environ 90%.

a. Était-ce prévisible ? Pourquoi ?

Ce problème était prévisible étant donné que les variables ne sont pas de même dimension.

b. Quel est cet axe ?

d'axe correspondant à la ~~corrélation~~ entre le PIB et les dépenses ~~sera~~ prédominant.

c. Quel problème soulève ce phénomène ?

Ce phénomène implique un écart-type ~~très~~ grand entre les variables.

d. Comment y remédier ?

Il suffit de centrer et réduire les valeurs afin d'obtenir des valeurs plus faciles à analyser.



| $r$ | $\chi^2_{0,995}$ | $\chi^2_{0,99}$ | $\chi^2_{0,975}$ | $\chi^2_{0,95}$ | $\chi^2_{0,90}$ | $\chi^2_{0,75}$ | $\chi^2_{0,50}$ | $\chi^2_{0,25}$ | $\chi^2_{0,10}$ | $\chi^2_{0,05}$ | $\chi^2_{0,025}$ | $\chi^2_{0,01}$ | $\chi^2_{0,005}$ |
|-----|------------------|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|-----------------|------------------|
| 23  | 44,2             | 41,6            | 38,1             | 35,2            | 32,0            | 27,1            | 22,3            | 18,1            | 14,8            | 13,1            | 11,7             | 10,2            | 9,26             |
| 24  | 45,6             | 43,0            | 39,4             | 36,4            | 33,2            | 28,2            | 23,3            | 19,0            | 15,7            | 13,8            | 12,4             | 10,9            | 9,89             |
| 25  | 46,9             | 44,3            | 40,6             | 37,7            | 34,4            | 29,3            | 24,3            | 19,9            | 16,5            | 14,6            | 13,1             | 11,5            | 10,5             |
| 26  | 48,3             | 45,6            | 41,9             | 38,9            | 35,6            | 30,4            | 25,3            | 20,8            | 17,3            | 15,4            | 13,8             | 12,2            | 11,2             |
| 27  | 49,6             | 47,0            | 43,2             | 40,1            | 36,7            | 31,5            | 26,3            | 21,7            | 18,1            | 16,2            | 14,6             | 12,9            | 11,8             |
| 28  | 51,0             | 48,3            | 44,5             | 41,3            | 37,9            | 32,6            | 27,3            | 22,7            | 18,9            | 16,9            | 15,3             | 13,6            | 12,5             |
| 29  | 52,3             | 49,6            | 45,7             | 42,6            | 39,1            | 33,7            | 28,3            | 23,6            | 19,8            | 17,7            | 16,0             | 14,3            | 13,1             |
| 30  | 53,7             | 50,9            | 47,0             | 43,8            | 40,3            | 34,8            | 29,3            | 24,5            | 20,6            | 18,5            | 16,8             | 15,0            | 13,8             |

$$\begin{array}{r} 200 \\ 1000 \\ - 1870 \\ \hline 1300 \\ - 1122 \\ \hline 178 \end{array} \quad \begin{array}{r} 1374 \\ 0,53 \end{array} \quad \begin{array}{r} 119 \\ 1190 \\ - 1122 \\ \hline 680 \\ - 374 \\ \hline 306 \end{array}$$

8. Justifier le choix des deux premiers axes.

En prenant les 2 premiers axes, on remarque que l'axe 1 compose 53% de l'information.

9. Quelle est la différence entre « qlt » et « cr » ? A quoi servent ces paramètres ? et que l'axe 2 pèse 31%.

qlt : qualité de représentation

cr : contribution relative

qlt est établie pour pouvoir choisir les individus. Si la qlt est élevée, on choisira

ce qui représente le poids qu'apporte la modalité <sup>selon</sup> particulière d'information.

A elles deux, nous avons 84% de l'information ce qui est suffisant.  
alors-ci  
à coordonnées  
équivalentes

10. On donne ci - après les tableaux de composantes principales des profils.

Composantes principales des lignes

|      | axe 1  |       |       | axe 2  |       |       |
|------|--------|-------|-------|--------|-------|-------|
|      | coord  | qlt   | cr    | coord  | qlt   | cr    |
| agri | 0,861  | 0,862 | 0,649 | 0,343  | 0,137 | 0,173 |
| acce | -0,269 | 0,193 | 0,050 | 0,149  | 0,059 | 0,026 |
| cpis | -0,640 | 0,493 | 0,172 | 0,604  | 0,440 | 0,258 |
| pi   | -0,425 | 0,599 | 0,094 | 0,253  | 0,212 | 0,056 |
| emp  | -0,246 | 0,549 | 0,034 | -0,073 | 0,048 | 0,005 |
| ouv  | 0,030  | 0,006 | 0,002 | -0,383 | 0,978 | 0,482 |



3

## Exercice n°3 :

Le tableau donné ci-dessous indique pour certains pays certains caractères socio-politiques.

| pays            | continent        | religion majoritaire | peine de mort |
|-----------------|------------------|----------------------|---------------|
| Arabie saoudite | Asie             | islam                | oui           |
| Brésil          | Amérique du Sud  | christianisme        | non           |
| Chine           | Asie             | bouddhisme           | oui           |
| France          | Europe           | christianisme        | non           |
| Égypte          | Afrique          | islam                | oui           |
| Canada          | Amérique du Nord | christianisme        | non           |
| Sénégal         | Afrique          | islam                | non           |
| Etats-Unis      | Amérique du Nord | christianisme        | oui           |
| Russie          | Europe           | christianisme        | oui           |
| Colombie        | Amérique du Sud  | christianisme        | oui           |
| Japon           | Asie             | bouddhisme           | non           |
| Iran            | Asie             | islam                | oui           |

1. Quel est le type d'analyse de données que l'on peut effectuer sur ce type de données ?

on va effectuer une ACM étant donné qu'il n'y a que des données qualitatives

2. Déterminer le tableau disjonctif complet associé à ce tableau de données. Attention, le tableau fourni n'a pas forcément la bonne dimension !

|                 | continent |        |        |         |          | religion |         |       | peine mort |     |  |  |
|-----------------|-----------|--------|--------|---------|----------|----------|---------|-------|------------|-----|--|--|
|                 | Asie      | Am. S. | Europe | Afrique | A. du N. | islam    | chrest. | Boud. | oui        | non |  |  |
| Arabie saoudite | 1         | 0      | 0      | 0       | 0        | 1        | 0       | 0     | 1          | 0   |  |  |
| Brésil          | 0         | 1      | 0      | 0       | 0        | 0        | 1       | 0     | 0          | 1   |  |  |
| Chine           | 1         | 0      | 0      | 0       | 0        | 0        | 0       | 1     | 1          | 0   |  |  |
| France          | 0         | 0      | 1      | 0       | 0        | 0        | 1       | 0     | 0          | 1   |  |  |
| Égypte          | 0         | 0      | 0      | 1       | 0        | 1        | 0       | 0     | 1          | 0   |  |  |
| Canada          | 0         | 0      | 0      | 0       | 1        | 0        | 1       | 0     | 0          | 1   |  |  |
| Sénégal         | 0         | 0      | 0      | 1       | 0        | 1        | 0       | 0     | 0          | 1   |  |  |
| Etats-Unis      | 0         | 0      | 0      | 0       | 1        | 0        | 1       | 0     | 1          | 0   |  |  |
| Russie          | 0         | 0      | 1      | 0       | 0        | 0        | 1       | 0     | 1          | 0   |  |  |
| Colombie        | 0         | 1      | 0      | 0       | 0        | 0        | 1       | 0     | 1          | 0   |  |  |
| Japon           | 1         | 0      | 0      | 0       | 0        | 0        | 0       | 1     | 0          | 1   |  |  |
| Iran            | 1         | 0      | 0      | 0       | 0        | 1        | 0       | 0     | 1          | 0   |  |  |

3. Quel est l'inertie totale de ces données ?

$$R = \sum_{up} - 1$$

$$= \frac{m}{q} - 1$$

$$= \frac{10}{3} - 1 = 3,333 - 1$$

$$= 2,333$$

$$\sum_{up} = \frac{m}{q} \leftarrow \begin{array}{l} \text{nombre de modalités} \\ \text{nombre de variables} \end{array}$$



justification :

$$\begin{array}{r} 258 \\ 2580 \\ - 2400 \\ \hline 1800 \end{array}$$

$$\frac{300}{0,86}$$

Tableau des profils ..... *données*

| PFC  | AGRI  | ACCE  | CPIS  | PI           | EMP   | OUV   | PFC moyen |
|------|-------|-------|-------|--------------|-------|-------|-----------|
| agri | 0,860 | 0,119 | 0,094 | 0,109        | 0,148 | 0,191 | 0,17      |
| acce | 0,047 | 0,362 | 0,156 | 0,120        | 0,099 | 0,087 | 0,13      |
| cpis | 0,010 | 0,079 | 0,230 | <u>0,074</u> | 0,074 | 0,018 | 0,08      |
| pi   | 0,017 | 0,068 | 0,195 | 0,136        | 0,107 | 0,051 | 0,10      |
| emp  | 0,003 | 0,072 | 0,128 | 0,154        | 0,130 | 0,094 | 0,11      |
| ouv  | 0,063 | 0,300 | 0,197 | 0,406        | 0,442 | 0,559 | 0,38      |
|      |       |       |       |              |       |       |           |

3. Quelle est la signification des nombres soulignés dans les 3 tableaux précédents ?

0,107 et 0,07 correspondent aux nombres les plus proches des PFC / PFC moyens.

4. Faire une analyse des PFC et des PFL. Vous pourrez utiliser les tableaux précédents et faire une synthèse.

PFL

⊕

⊖

5. Va-t-on effectuer l'AFC à partir des PFL ou à partir des PFC ? Argumenter.

6. Quelle est le nombre de valeurs propres que l'on va obtenir ?

En AFC, le nombre de variables donne le nombre de valeurs propres : 6

7. On obtient alors les valeurs propres suivantes : 0,200 ; 0,119 ; 0,050 ; 0,005. Effectuer un test d'indépendance de ces données. Que peut-on en déduire? *il y a 2 valeurs propres faibles 0 et 0.*

$$R = \sum \lambda_p = 0,2 + 0,119 + 0,050 + 0,005$$

$$= 0,374$$

$$\chi^2 = R \times N$$

$$= 0,374 \times 6012$$

$$= 2253,488$$

Pour n'importe quel seuil,  $\chi^2_{obs} > \chi^2_{théorique}$ .  
des données sont donc dépendantes.

m<sup>1</sup> de degré?   
 X



## 4. Construire le tableau de BURT associé à ces données.

|         | Asie | A.S | Europe | Afrique | A-Nord | Islam | christ | Bud | oui | Non |
|---------|------|-----|--------|---------|--------|-------|--------|-----|-----|-----|
| Asie    | 4    | 0   | 0      | 0       | 0      | 2     | 0      | 2   | 3   | 1   |
| A.Sid   | 0    | 2   | 0      | 0       | 0      | 0     | 2      | 0   | 1   | 1   |
| Europe  | 0    | 0   | 2      | 0       | 0      | 0     | 2      | 0   | 1   | 1   |
| Afrique | 0    | 0   | 0      | 2       | 0      | 2     | 0      | 0   | 1   | 1   |
| A-Nord  | 0    | 0   | 0      | 0       | 2      | 0     | 2      | 0   | 1   | 1   |
| Islam   | 2    | 0   | 0      | 2       | 0      | 4     | 0      | 0   | 3   | 1   |
| christ  | 0    | 2   | 2      | 0       | 2      | 0     | 6      | 0   | 3   | 3   |
| Bud     | 2    | 0   | 0      | 0       | 0      | 0     | 0      | 2   | 1   | 1   |
| oui     | 3    | 1   | 1      | 1       | 1      | 3     | 3      | 1   | 7   | 0   |
| Non     | 1    | 1   | 1      | 1       | 1      | 1     | 3      | 1   | 0   | 5   |

## 5. Lorsque l'on réalise cette analyse, est-il préférable de travailler à partir du tableau disjonctif complet ou à partir du tableau de Burt ? Argumenter.

Il est préférable de travailler avec le tableau de Burt car elle correspond à une matrice carrée et symétrique.

## Exercice n°4 :

On effectue une AFC sur une population de jeunes adultes, les deux variables étudiées étant la profession du père dont les modalités sont en majuscule et la profession du jeune dont les modalités sont en minuscule. Les diverses modalités sont :

- **AGRI**culteur
- Artisan, Commerçant, Chef d'Entreprise
- Cadre et Professions Intellectuelles Supérieures
- Profession Intermédiaire
- **EMP**loyé
- **OUV**rier

| effectifs | AGRI | ACCE | CPIS | PI   | EMP | OUV  | total |
|-----------|------|------|------|------|-----|------|-------|
| agri      | 258  | 81   | 108  | 153  | 84  | 365  | 1049  |
| acce      | 14   | 246  | 180  | 168  | 56  | 167  | 831   |
| cpis      | 3    | 54   | 266  | 104  | 42  | 34   | 503   |
| pi        | 5    | 46   | 225  | 190  | 61  | 97   | 624   |
| emp       | 1    | 49   | 148  | 215  | 74  | 180  | 667   |
| ouv       | 19   | 204  | 228  | 568  | 251 | 1068 | 2338  |
| total     | 300  | 680  | 1155 | 1398 | 568 | 1911 | 6012  |



| fréquence | AGRI  | ACCE  | CPIS  | PI    | EMP   | OUV          | total |
|-----------|-------|-------|-------|-------|-------|--------------|-------|
| agri      | 0,043 | 0,013 | 0,018 | 0,025 | 0,014 | 0,061        | 0,174 |
| acce      | 0,002 | 0,041 | 0,030 | 0,028 | 0,009 | 0,028        | 0,138 |
| cpis      | 0,000 | 0,009 | 0,044 | 0,017 | 0,007 | <u>0,006</u> | 0,084 |
| pi        | 0,001 | 0,008 | 0,037 | 0,032 | 0,010 | 0,016        | 0,104 |
| emp       | 0,000 | 0,008 | 0,025 | 0,036 | 0,012 | 0,030        | 0,111 |
| ouv       | 0,003 | 0,034 | 0,038 | 0,094 | 0,042 | 0,178        | 0,389 |
| total     | 0,050 | 0,113 | 0,192 | 0,233 | 0,094 | 0,318        | 1,000 |

1. Quelle est la problématique de l'étude ? On s'intéresse à l'indépendance  de la profession père influence celle des fils.
2. Les tableaux suivants sont les tableaux des profils : l'un est celui des PFL, l'autre celui des PFC.
  - a. Rappeler ce que sont les PFL et les PFC.

des PFL correspondent aux profils lignes

des PFC correspondent aux profils colonnes

- b. Identifier chacun de ces tableaux en expliquant vos raisons et vous indiquerez les PFC ou PFL moyens dans le tableau correspondant.

Tableau des profils ..... lignes

| PFL       | AGRI  | ACCE         | CPIS  | PI    | EMP   | OUV   |  |
|-----------|-------|--------------|-------|-------|-------|-------|--|
| agri      | 0,246 | 0,077        | 0,103 | 0,146 | 0,080 | 0,348 |  |
| acce      | 0,017 | 0,296        | 0,217 | 0,202 | 0,067 | 0,201 |  |
| cpis      | 0,006 | <u>0,107</u> | 0,529 | 0,207 | 0,083 | 0,068 |  |
| pi        | 0,008 | 0,074        | 0,361 | 0,304 | 0,098 | 0,155 |  |
| emp       | 0,001 | 0,073        | 0,222 | 0,322 | 0,111 | 0,270 |  |
| ouv       | 0,008 | 0,087        | 0,098 | 0,243 | 0,107 | 0,457 |  |
| PFL moyen | 0,049 | 0,11         | 0,19  | 0,23  | 0,09  | 0,31  |  |

justificatif :

$$\begin{array}{r}
 258 \\
 2580 \\
 - 2098 \\
 \hline
 4820 \\
 - 4196 \\
 \hline
 06240 \\
 - 5255 \\
 \hline
 0985
 \end{array}$$

11049  
0,245  $\approx$  0,246

