Examen MAT 5201 DATA MINING

Mardi 08 Novembre 2011

Première Partie : 15 minutes (7 points)

Enseignant responsable : Frédéric Bertrand

Remarque importante : les questions de ce questionnaire sont posées dans le contexte d'un cours de DATA MINING. Une seule réponse est correcte par question.

1. **Que signifie ACP ?**
   a) **Analyse des correspondances partielles**
   b) **Analyse de classification prépondérante**
   c) **Analyse en composantes principales**
2. **Que signifie CRM ?**
   a) **Centre de Recherche en Mathématiques**
   b) **Classification des Relations Maximales**
   c) **Customer Relationship Management**
3. **Que signifie PLS ?**
   a) **Partial least squares**
   b) **Prévision linéaire simple**
   c) **Partitionnements logiques successifs**
4. **Le Data Mining est-il utilisé en CRM ? Donnez un exemple de problématique liée à son utilisation.**
   a) **Non. ……**
   b) **Oui. ……..**
5. **Combien de grandes familles de techniques de DATA MINING sont présentées dans ce cours ? Citez-les dans le cas que vous avez choisi.**
   a) **Une. ……**
   b) **Deux. ……**
   c) **Trois. ……**
6. **Les techniques factorielles sont-elles utiles en Data Mining ? Si oui, donnez un exemple de problématique liée à son utilisation.**
   a) **Non. ……**
   b) **Oui. ……..**
7. **Pour étudier les habitudes de consommation des clients d'un supermarché, on utilisera**
   a) **Des règles d'association**
   b) **Des règles de dissociation**
   c) **Une technique prédictive**
8. **La confiance d'une règle d'association est la proportion de transactions contenant les items impliqués auxquels s'applique la règle :**
   a) **Vrai**

b) **Faux**

9. **Les arbres de décisions permettent d'obtenir des prédictions pour une variable**
   a) **Quantitative**
   b) **Qualitative**
   c) **Quantitative et qualitative.**
10. **Les arbres de décisions permettent d'obtenir des prédictions à l'aide de variables**
   a) **Quantitatives**
   b) **Qualitatives**
   c) **Quantitatives et qualitatives.**
11. **Une ACP se réalise**
   a) **Sur des variables qualitatives**
   b) **Sur des variables quantitatives**
   c) **Sur des variables mixtes**
12. **La commande sous R pour ajuster un arbre de régression est**
   a) **tree()**
   b) **tree.cv()**
   c) **prune.tree()**
13. **La commande sous R pour extraire des règles d'association est**
   a) **read.translations()**
   b) **inspect()**
   c) **apriori()**
14. **La commande sous R pour réaliser une ACP est**
   a) **res.pca()**
   b) **PCA()**
   c) **plot()**

E.S.I.E.A Paris
Année scolaire 2011/2012

UE de cinquième année : **MAT 5201 - Data Mining**
Enseignant Responsable : F. Bertrand

Chaque réponse devra être justifiée précisément. Dans le corps du texte et en annexe sont donnés le journal et la sortie d'un traitement avec le logiciel R.

---

1 L'immobilier à Boston (4 points)

Nous disposons de plusieurs variables qui ont été relevées dans 506 localités des environs de Boston. Nous souhaitons nous servir pour prédire la valeur médiane des logements, exprimées en milliers de dollards US. Les Figures 1 et 2, montrent les résultats obtenus lorsque nous n'utilisons que la position géographique, longitude et latitude. Les Figures 3 et 4, montrent les résultats obtenus lorsque nous ajoutons toutes les variables suivantes à la longitude et à la lattitude :

- `crim`, le nombre d'acte de criminalité par habitant ;
- `zn`, la proportion de la surface résidentielle occupée par des habitations de plus de 2300 m^2 ;
- `indus`, la proportion de la surface de la localité occupée par des industries ;
- `chas`, une variable indiquant si la localité est située au bord de la rivière Charles, 0 sinon ;
- `nox`, la concentration monoxide d'azote (par 10 million) ;
- `rm`, le nombre moyen de pièces par habitation ;
- `age`, proportion des logements construits avant 1940 et occupés par leurs propriétaires ;
- `dis`, distances pondérées jusqu'à cinq foyers d'emplois ;
- `rad`, indice d'accessibilité aux autoroutes périphériques ;
- `tax`, tax-foncière en 10000 dollards US ;
- `ptratio`, nombre moyen d'élève par enseignant ;
- `lstat`, pourcentage de la population ayant de faibles revenus présent dans la ville.

1. Combien de noeuds intérieurs possède l'arbre `treefit` ? Combien de feuilles a l'arbre `treefit2` ?

2. Déterminer les valeurs médianes prédites pour les deux arbres pour la localité 1 et la localité 100 dont les caractéristiques sont les suivantes :

```
       crim zn indus chas   nox    rm  age    dis rad tax ptratio
1   0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3
```

```
100 0.06860  0  2.89    0 0.445 7.416 62.5 3.4952   2 276    18.0
    lstat     lon    lat
1    4.98 -70.955 42.255
100  6.19 -71.080 42.268
```

3. Comparer les valeurs médianes prédites à la question **2.** avec les valeurs médianes réelles de la localité 1 et de la localité 100 qui sont données ci-dessous :

```
    ValeurMédiane
1        3.178054
100      3.502550
```

Quelles sont les meilleures prédictions ? En se basant sur ce constat, quel est l'arbre que vous préférerez utiliser ?

4. Parmi toutes les variables explicatives proposées quelles sont celles qui sont utilisées par le modèle que vous avez retenu à la question **3.** ?

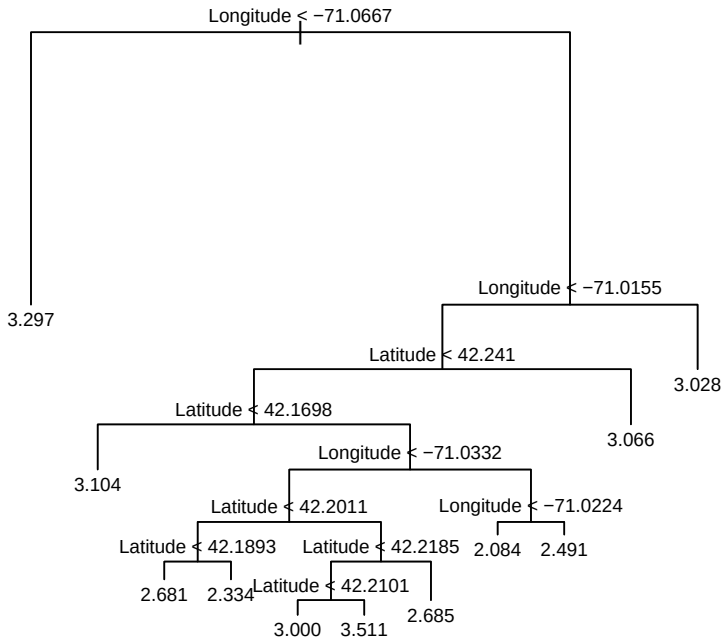

FIGURE 1 – Arbre de régression pour prédire le prix médian d'un logement à Boston à partir de l'emplacement géographique. Les feuilles sont étiquettées avec le logarithme népérien du prix médian d'un logement.

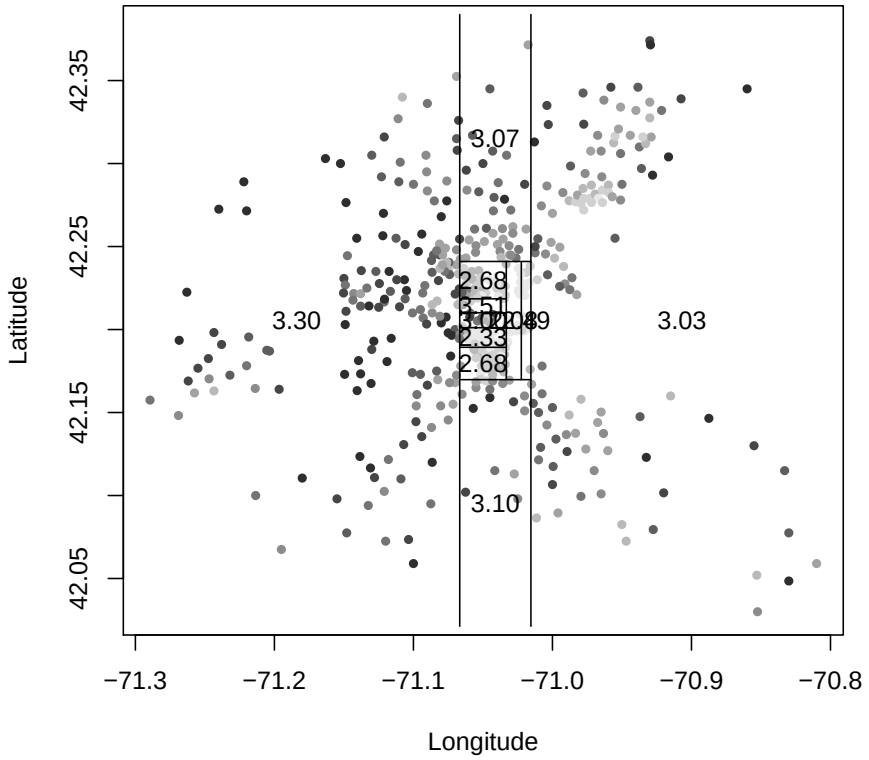

FIGURE 2 – Carte de la valeur réelle des logements (les nuances de gris correspondent aux déciles, plus le gris est foncé plus le prix est élevé) et la partition de l'arbre `treefit`.

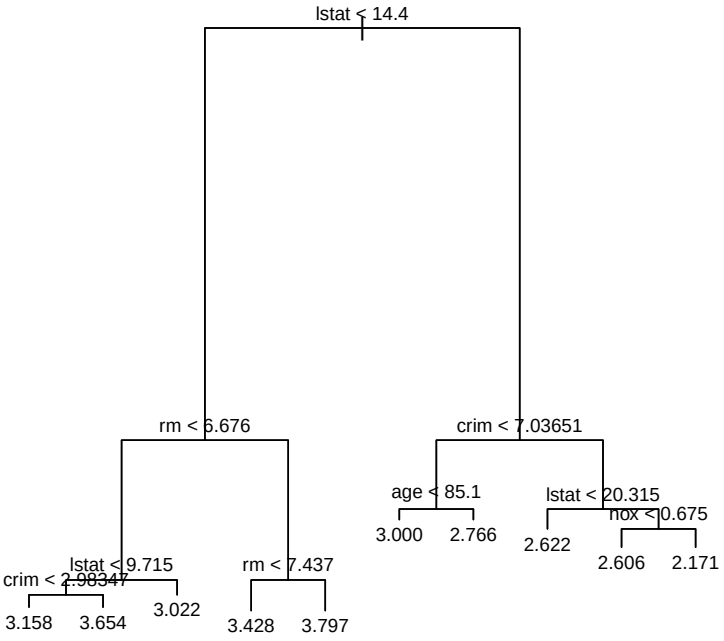

FIGURE 3 – Arbre de régression pour prédire le prix médian des logements à Boston à partir de l'emplacement géographique. Les feuilles sont étiquettées avec le logarithme népérien du prix médian.

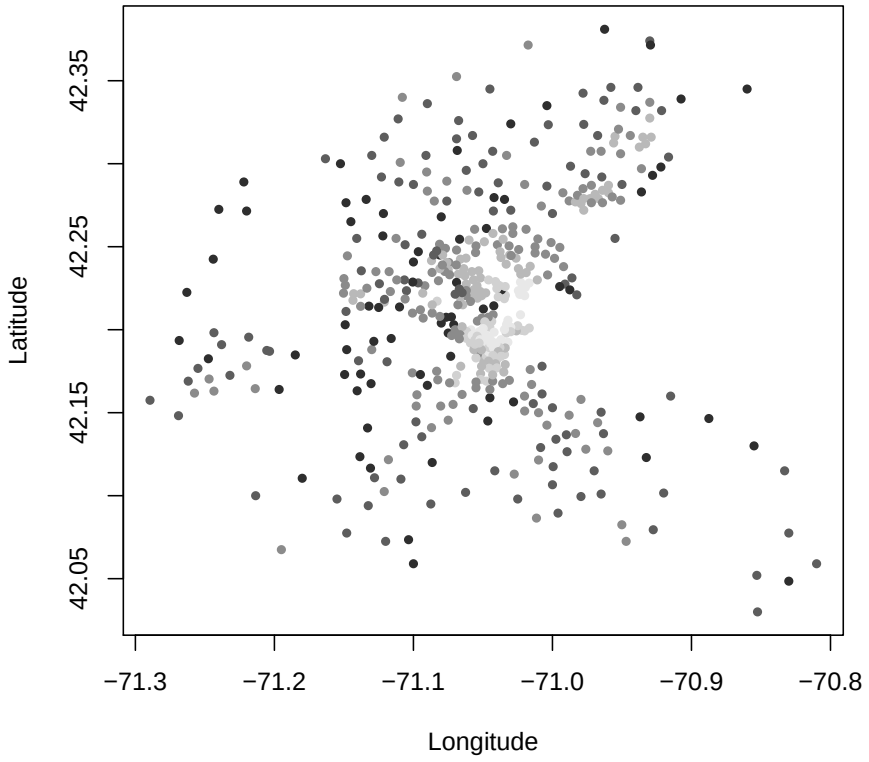

FIGURE 4 – Carte des prédiction du prix médian des logements associée à l'arbre `treefit2` créé à partir de toutes les variables disponibles.

2 Visualisation des règles d'association (4 points)

Nous avons procédé à l'analyse des tickets de caisse des ventes dans un supermarché en Angleterre.

```
parameter specification:
 confidence minval smax arem  aval originalSupport support minlen
        0.8    0.1    1 none FALSE            TRUE   0.001      1
 maxlen target   ext
     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 done [0.02s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].

set of 410 rules
```

Grâce à la Figure 5, nous remarquons qu'un petit nombre de règles, celles associées aux points les plus foncés, ont un lift plus élevé que les autres. Voici les trois règles ayant le lift le plus élevé. Une représentation graphique des objets qui composent ces règles a été construite à la Figure 6.

```
  lhs                          rhs                    support confidence     lift
1 {liquor,
   red/blush wine}          => {bottled beer}     0.001931876  0.9047619 11.23527
2 {citrus fruit,
   other vegetables,
   soda,
   fruit/vegetable juice} => {root vegetables} 0.001016777  0.9090909  8.34040
3 {tropical fruit,
   other vegetables,
   whole milk,
   yogurt,
   oil}                   => {root vegetables} 0.001016777  0.9090909  8.34040
```

1. Combien de transactions ont-elles été étudiées ? Sur combien de produits portaient-elles ?
2. Combien de règles ont-elles été trouvées ? Quelles étaient les valeurs de confiance et de support que devaient vérifier ces règles ?

3. Quel est la règle ayant le lift le plus élevé ? Donner son support et sa confiance.

4. Quel(s) est(sont) le(s) objet(s) dont l'achat entraine celui de `root vegetables` ?

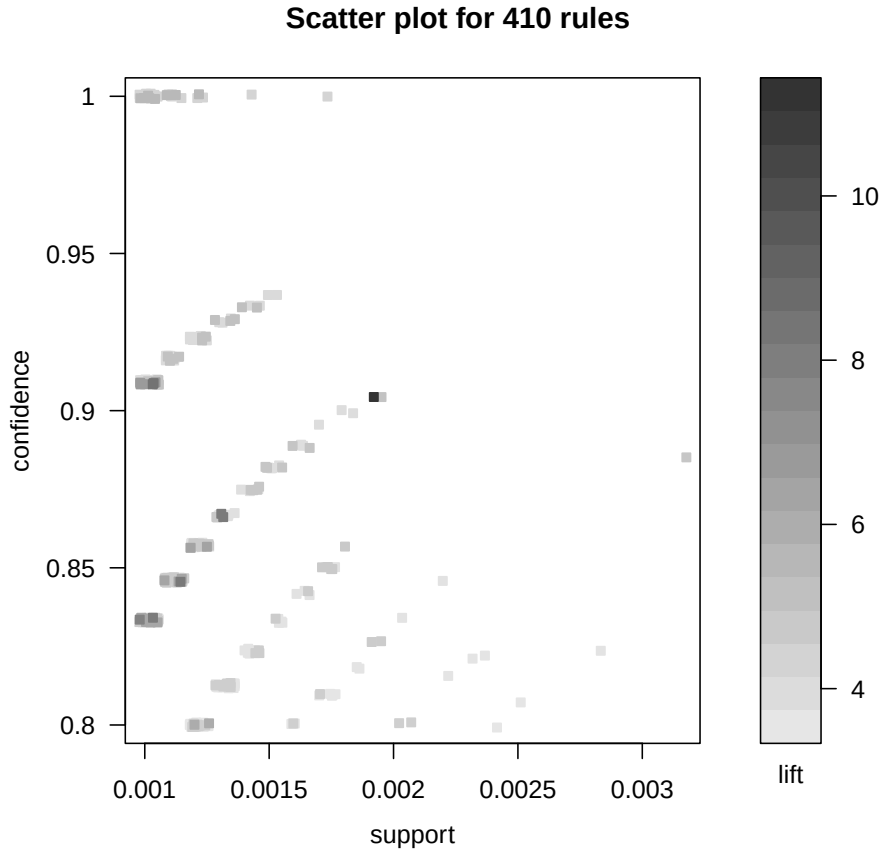

FIGURE 5 – Représentation graphique du lift des règles d'association en fonction de leur support et de leur confiance.

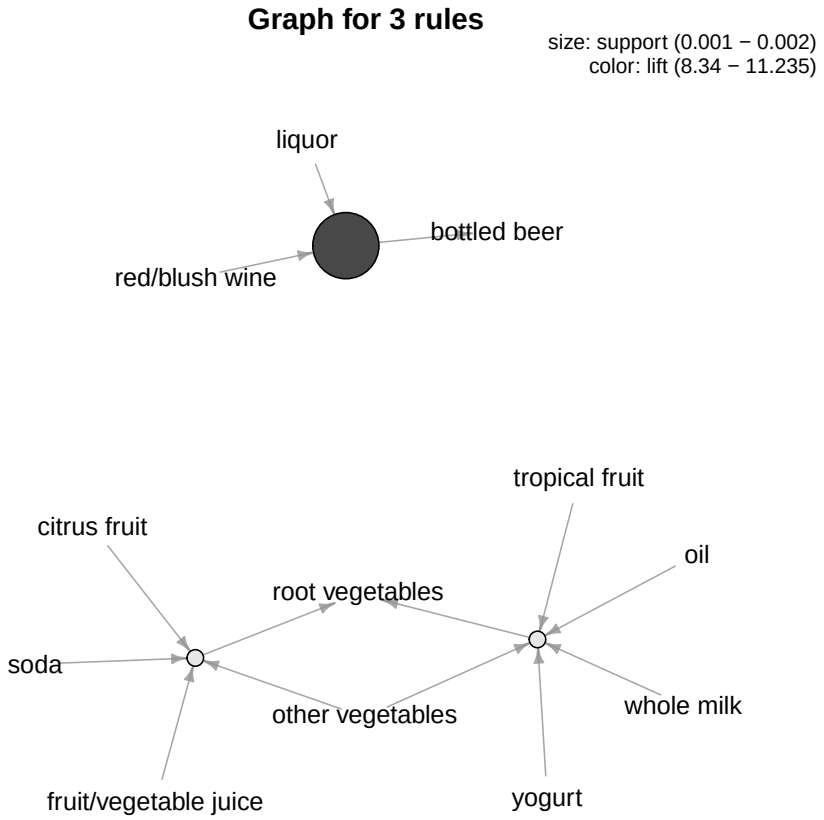

FIGURE 6 – Représentation graphique des trois règles ayant le lift le plus élevé et des objets qui les composent.

3 Inflation et chômage en Europe (5 points)

Il est souvent affirmé par les économistes qu'il existe un lien entre inflation et chômage. Pour en avoir le coeur net, nous avons réuni, pour les années allant de 1998 à 2004, le taux d'inflation des prix à la consommation et le taux de chômage pour 18 pays industrialisés (source Eurostat). Ces données sont regroupées dans le Tableau 1.

1. Décrire le jeu de données (nombre d'individus, nombre de variables, nature des variables). Quelles sont les variables qui ont été utilisées pour réaliser l'ACP ?
2. On veut effectuer une ACP sur ce jeu de données : quels sont les objectifs d'une telle analyse ?
3. Les variables ont été centrées et réduites avant l'analyse. La réduction était-elle indispensable ? Vous pourrez justifier ce choix en utilisant, par exemple, les informations contenues dans les tableaux 2, 6, 3, 4 et 5.

Les tableaux 7, 8, 9, 10, 11, 12 et 13 donnent les PRINCIPAUX résultats de l'ACP sur les individus, les variables et les variables mises en supplémentaire. La figure 7 donne le graphe des individus de l'ACP. La figure 8 donne le graphe des variables.

4. Quelle est l'inertie expliquée par le premier axe de l'ACP ? Et par le premier plan ?
5. VRAI ou FAUX ? Si FAUX, corriger la phrase proposée.
   - Le Japon est bien représenté sur l'axe 2.
   - La variable i2003 a joué le rôle le plus important dans la construction de l'axe 1.
   - Pour chacune des pays, le taux d'inflation en 1999 est fortement corrélée au taux d'inflation en 2001.
   - Si l'inflation a été élevée en 1998, elle l'a aussi été en 2004.
   - Le Japon et l'Irlande ont des profils d'inflation similaires.

	i1998	i1999	i2000	i2001	i2002	i2003	i2004	c1998	c1999	c2000	c2001	c2002	c2003	c2004
Allemagne	0.60	0.60	1.40	1.90	1.30	1.00	1.80	8.80	7.90	7.20	7.40	8.20	9.00	9.50
Autriche	0.80	0.50	2.00	2.30	1.70	1.30	2.00	4.50	3.90	3.70	3.60	4.20	4.30	4.50
Belgique	0.90	1.10	2.70	2.40	1.60	1.50	1.90	9.30	8.60	6.90	6.70	7.30	8.00	7.80
Danemark	1.30	2.10	2.70	2.30	2.40	2.00	0.90	4.90	4.80	4.40	4.30	4.60	5.60	5.40
Espagne	1.80	2.20	3.50	2.80	3.60	3.10	3.10	15.20	12.80	11.30	10.60	11.30	11.30	10.80
Etats-Unis	1.60	2.20	3.40	2.80	1.60	2.30	3.30	4.50	4.20	4.00	4.80	5.80	6.00	5.50
Finlande	1.40	1.30	3.00	2.70	2.00	1.30	0.10	11.40	10.20	9.80	9.10	9.10	9.00	8.80
France	0.70	0.60	1.80	1.80	1.90	2.20	2.30	11.10	10.50	9.10	8.40	8.90	9.50	9.70
Grèce	4.50	2.10	2.90	3.70	3.90	3.40	3.00	10.90	12.00	11.30	10.80	10.30	9.70	10.50
Irlande	2.10	2.50	5.30	4.00	4.70	4.00	2.30	7.50	5.60	4.30	3.90	4.30	4.60	4.50
Italie	2.00	1.70	2.60	2.30	2.60	2.80	2.30	11.30	10.90	10.10	9.10	8.60	8.40	8.00
Japon	0.60	-0.30	-0.70	-0.70	-0.90	-0.30	-0.10	4.10	4.70	4.70	5.00	5.40	5.30	4.70
Luxembourg	1.00	1.00	3.80	2.40	2.10	2.50	3.20	2.70	2.40	2.30	2.10	2.80	3.70	4.20
Norvège	2.00	2.10	3.00	2.70	0.80	2.00	0.60	3.20	3.20	3.40	3.60	3.90	4.50	4.40
Pays-Bas	1.80	2.00	2.30	5.10	3.90	2.20	1.40	3.80	3.20	2.80	2.20	2.80	3.70	4.60
Portugal	2.20	2.20	2.80	4.40	3.70	3.30	2.50	5.10	4.50	4.10	4.00	5.00	6.30	6.70
Royaume-Unis	1.60	1.30	0.80	1.20	1.30	1.40	1.30	6.20	5.90	5.40	5.00	5.10	4.90	4.70
Suède	1.00	0.60	1.30	2.70	2.00	2.30	1.00	8.20	6.70	5.60	4.90	4.90	5.60	6.30
UE(15pays)	1.30	1.20	1.90	2.20	2.10	2.00	2.00	9.30	8.50	7.60	7.20	7.60	7.90	8.00
UE(25pays)	2.10	1.60	2.40	2.50	2.10	1.90	2.10	9.50	9.10	8.60	8.40	8.70	8.90	9.00

TABLE 1 – Données brutes sur l'inflation et le chômage

	i1998	i1999	i2000	i2001
1	Min. :-1.041e+00	Min. :-2.163e+00	Min. :-2.444e+00	Min. :-2.614e+00
2	1st Qu. :-6.847e-01	1st Qu. :-9.150e-01	1st Qu. :-4.829e-01	1st Qu. :-2.376e-01
3	Median :-5.478e-02	Median : 8.318e-02	Median : 1.709e-01	Median :-3.960e-02
4	Mean :-3.108e-18	Mean : 4.626e-17	Mean :-8.791e-17	Mean :-8.944e-17
5	3rd Qu. : 4.382e-01	3rd Qu. : 8.318e-01	3rd Qu. : 4.017e-01	3rd Qu. : 1.584e-01
6	Max. : 3.232e+00	Max. : 1.331e+00	Max. : 2.171e+00	Max. : 1.980e+00

Table 2 – Statistiques descriptives sur les taux d'inflation

	i1998	i1999	i2000	i2001	i2002	i2003	i2004
i1998	1.00	0.68	0.40	0.54	0.60	0.65	0.34
i1999	0.68	1.00	0.75	0.73	0.72	0.77	0.39
i2000	0.40	0.75	1.00	0.67	0.70	0.78	0.54
i2001	0.54	0.73	0.67	1.00	0.86	0.75	0.39
i2002	0.60	0.72	0.70	0.86	1.00	0.88	0.53
i2003	0.65	0.77	0.78	0.75	0.88	1.00	0.66
i2004	0.34	0.39	0.54	0.39	0.53	0.66	1.00

Table 3 – Corrélation entre les taux l'inflation

	c1998	c1999	c2000	c2001	c2002	c2003	c2004
c1998	1.00	0.97	0.95	0.92	0.92	0.91	0.89
c1999	0.97	1.00	0.99	0.97	0.96	0.94	0.92
c2000	0.95	0.99	1.00	0.99	0.97	0.94	0.92
c2001	0.92	0.97	0.99	1.00	0.99	0.95	0.93
c2002	0.92	0.96	0.97	0.99	1.00	0.98	0.95
c2003	0.91	0.94	0.94	0.95	0.98	1.00	0.98
c2004	0.89	0.92	0.92	0.93	0.95	0.98	1.00

Table 4 – Corrélation entre les taux de chômage

	i1998	i1999	i2000	i2001	i2002	i2003	i2004	c1998	c1999	c2000	c2001	c2002	c2003	c2004
Allemagne	-1.04	-1.04	-0.83	-0.55	-0.69	-1.11	-0.03	0.40	0.34	0.36	0.56	0.76	1.00	1.20
Autriche	-0.82	-1.16	-0.37	-0.24	-0.40	-0.81	0.17	-0.80	-0.87	-0.82	-0.82	-0.79	-0.99	-0.94
Belgique	-0.71	-0.42	0.17	-0.16	-0.47	-0.62	0.07	0.54	0.55	0.26	0.31	0.41	0.58	0.47
Danemark	-0.27	0.83	0.17	-0.24	0.12	-0.13	-0.89	-0.69	-0.60	-0.58	-0.57	-0.64	-0.44	-0.56
Espagne	0.27	0.96	0.79	0.16	1.02	0.96	1.22	2.19	1.82	1.74	1.72	1.96	1.98	1.76
Etats-Unis	0.05	0.96	0.71	0.16	-0.47	0.17	1.41	-0.80	-0.78	-0.72	-0.39	-0.17	-0.27	-0.51
Finlande	-0.16	-0.17	0.40	0.08	-0.17	-0.81	-1.66	1.13	1.03	1.23	1.18	1.10	1.00	0.90
France	-0.93	-1.04	-0.52	-0.63	-0.25	0.07	0.45	1.04	1.13	1.00	0.92	1.03	1.22	1.29
Grèce	3.23	0.83	0.32	0.87	1.24	1.25	1.13	0.99	1.58	1.74	1.80	1.57	1.30	1.63
Irlande	0.60	1.33	2.17	1.11	1.83	1.84	0.45	0.04	-0.36	-0.62	-0.71	-0.76	-0.86	-0.94
Italie	0.49	0.33	0.09	-0.24	0.27	0.66	0.45	1.10	1.25	1.33	1.18	0.91	0.75	0.56
Japon	-1.04	-2.16	-2.44	-2.61	-2.33	-2.39	-1.85	-0.91	-0.63	-0.48	-0.31	-0.33	-0.57	-0.86
Luxembourg	-0.60	-0.54	1.02	-0.16	-0.10	0.37	1.32	-1.31	-1.32	-1.29	-1.37	-1.34	-1.25	-1.07
Norvège	0.49	0.83	0.40	0.08	-1.07	-0.13	-1.18	-1.17	-1.08	-0.92	-0.82	-0.91	-0.91	-0.99
Pays-Bas	0.27	0.71	-0.14	1.98	1.24	0.07	-0.41	-1.00	-1.08	-1.12	-1.33	-1.34	-1.25	-0.90
Portugal	0.71	0.96	0.25	1.43	1.09	1.15	0.65	-0.64	-0.69	-0.68	-0.68	-0.48	-0.14	0.00
Royaume-Unis	0.05	-0.17	-1.29	-1.11	-0.69	-0.72	-0.51	-0.33	-0.27	-0.25	-0.31	-0.45	-0.74	-0.86
Suède	-0.60	-1.04	-0.91	0.08	-0.17	0.17	-0.80	0.23	-0.02	-0.18	-0.35	-0.52	-0.44	-0.17

TABLE 5 – Données centrées-réduites sur l'inflation et le chômage

	i2002	i2003	i2004
1	Min. :-2.329e+00	Min. :-2.386e+00	Min. :-1.852e+00
2	1st Qu. :-4.708e-01	1st Qu. :-6.908e-01	1st Qu. :-7.233e-01
3	Median :-1.735e-01	Median : 7.099e-02	Median : 1.174e-01
4	Mean :-3.703e-17	Mean : 1.565e-18	Mean : 8.674e-18
5	3rd Qu. : 8.301e-01	3rd Qu. : 5.870e-01	3rd Qu. : 5.978e-01
6	Max. : 1.834e+00	Max. : 1.840e+00	Max. : 1.414e+00

TABLE 6 – Statistiques descriptives sur les taux de chômage

	comp 1	comp 2	comp 3	comp 4
eigenvalue	4.88	0.78	0.57	0.39
percentage of variance	69.75	11.09	8.19	5.57
cumulative percentage of variance	69.75	80.84	89.03	94.60

	comp 5	comp 6	comp 7
eigenvalue	0.18	0.13	0.07
percentage of variance	2.50	1.88	1.01
cumulative percentage of variance	97.10	98.99	100.00

TABLE 7 – Inertie des sept axes

	Dim.1	Dim.2	Dim.3
Allemagne	-2.11	0.60	-0.11
Autriche	-1.47	0.75	-0.19
Belgique	-0.86	0.53	-0.46
Danemark	-0.05	-0.75	-0.67
Espagne	2.07	0.78	0.17
Etats-Unis	1.05	0.98	0.25
Finlande	-0.86	-1.22	-1.05
France	-1.13	1.15	0.09
Grèce	3.31	-0.90	2.25
Irlande	3.77	0.09	-0.93
Italie	0.79	0.17	0.58
Japon	-5.84	-0.54	0.76
Luxembourg	0.43	1.80	-0.19
Norvège	-0.17	-1.32	-0.30
Pays-Bas	1.56	-1.15	-0.82
Portugal	2.46	-0.25	0.05
Royaume-Unis	-1.76	-0.50	0.83
Suède	-1.19	-0.22	-0.28

TABLE 8 – Coordonnées des individus sur les trois premiers axes

	Dim.1	Dim.2	Dim.3
Allemagne	0.86	0.07	0.00
Autriche	0.66	0.17	0.01
Belgique	0.52	0.20	0.15
Danemark	0.00	0.32	0.25
Espagne	0.80	0.11	0.01
Etats-Unis	0.28	0.25	0.02
Finlande	0.19	0.38	0.28
France	0.42	0.43	0.00
Grèce	0.63	0.05	0.29
Irlande	0.89	0.00	0.05
Italie	0.51	0.02	0.28
Japon	0.97	0.01	0.02
Luxembourg	0.05	0.85	0.01
Norvège	0.01	0.45	0.02
Pays-Bas	0.37	0.20	0.10
Portugal	0.89	0.01	0.00
Royaume-Unis	0.70	0.06	0.16
Suède	0.45	0.02	0.02

TABLE 9 – Cosinus^2 des individus sur les trois premiers axes

	Dim.1	Dim.2	Dim.3
Allemagne	5.07	2.60	0.12
Autriche	2.45	4.00	0.35
Belgique	0.83	2.01	2.06
Danemark	0.00	4.07	4.32
Espagne	4.90	4.37	0.28
Etats-Unis	1.25	6.89	0.60
Finlande	0.85	10.61	10.61
France	1.46	9.46	0.08
Grèce	12.50	5.84	49.19
Irlande	16.16	0.06	8.31
Italie	0.70	0.21	3.24
Japon	38.77	2.11	5.65
Luxembourg	0.21	23.28	0.35
Norvège	0.03	12.47	0.85
Pays-Bas	2.77	9.45	6.51
Portugal	6.91	0.44	0.02
Royaume-Unis	3.54	1.77	6.72
Suède	1.60	0.36	0.74

TABLE 10 – Contributions des individus sur les trois premiers axes

	Dim.1	Dim.2	Dim.3
i1998	0.72	-0.41	0.54
i1999	0.87	-0.25	-0.07
i2000	0.84	0.19	-0.32
i2001	0.86	-0.18	-0.26
i2002	0.92	-0.03	-0.09
i2003	0.95	0.11	0.05
i2004	0.64	0.68	0.31

TABLE 11 – Coordonnées des variables sur les trois premiers axes

	Dim.1	Dim.2	Dim.3
i1998	0.52	0.17	0.29
i1999	0.76	0.06	0.00
i2000	0.71	0.04	0.10
i2001	0.74	0.03	0.07
i2002	0.84	0.00	0.01
i2003	0.90	0.01	0.00
i2004	0.42	0.46	0.10

TABLE 12 – Cosinus^2 des variables sur les trois premiers axes

	Dim.1	Dim.2	Dim.3
i1998	10.56	21.89	50.59
i1999	15.65	8.08	0.79
i2000	14.45	4.56	18.30
i2001	15.23	4.13	11.69
i2002	17.28	0.15	1.40
i2003	18.33	1.56	0.49
i2004	8.50	59.64	16.74

TABLE 13 – Contributions des variables sur les trois premiers axes

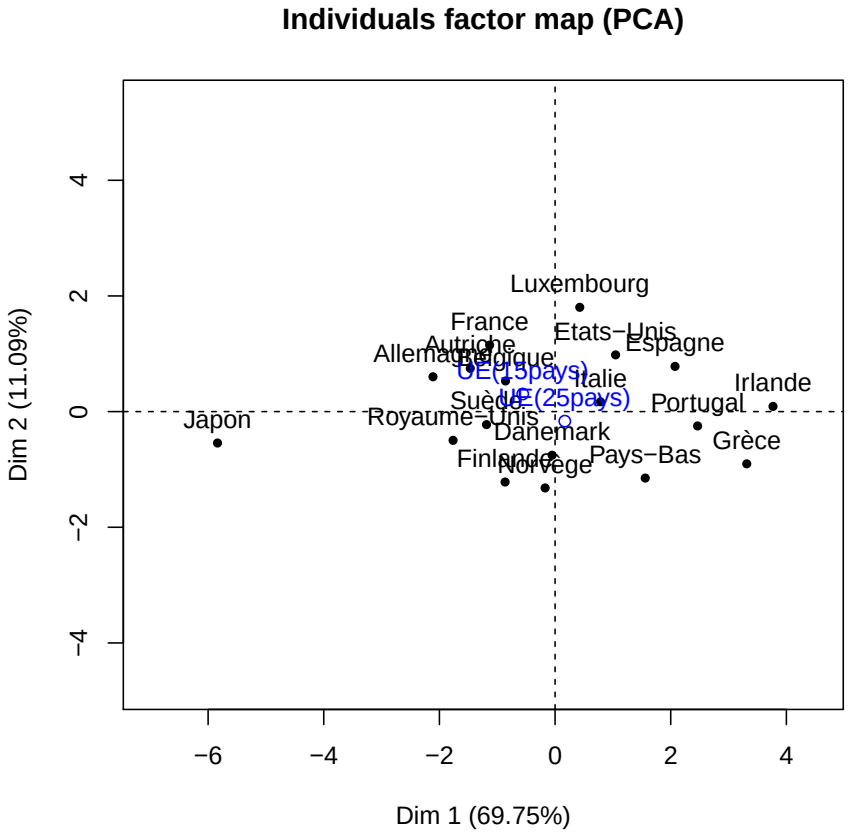

FIGURE 7 – Représentation des individus sur le premier plan factoriel

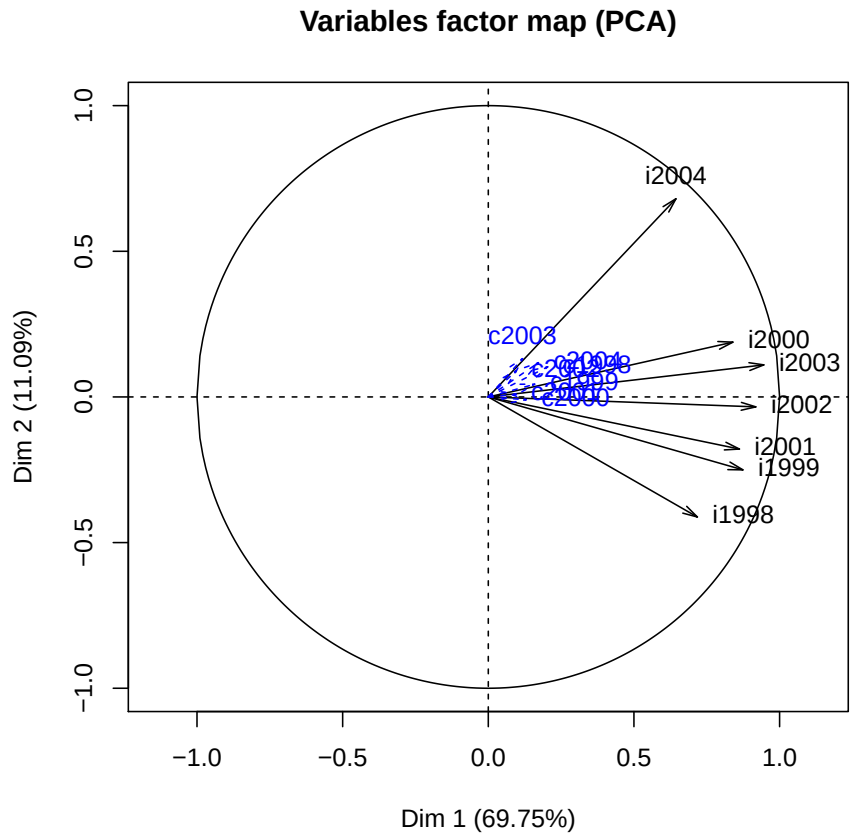

FIGURE 8 – Carte des variables