Semaine d'Etude Mathématiques et Entreprises 2 : Analyse multivariées pour la production d'aluminium

Thomas Auphan, Pierre Bochard, Juliette Bouhours, Blanche Buet, Julien Claisse, Anaïs Crestetto, Yannick Deleuze

Semaine d'Etude Maths–Entreprises 2

28 novembre – 2 décembre 2011, Université Lyon I

Analyses multivariées pour la production d'aluminium

T. Auphan[a] P. Bochard[b]
J. Bouhours[c] B. Buet[d]
J. Claisse[e] A. Crestetto[f]
Y. Deleuze[g,h]

[a] *Laboratoire d'Analyse Topologie Probabilités, Aix Marseille Université, 13453 Marseille, France*
[b] *Laboratoire de Mathématiques d'Orsay, Université Paris 11, 91405 Orsay, France*
[c] *Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 75005 Paris, France*
[d] *Institut Camille Jordan, Université Lyon 1, 69622 Villeurbanne, France*
[e] *INRIA Sophia Antipolis et Université de Nice Sophia Antipolis, 06902 Sophia Antipolis, France*
[f] *INRIA Nancy - Grand Est et Institut de Recherche Mathématique Avancée, Université de Strasbourg, 67084 Strasbourg, France*
[g] *Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 75005 Paris, France*
[h] *Scientific Computing and Cardiovascular Simulation Laboratory, National Taiwan University, Taiwan*

Sujet proposé par :

Rio Tinto Alcan

Correspondant : A. Augé (Rio Tinto Alcan)

Maths & Entreprises — cnrs dépasser les frontières — a m i e s

Résumé

Ce rapport présente l'étude statistique, menée au cours de la deuxième Semaine d'Étude Maths–Entreprises, d'un problème industriel rencontré par Rio Tinto Alcan [3]. Productrice d'aluminium par électrolyse, cette entreprise cherche à expliquer des fluctuations de procédé. À partir d'un ensemble de mesures sur les anodes et sur les cuves à électrolyse, nous proposons d'utiliser des méthodes d'analyse multivariée pour construire des modèles explicatifs. Le but étant de permettre aux usines d'éviter les périodes avec des fluctuations. Dans une première section, nous présentons le problème et ses enjeux. Nous détaillons dans les sections suivantes les différentes méthodes explorées et les résultats obtenus : l'analyse du coefficient de corrélation en présence d'un déphasage et l'auto-corrélation (Section 2), l'analyse en composantes principales (Section 3), les arbres de décisions (Section 4), le clustering et la régression linéaire (Section 5). Des résultats complémentaires sont donnés en annexe.

Table des matières

1 Présentation du modèle et enjeux du projet

1.1 Présentation du modèle

Rio Tinto Alcan (RTA) [3] est une filiale du groupe minier Anglo–Australien Rio Tinto. RTA produit et commercialise de la bauxite (minerai d'aluminium), de l'alumine (oxyde d'aluminium Al_2O_3) et de l'aluminium. L'aluminium est produit dans des cuves par électrolyse de l'alumine à 950^oC dans un bain de fluorures fondus. Le métal se dépose à la cathode (réduction des ions Al_3^+) et il y a un dégagement de gaz carbonique CO_2 à l'anode en carbone (oxydation des ions O_2^-).

L'anode est constituée de plusieurs blocs de carbone qui sont remplacés régulièrement (environ toutes les 4 semaines) quand la quantité de carbone restante n'est plus suffisante pour l'électrolyse. Environ un bloc de carbone est remplacé par jour et par cuve.

Il est observé parfois des fluctuations dans la consommation des anodes. Ces fluctuations apparaissent rarement de manière isolée. Les graphiques ci-dessous montrent l'évolution chronologique des fluctuations observées dans une usine (un point hebdomadaire en abscisse, l'unité de l'axe des ordonnées est centrée réduite). Nous représentons cette évolution pendant environ 200 semaines sur la figure 1 et nous restreignons à une année sur la figure 2.

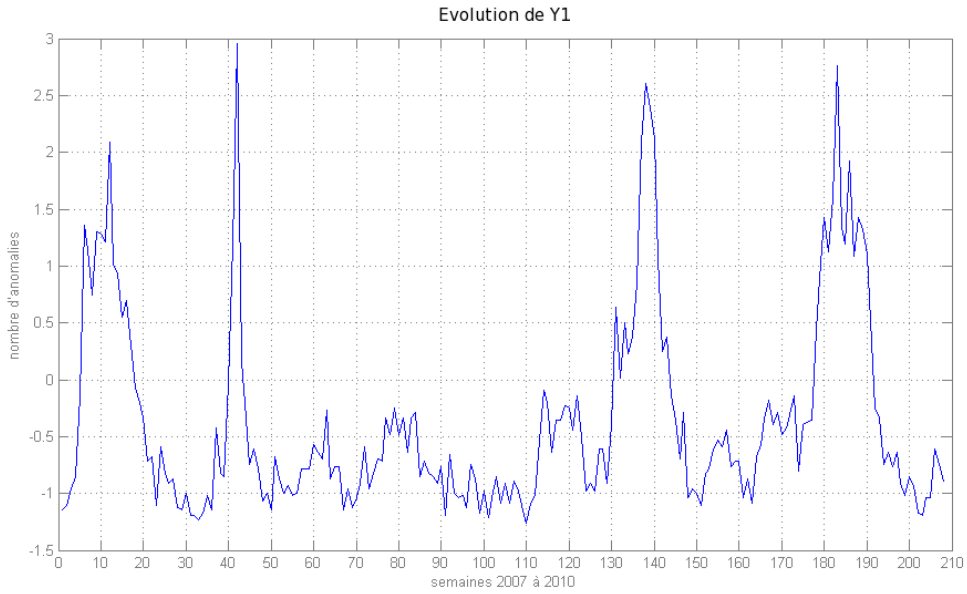


FIGURE 1 – Fluctuations en fonction du temps sur les années 2007, 2008, 2009 et 2010.

Un suivi rigoureux du procédé est fait et de nombreuses variables sont enregistrées.

L'objectif de cette étude est de mettre en évidence les variables ou les combinaisons de variables expliquant les fluctuations observées afin de pouvoir d'une part anticiper leur apparition et d'autre part les faire disparaître le cas échéant.

Dans la suite, les fluctuations sont suivies avec la variable Y_1 qui est la variable à expliquer.

1.2 Problématiques

André Augé, notre correspondant de Rio Tinto Alcan, a donc mis à notre disposition deux jeux de données :

- le premier avec la variable à expliquer Y_1 et un jeu de 73 variables mesurées sur 43 semaines (grandeurs, comptage, durée...),

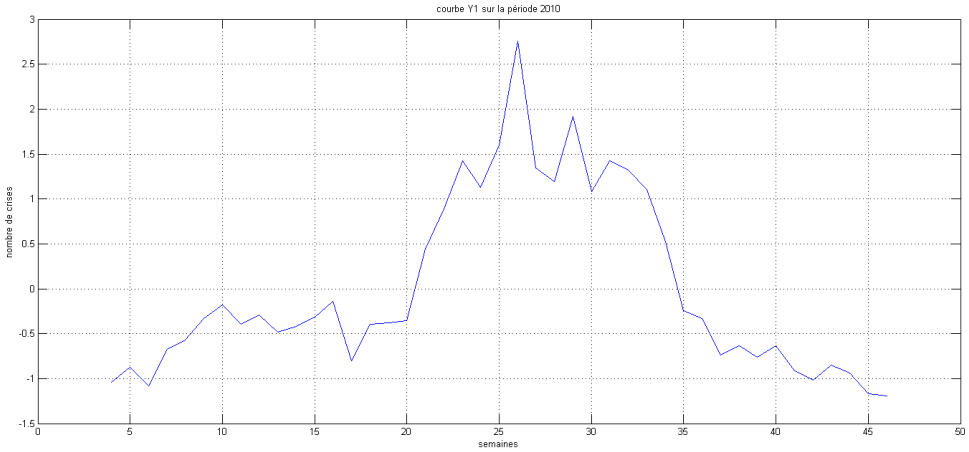


FIGURE 2 – Fluctuations en fonction du temps sur l'année 2010.

– le deuxième avec toujours cette même variable à expliquer Y_1 et un jeu de 50 variables mesurées sur quatre années consécutives (2007, 2008, 2009 et 2010).

Dans un souci de confidentialité, ces variables ont été renommées (Y_1, Y_2, X_{101}, $X_{2\star\star}$, et $X_{3\star\star}$) et sont centrées réduites.

Quelques remarques :

– Il manquait certaines valeurs pour les variables X_{201}, X_{202} et X_{205}. N'étant pas habitués au traitement de données manquantes, et dans un souci de rapidité (en accord avec le temps qui nous était alloué), nous avons choisi d'exclure ces variables. Il convient donc de préciser que ces variables mises de côté peuvent avoir un rôle important dans ces fluctuations.
– André Augé nous a indiqué que les variables X_{301}, X_{302} et X_{336} n'étaient pas intéressantes. Nous les avons par conséquent retirées pour notre étude.
– Les études menées auparavant par André Augé avaient fait ressortir cinq variables significatives (X_{203}, X_{210}, X_{218}, X_{223} et X_{315}) que nous avons privilégiées dans certaines expériences.

Un des problèmes pour notre étude est le faible nombre de données disponibles, qui fait que l'on est souvent à la limite du domaine de validité des méthodes utilisées (mais c'est souvent le cas dans les études industrielles).

1.3 Méthodes utilisées

Pour traiter ce problème, nous avons principalement eu recours aux cinq méthodes suivantes, que nous décrirons plus en détails dans les prochaines sections :

– l'analyse du coefficient de corrélation en présence d'un déphasage et auto-corrélation sur les deux jeux de données (Section 2),
– l'analyse en composantes principales (ACP) sur les deux jeux de données (Section 3),
– les arbres de décision sur les deux jeux de données (Section 4),
– la régression linéaire et le clustering sur le premier jeu de données (Section 5).

2 Méthode de corrélation-déphasage

Nous présentons tout d'abord les méthodes de corrélation-déphasage et d'auto-régression.

2.1 Définition de la méthode

Le principe de cette méthode est d'introduire un déphasage en temps sur chacune des variables considérées et d'étudier l'évolution du coefficient de corrélation de la variable déphasée avec la variable à expliquer. Pour cela on considère les variables une par une et pour chaque variable on décale les données de n semaines, soit vers la gauche pour déterminer les variables conséquences, soit vers la droite pour déterminer les variables causes. On regarde ensuite pour quelles valeurs le coefficient de corrélation est le plus élevé (en valeur absolue), si le déphasage "optimal" est négatif cela signifie que la variable considérée pourrait être cause, si le déphasage "optimal" est positif alors la variable serait conséquence.

Il faut néanmoins être vigilant quant au choix des valeurs maximales et minimales du déphasage. En effet pour pouvoir calculer le coefficient de corrélation entre la variable déphasée et la variable à expliquer, il faut qu'elles aient le même nombre d'observations. Ainsi si on déphase une variable de n semaines on perd n observations (les n premières ou dernières en fonction du signe du déphasage). Donc par exemple pour l'étude faite avec le premier jeu de données, si le déphasage est supérieur à 5 semaines les résultats obtenus ne sont plus intéressants car la durée de séjour d'un bloc de carbone (anode) sur cuve est d'environ 4 semaines.

On a alors appliqué cette méthode sur plusieurs variables :

- dans un premier temps, nous avons utilisé cette méthode sur les cinq variables significatives données par André Augé : X_{203}, X_{210}, X_{218}, X_{223} et X_{315},
- puis on a effectué le même raisonnement sur le deuxième jeu de données (étendu aux quatre années), sur chacune des variables explicatives cette fois-ci.

On obtient alors des conclusions variées que nous présentons maintenant.

2.2 Résultats sur le premier jeu de données

Pour ce premier jeu, nous rappelons que nous avons étudié seulement les cinq variables significatives données par André Augé. Vous pourrez trouver toutes les courbes correspondantes en Annexe A. Les résultats obtenus pour la variable X_{315} sont présentés sur la Figure 3. Nous observons que cette variable aurait tendance a être une variable conséquence (d'environ une ou deux semaines). Nous avons ensuite, par souci de clarté, réuni les coefficients de corrélation de chaque variable sur un même graphique (Figure 4).

Nous observons alors que les variables X_{203} et X_{315} ne peuvent être que des variables conséquences, alors que les variables X_{218} et X_{223} ne peuvent être que des variables causes ou conséquences immédiates. On ne peut pas conclure pour la variable X_{210}.
Attention, cette analyse ne permet pas de conclure à un lien direct cause/conséquence entre Y_1 et $X_{\star\star\star}$ (les deux variables peuvent être liées à une autre variable extérieure).

2.3 Résultats sur le deuxième jeu de données

Pour ce jeu de données, on a repéré plusieurs variables qui semblent significatives : X_{304} et X_{345}, d'autres relativement significatives, à surveiller : X_{308}, X_{309}, X_{312}, X_{319}, X_{321}, X_{329}, X_{330}, X_{333}, X_{340}, X_{344} et X_{346}. Ces dernières variables ne se comportent pas toujours de la même manière mais semblent bien corrélées à deux ou trois reprises.

On va présenter l'évolution du coefficient de corrélation entre la variable $X_{\star\star\star}$ et la variable à expliquer Y_1 pour trois variables types : une variable qui apparaît nettement significative, une variable qui apparaît parfois significative et une variable qui ne paraît pas significative.

Figure 5, nous représentons le coefficient de corrélation entre les variables Y_1 et X_{304} en fonction du déphasage, ainsi que leur allure en fonction du temps. Cette variable apparaît comme une cause ou conséquence immédiate de la variable à expliquer. Son coefficient de corrélation augmente très fortement quand le déphasage est aux alentours de 0.

Figure 6, nous représentons le coefficient de corrélation entre les variables Y_1 et X_{340} en fonction du déphasage, ainsi que leur allure en fonction du temps. Cette variable apparaît comme une cause. Mais les résultats sont quand même moins flagrants que pour la variable précédente X_{304}.

Figure 7, nous représentons le coefficient de corrélation entre les variables Y_1 et X_{315} en fonction du déphasage, ainsi que leur allure en fonction du temps. Cette variable est très

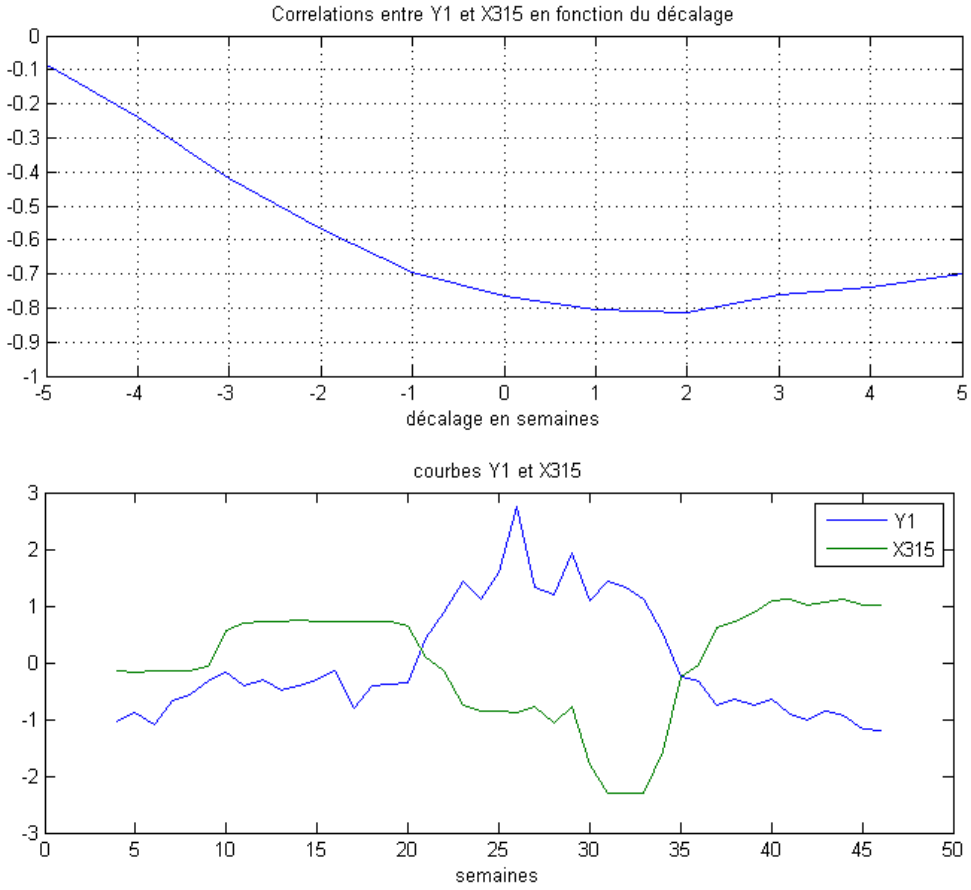


FIGURE 3 – Corrélation entre Y_1 et X_{315} en fonction du déphasage (haut) et tracés de ces variables en fonction du temps (bas).

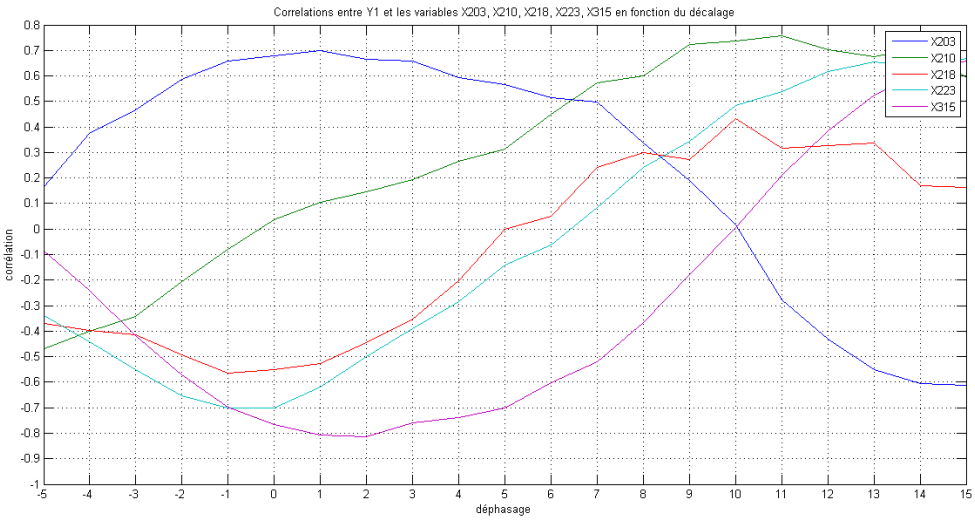


FIGURE 4 – Coefficients de corrélation en fonction du déphasage, variables X_{203} (bleu foncé), X_{210} (vert), X_{218} (rouge), X_{223} (bleu clair) et X_{315} (violet) comparées à Y_1.

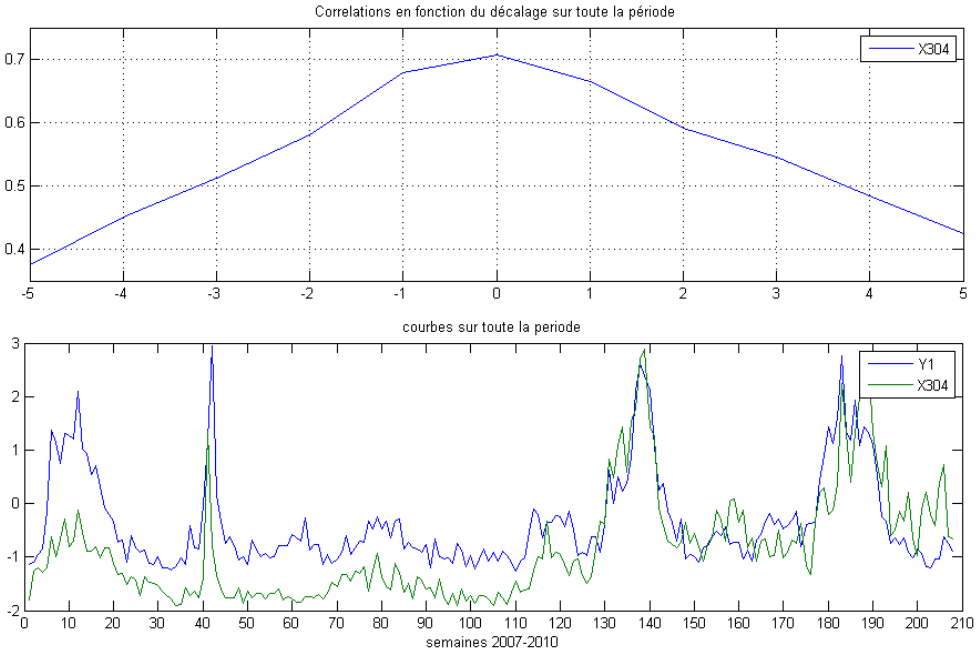


FIGURE 5 – Corrélation entre Y_1 et X_{304} en fonction du déphasage (haut) et tracés de ces variables en fonction du temps (bas).

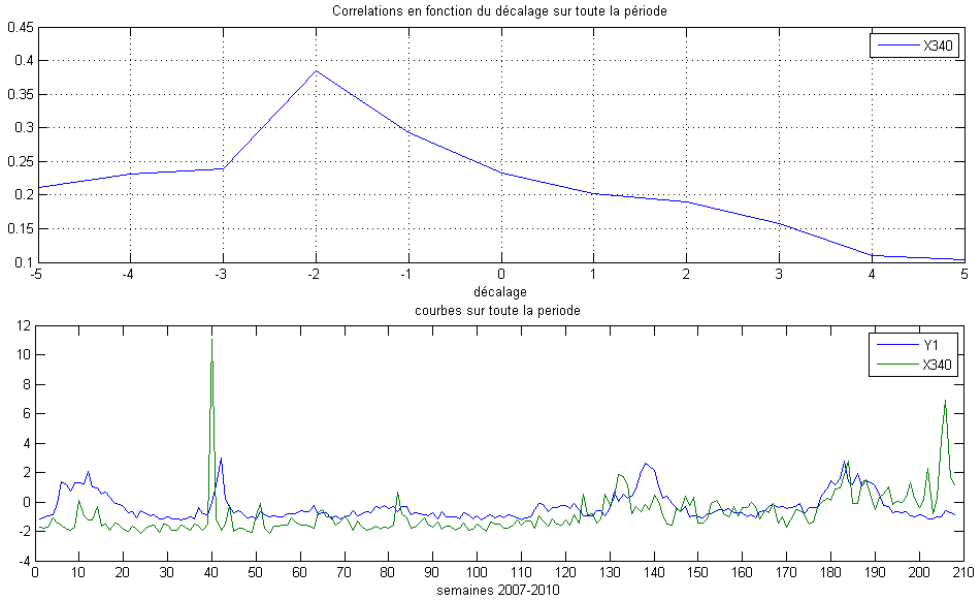


FIGURE 6 – Corrélation entre Y_1 et X_{340} en fonction du déphasage (haut) et tracés de ces variables en fonction du temps (bas).

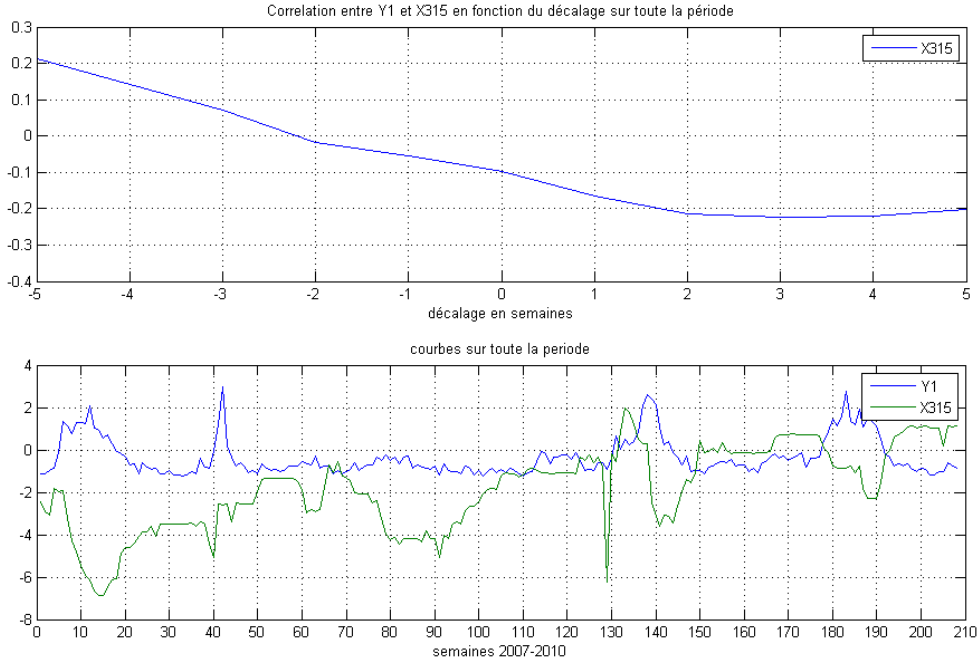


FIGURE 7 – Corrélation entre Y_1 et X_{315} en fonction du déphasage (haut) et tracés de ces variables en fonction du temps (bas).

intéressante et illustre bien les limites de notre méthode. En effet, considérée sur l'année 2010 seulement cette variable nous apparaissait comme significative alors que lorsqu'on l'étudie sur une période étendue cette variable ne semble plus du tout significative.

2.4 Méthode d'auto-corrélation pour construire un indicateur

Cette méthode requiert de connaître une ou de préférence plusieurs courbes significatives. L'objectif est d'éliminer ou simplifier l'étape de la lecture de la courbe significative par un œil humain. En effet lorsque l'on dispose d'une ou plusieurs courbes significatives, il faut pouvoir les lire et les analyser pour déterminer si on entre ou non dans une période de fluctuations. Pour cela, on essaie de débruiter la courbe pour ne garder que les variations principales puisqu'on détecte une période de fluctuations précisément lorsque les courbes significatives présentent des variations élevées pendant quelques semaines. On recherche donc un indicateur de ces variations. On a pour cela testé une méthode dite d'auto-corrélation.

On se donne une variable X aux temps n allant de 1 à N. L'idée est de trouver des coefficients $a_1 \dots a_p$ vérifiant

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + \varepsilon_n.$$

Plus précisément, on choisit un entier dn (supérieur à l'ordre p de l'autorégression) et on calcule pour chaque temps n entre 1 et $N - dn$ les coefficients $a_1 \dots a_p$ qui minimisent

$$\sum_{i=n+p}^{n+dn} \varepsilon_i^2 = \sum_{i=n+p}^{n+dn} \left(X_i - \sum_{k=1}^{p} a_k X_{i-k} \right)^2.$$

On peut alors calculer les *coefficients de réflexion* associés à ce modèle d'auto-corrélation : si on représente les coefficients trouvés pour n entre 1 et $N - dn$ pour la variable X_{304} avec un ordre $p = 2$, on obtient la courbe $k2$ (en vert) sur la Figure 8.

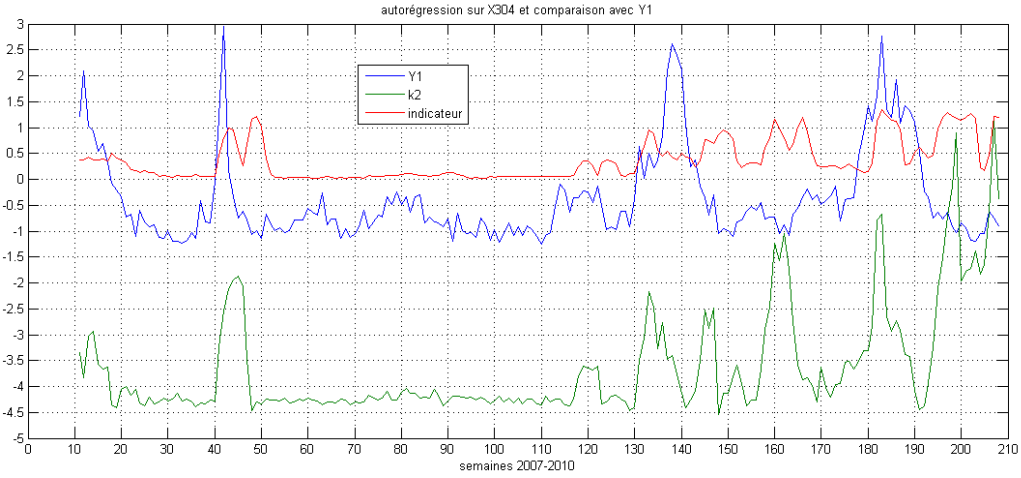


FIGURE 8 – Comparaison entre la courbe de Y_1 (en bleu) et deux indicateurs.

On voit assez nettement apparaître des pics sur les courbes rouge et verte. L'idée était alors simple, lorsque les indicateurs liés à plusieurs courbes significatives dépassent un certain seuil, Y_1 devrait croître. Le problème ici, c'est que les pics observés sur les indicateurs sont en retard par rapport aux pics de Y_1. Cela provient du fait que pour calculer les coefficients, on a besoin des données sur dn temps différents, d'où le retard au niveau de la courbe verte ($k2$ coefficient de réflexion), et la courbe rouge *indicateur* est obtenue en moyennant la courbe de $k2$ d'où un retard encore plus marqué. Sachant qu'il s'agit de semaines, cette méthode n'est pas efficace

pour anticiper une variation de la réponse Y_1. Peut-être qu'avec une méthode basée sur des transformées en ondelettes on aurait obtenu des résultats plus précis en temps.

3 Analyse en Composante Principale (ACP)

La deuxième méthode étudiée est l'analyse en composante principale (ACP) [1]. Nous la présentons dans cette section.

3.1 Objectifs de l'ACP

Considérons X_1, X_2, ...,X_n n variables aléatoires dont on connaît p réalisations conjointes. Géométriquement, on peut les visualiser comme un nuage de p points de $\mathbb{R}^n$. Le but de l'analyse en composante principale est de trouver une base adaptée à la représentation du nuage de points. Plus précisément, on va chercher des vecteurs de $\mathbb{R}^n$ tels que la projection du nuage de points sur ces vecteurs maximise la variance, ce sont ces vecteurs qu'on appelle composantes principales. Ceci permet deux choses :

- l'extraction d'informations signifiantes : les composantes principales donnent des directions qui expliquent bien la forme du nuage de points. Pour nos données, cela signifie qu'il peut être plus intéressant de regarder l'évolution des composantes principales au cours du temps plutôt que l'évolution de chacune des variables initiales séparément.
- la hiérarchisation des informations : chacune des composantes principales est associée à un poids représentant combien la composante explique la forme du nuage de points. En pratique, cela signifie qu'un grand nombre de variables aléatoires va pouvoir être remplacé par un nombre beaucoup plus faible.

Bien sûr, la méthode avait déjà été utilisée par André Augé. N'étant pas spécialistes en statistiques, il nous a néanmoins semblé intéressant de la reprendre par nous-mêmes comme point de départ. Cette méthode nous a paru pertinente à plusieurs égards pour répondre à la question posée par Rio Tinto Alcan. D'une part, on disposait d'une masse importante de données (75 variables aléatoires différentes à 43 instants distincts) dont on ne connaissait pas la signification de sorte qu'il n'était pas possible de faire une sélection empirique des données qui pourraient être pertinentes ou non pour expliquer Y_1. L'ACP nous a permis de garder des données qui semblaient pertinentes, et d'oublier les autres. On a appliqué par la suite d'autres méthodes statistiques à ces données selectionnées. La projection sur les premières composantes principales nous a également permis de séparer les variables suivant des périodes de temps significatives. Dans ce qui suit, on rappelle brièvement le principe mathématique sur lequel se base l'ACP, puis on présente nos applications de cette méthode aux données de Rio Tinto Alcan.

3.2 Description mathématique

On va représenter les p réalisations de nos n variables aléatoires par une matrice M (on supposera par la suite les variables centrées et réduites) :

$$M = \begin{pmatrix} X_{1,1} & X_{2,1} & \dots & X_{n,1} \\ X_{1,2} & X_{2,2} & \dots & X_{n,2} \\ \vdots & & & \vdots \\ X_{1,p} & X_{2,p} & \dots & X_{n,p} \end{pmatrix}.$$

On cherche maintenant un vecteur u de $\mathbb{R}^n$ tel que la projection sur u soit de variance maximale. Or la projection sur u est donnée par $\pi_u(M) = M \cdot u$ dont la variance vaut :

$$\frac{1}{p} \cdot \pi_u(M)^t \cdot \pi_u(M) = \frac{1}{p} \cdot u^t \cdot M^t \cdot M \cdot u.$$

La matrice $M^t \cdot M$ est symétrique définie positive, donc diagonalisable dans une base orthonormée et il existe une matrice diagonale $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ et une matrice orthogonale P

dans $\mathcal{M}_n(\mathbb{R})$ telles que :

$$\frac{1}{p} \cdot \pi_u(M)^t \cdot \pi_u(M) = (Pu)^t \cdot D \cdot Pu.$$

Notant $Pu = v$, ceci se réécrit :

$$\frac{1}{p} \cdot \pi_u(M)^t \cdot \pi_u(M) = \sum_{i=1}^{n} v_i^2 \lambda_i.$$

D étant définie positive, les λ_i sont positifs ; si on les suppose ordonnés par ordre décroissant, il est clair que les vecteurs unitaires v maximisant la quantité précédente sont les vecteurs propres associés à D et les composantes principales u sont déterminées par la formule $u = P^{-1}v$. Le poids associé à u est alors la valeur propre λ_i associée à v.

En pratique, on se fixe d'avance un seuil explicatif (de l'ordre de 70%) et on ne garde que les k premières composantes principales telles que $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \geq \frac{70}{100}$.

3.3 Applications

3.3.1 Détermination de périodes

Notre idée n'est pas de voir les semaines comme une évolution continue dans le temps mais comme appartenant à une classe correspondant à des périodes d'avant, pendant et après une période avec des fluctuations.

Nous avons subdivisé les 43 semaines en un certain nombre de périodes que nous avons projetées en effectuant une ACP. En observant les résultats obtenus nous avons déterminé quatre périodes qui nous semblent les plus importantes et que nous utiliserons pour observer les résultats des différentes ACP effectuées ainsi que pour utiliser d'autre outils statistiques. Ces périodes sont les suivantes :

- avant : semaines 4 à 19,
- pré : semaines 20 à 22,
- pendant : semaines 23 à 33, et
- après : semaines 34 à 46.

3.3.2 Méthodes

On utilise le package `prcomp` du logiciel `R` afin d'effectuer l'ACP du jeu de données. Le jeu de données contient 73 variables avec 43 observations. Ces données sont stockées dans une matrice. Puis les trois variables dont il manque des mesures sont tronquées ainsi que les trois variables dont il nous a été conseillé de ne pas tenir compte. Nous avons donc 67 variables et 43 observations correspondant à 43 semaines sur lesquelles ont été prises les mesures.

Pour chaque expérience :

- on distingue des périodes que l'on labellise par des couleurs : avant (semaines 4 à 19), pré (20 à 22), pendant (23 à 33) et après (34 à 46),
- on choisit q composantes principales représentant 75% de l'inertie,
- on trace les nuages de points dans les plans factoriels associés aux q composantes principales,
- on cherche à interpréter et déterminer les variables significatives.

3.3.3 Expérience 1

Nous réalisons tout d'abord une ACP sur toutes les variables ainsi que sur toutes les 43 observations labellisées en 4 périodes.

On retient les 4 premières composantes principales qui représentent les 75% de l'inertie totale, mais on ne s'intéresse qu'à PC1 et PC2 qui semblent contenir des informations exploitables (voir l'Annexe B pour les graphiques complémentaires).

En projetant sur les deux premières composantes principales (Figure 9), on distingue 3 nuages de points qui se démarquent nettement représentant bien les périodes d'avant, de pendant et d'après. PC1 semble donner une information chronologique. Sur la figure 9, on

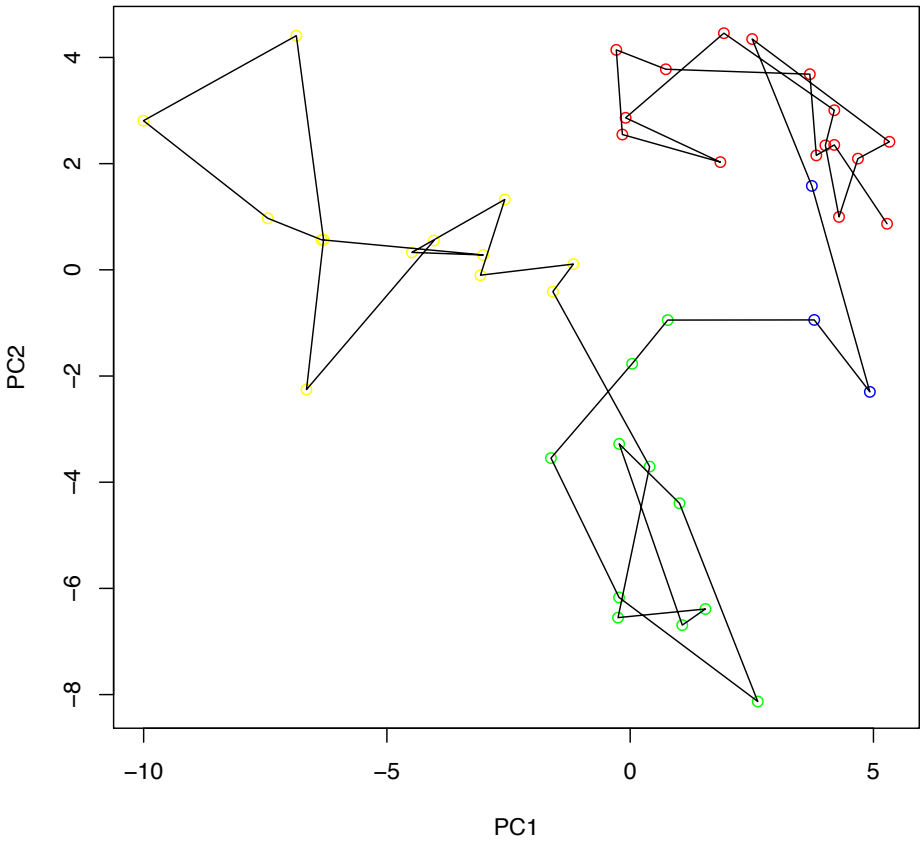


FIGURE 9 – Expérience 1 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps. L'axe PC1 semble donner une information chronologique alors que l'axe PC2 semble distinguer deux périodes.

peut voir de droite à gauche, les semaines d'avant, puis de pendant et enfin d'après. PC2 nous semble plus importante dans le sens où elle permet de discriminer les semaines de fluctuation des autres semaines.

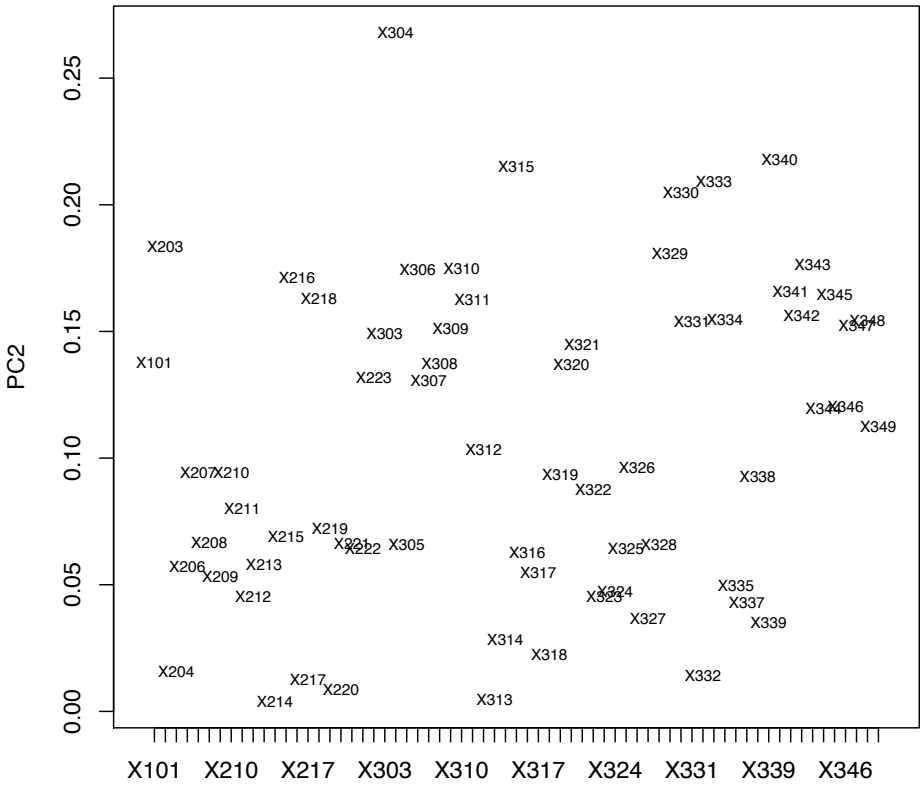


FIGURE 10 – Expérience 1 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2. La variable X_{304} est celle qui participe le plus à la formation de l'axe PC2.

Pour chaque axe retenu, on regarde quelles sont les variables qui participent le plus à la formation de l'axe, c'est-à-dire celles qui ont une grande coordonnée en valeur absolue sur l'axe. Dans le cas de PC2 (Figure 10), ce sont les variables X_{304}, X_{340}, X_{315}, X_{333}, X_{330}, X_{203}, X_{329} et X_{343} qui sont les plus importantes sur l'axe. Dans le cas de PC1, on peut retenir les variables X_{217}, X_{212}, X_{215} et X_{210}.

Grâce à cette première ACP, nous avons pu mettre en évidence la période avec des fluctuations et la caractériser par les variables qui semblent les plus significatives, mais on ne peut pas identifier les facteurs explicatifs.

3.3.4 Expérience 2

À partir de la première expérience, nous avons sélectionné 12 variables aléatoires qui semblaient participer le plus à la formation des deux axes principaux : X_{304}, X_{340}, X_{315}, X_{333}, X_{330}, X_{203}, X_{329} et X_{343}, X_{217}, X_{212}, X_{215} et X_{210}. Nous réalisons une ACP avec ces 12 variables significatives uniquement et sur les 43 semaines.

On retient les deux premières composantes principales pour projeter les individus représentant les semaines (Figure 11). On observe que la composante principale PC1 semble traduire, non seulement une augmentation des fluctuations, mais aussi une irréversibilité dans le système.

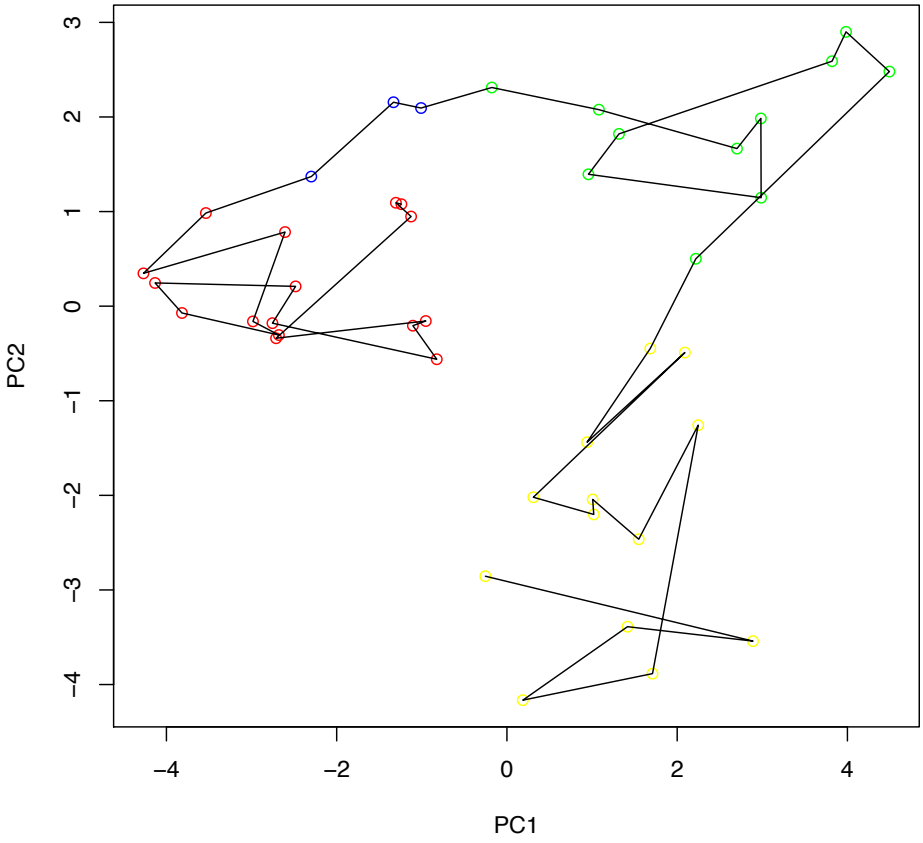


FIGURE 11 – Expérience 2 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

En se focalisant sur PC1 (Figure 12), on peut donc mettre en évidence les variables qui participent le plus, en valeur absolue, à la formation de l'axe. Les variables X_{329}, X_{330}, X_{340}, X_{333}, X_{304}, X_{210} et X_{215} auraient peut-être une influence sur Y_1.

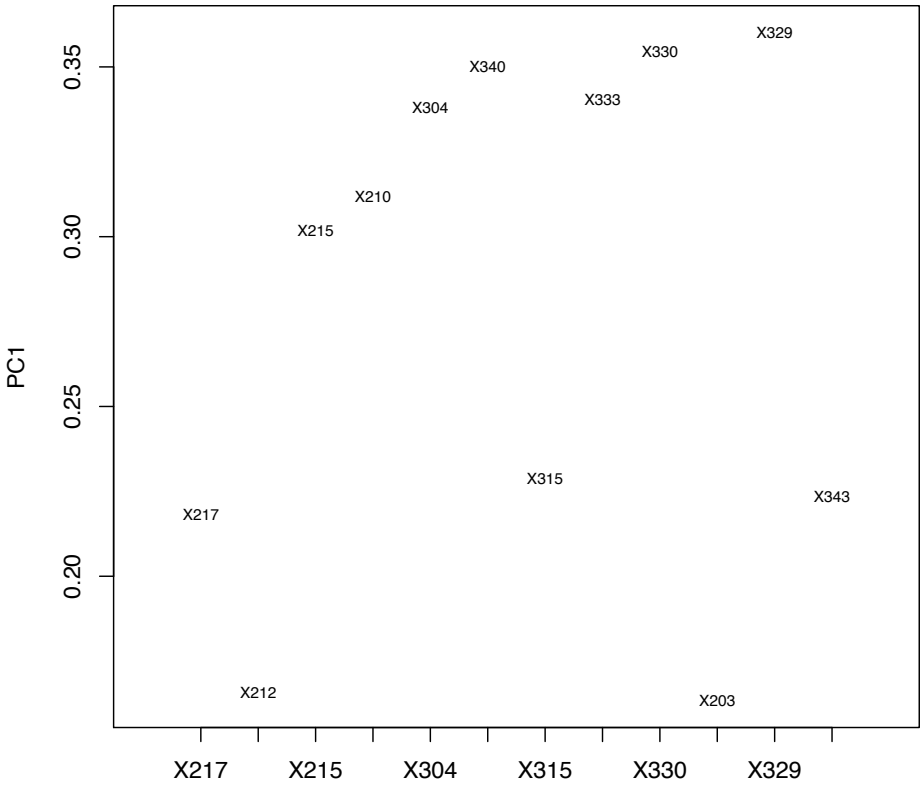


FIGURE 12 – Expérience 2 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1. La variable X_{329} est celle qui participe le plus à la formation de l'axe PC1.

3.3.5 Expérience 3

Après avoir mis en évidence des variables susceptibles d'être explicatives, nous nous sommes restreints aux 19 premières semaines (qui correspondent aux semaines avant) en considérant toutes les variables.

On retient les 4 premières composantes principales pour expliquer l'inertie totale du système, mais on se focalise sur la projection dans l'espace factoriel des trois axes PC2, PC3 et PC4 (Figure 13). L'axe PC3 permet d'expliquer au mieux la variabilité entre les nuages de points d'avant et de pré.

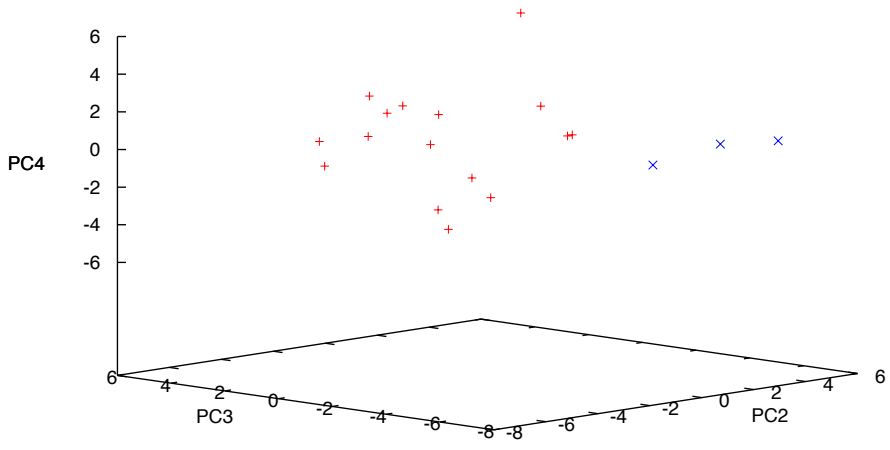


FIGURE 13 – Expérience 3 : Projection des individus sur l'espace factoriel PC2× PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. L'axe PC3 permet de différencier au mieux les périodes d'avant et de pré.

En regardant quelles variables ont le plus de poids (en valeur absolue) dans la formation de l'axe PC3 (Figure 14), on peut retenir les 12 variables suivantes : X_{348}, X_{333}, X_{334}, X_{343}, X_{204}, X_{322}, X_{215}, X_{212}, X_{216}, X_{308}, X_{338}, X_{217}. Celles-ci seraient donc les variables qui différencieraient au mieux le nuage de points d'avant de celui de pré.

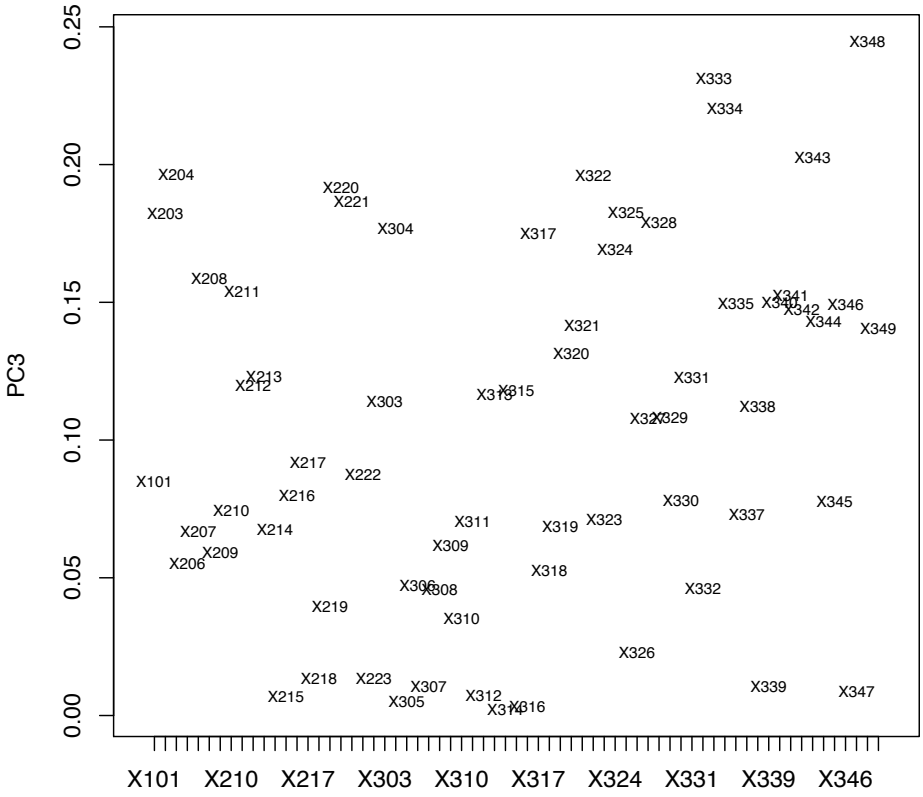


FIGURE 14 – Expérience 3 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3. La variable X_{348} est celle qui participe le plus à la formation de l'axe PC3.

3.3.6 Expérience 4

Dans l'expérience 3, nous avons mis en évidence 12 variables qui semblent jouer un rôle dans le passage de l'avant à celui de pré. Nous effectuons une ACP sur ces 12 variables en tenant compte des 19 premières semaines comme précédemment.

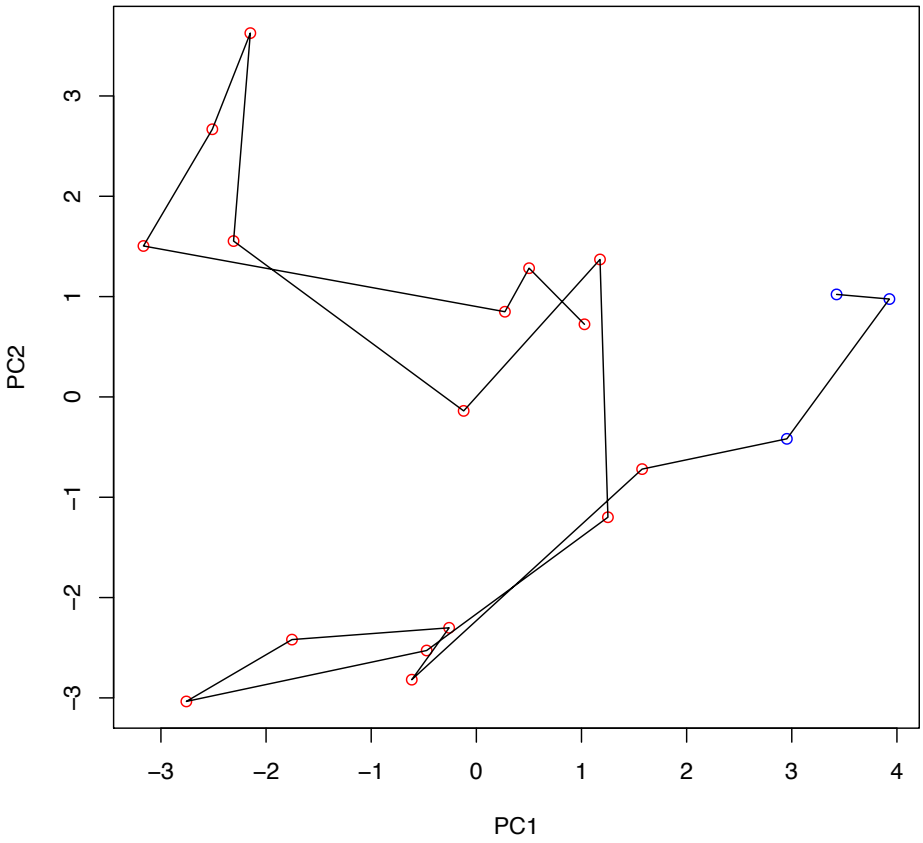


FIGURE 15 – Expérience 4 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

En projetant sur le plan factoriel PC1 × PC2 (Figure 15), les deux nuages de points d'avant et de pré sont bien différenciés. L'axe PC1 permet d'expliquer au mieux la variabilité entre les nuages de points d'avant et de pré.

En regardant quelles variables ont le plus de poids (en valeur absolue) dans la formation de l'axe PC1 (Figure 16), on peut retenir les 3 variables suivantes : X_{212}, X_{216} et X_{217}. Celles-ci seraient, parmi les 12 variables pré-selectionnées, les variables qui expliqueraient au mieux le passage d'avant à pré.

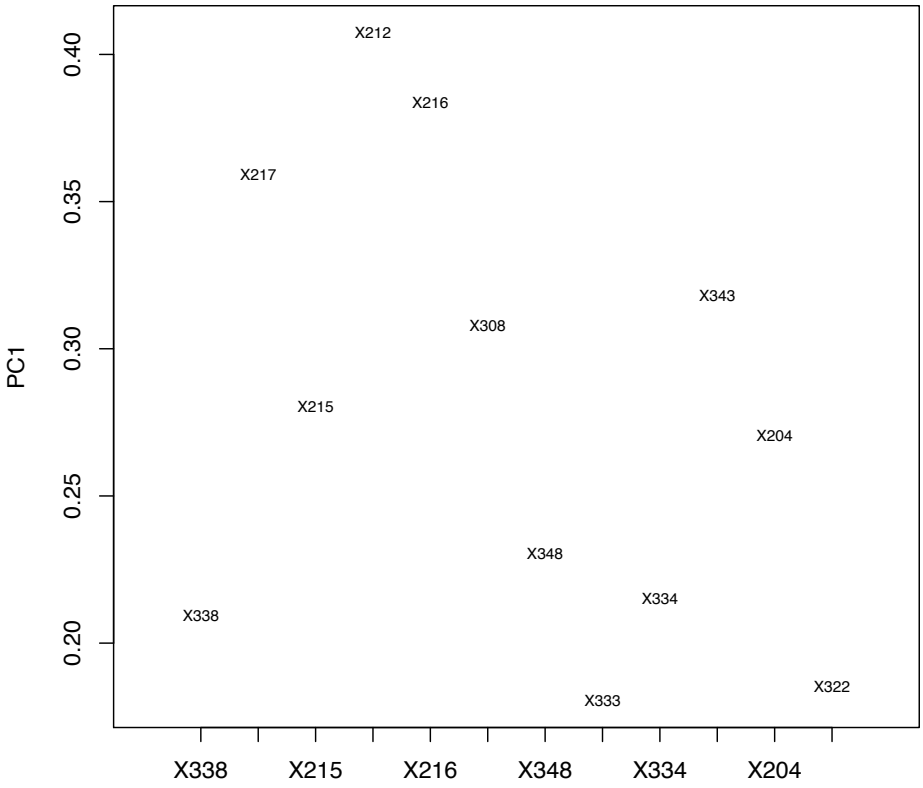


FIGURE 16 – Expérience 4 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1. La variable X_{212} est celle qui participe le plus à la formation de l'axe PC1.

3.3.7 Expérience 5

Cette fois nous choisissons de ne pas considérer la période d'après et de prendre en compte les 33 premières semaines seulement. Nous réalisons une ACP avec toutes les variables.

En projetant les observations sur le plan factoriel PC1×PC2 (Figure 17), on peut voir que PC1 nous donne des informations sur la période de fluctuations. En projetant les variables sur l'axe factoriel PC1 (Figure 18), nous pouvons mesurer le poids de chacune des variables initiales. On observe que les variables X_{304}, X_{315}, X_{333}, X_{203}, X_{210} et X_{340} semblent jouer un rôle pour discriminer la période de fluctuation. On retrouve certaines variables obtenues précédemment. Ce résultat est cohérent avec les résultats obtenus dans les expériences 1 et 2 mais n'apporte pas beaucoup d'informations supplémentaires.

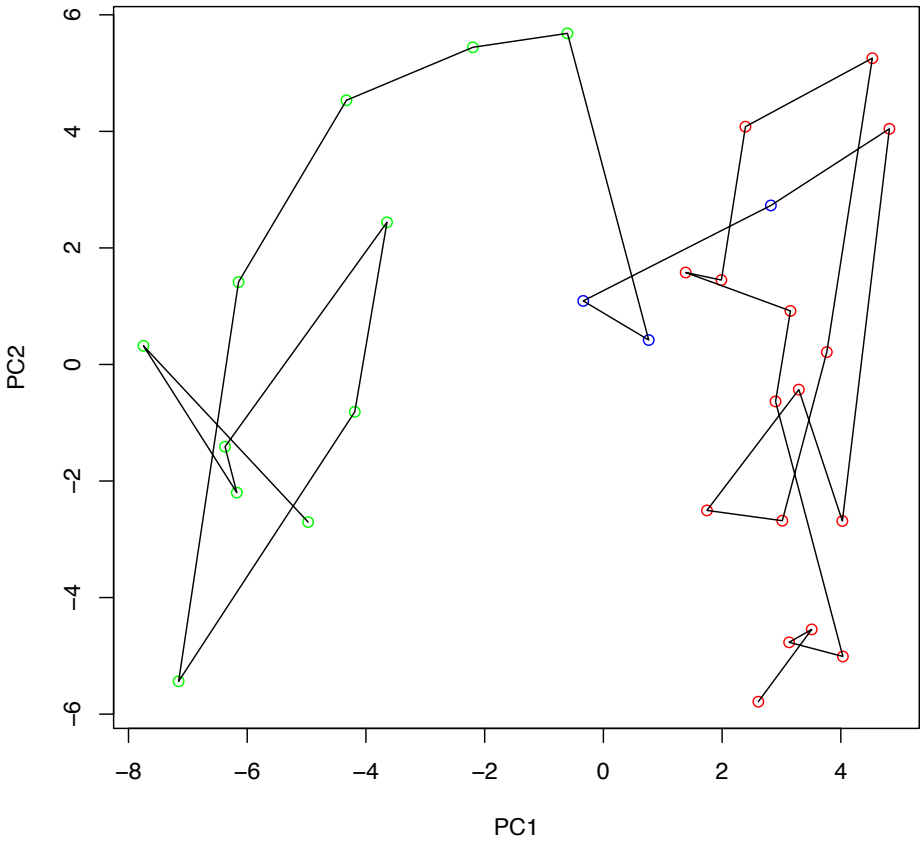


FIGURE 17 – Expérience 5 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

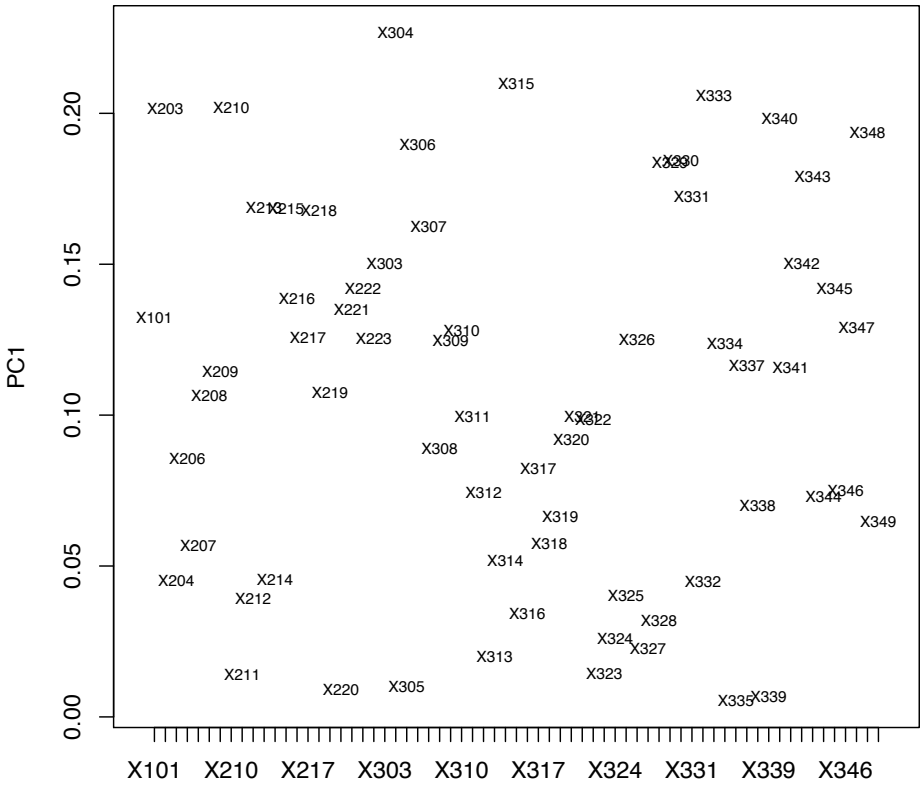


FIGURE 18 – Expérience 5 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1. La variable X_{304} est celle qui participe le plus à la formation de l'axe PC1.

3.3.8 Expérience 6

Nous avons voulu tester ces résultats sur le deuxième jeu de données. Malheureusement, ces données ne contiennent qu'une partie des variables et notamment que les variables de type $X_{3\star\star}$. Il manque toutes les informations sur les variables de type $X_{2\star\star}$. Nous avons tout de même réalisé une ACP sur ce jeu de données. Nous avons attribué aux semaines les 4 mêmes labels d'avant, de pré, de pendant et d'après en nous basant sur les résultats précédents. Nous obtenons le résultat présenté Figure 19 qui nous montre que l'on n'arrive plus à tirer d'informations pertinentes.

Nous concluons de cette expérience que les variables de type $X_{2\star\star}$ doivent avoir un rôle important dans les apparitions des fluctuations.

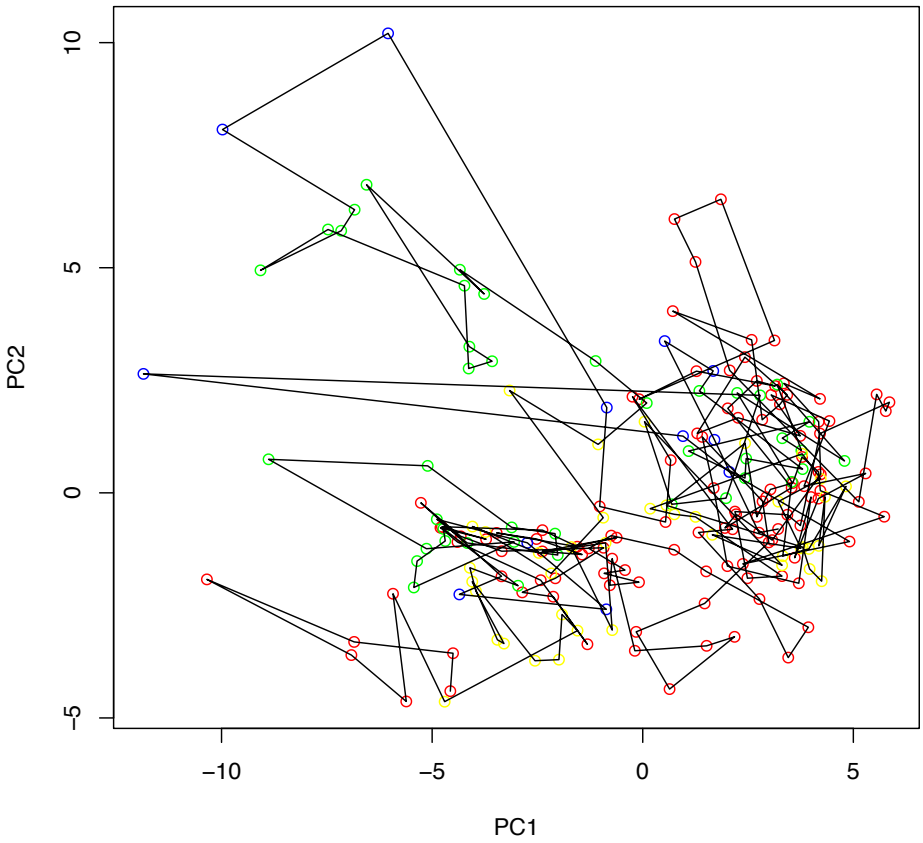


FIGURE 19 – Expérience 6 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

3.4 Conclusion de l'ACP

Nous avons appliqué la méthode d'ACP sur plusieurs sous-ensembles des données initiales afin :

- d'expliquer les périodes avec et les périodes sans fluctuations,
- de différencier les périodes d'avant et de pré,
- de se focaliser sur les périodes avant et pendant (33 premières semaines),
- de projeter d'autres mesures afin de prédire les fluctuations.

Nous avons essayé d'interpréter les résultats obtenus en relevant certaines variables pouvant expliquer les différentes périodes. Ces propositions sont à considérer avec prudence : il faut s'assurer qu'elles sont cohérentes avec le problème réel.

4 Arbres de décision

Dans cette section, nous nous intéressons à une autre méthode étudiée pour résoudre notre problème : les arbres de décision.

4.1 Description de la méthode

Les arbres de décision permettent de prédire les valeurs prises par une variable, dite cible, à partir d'un ensemble de variables, dites prédictives. Cette méthode nous a paru particulièrement attractive pour deux raisons : sa lisibilité et sa capacité à sélectionner des variables discriminantes dans un fichier de données contenant un très grand nombre de variables.

Parmi les différents algorithmes envisageables, nous avons utilisé la méthode CART (Classification And Regression Tree). Cette méthode est implémentée, entre autres, dans le package *rpart* du logiciel *R* (voir [2]).

Dans notre étude, la variable cible est l'état du système regroupé dans les 4 classes proposées lors de l'analyse en composantes principales. Pour rappel, il s'agit des périodes :

- avant : semaines 4 à 19 (Av),
- pré : semaines 20 à 22 (Pre),
- pendant : semaines 23 à 33 (Pdt),
- après : semaines 34 à 46 (Ap).

Les variables explicatives retenues sont les 67 variables correspondant aux mesures effectuées au cours de l'année 2010. Pour mettre en œuvre la méthode, il convient alors de séparer les données en deux classes : apprentissage et validation. Les premières permettent de construire l'arbre de décision, c'est la phase d'apprentissage. Cette construction repose sur les éléments suivants :

1. un critère de segmentation : le coefficent de Gini pour la méthode CART. Il permet de déterminer la variable explicative et le seuil qui séparent au "mieux" les individus d'une population, ici les données d'appentissage, dans les différentes classes de la variable cible. En pratique, on retient le couple variable–seuil qui maximise ce coefficient,
2. l'élagage : il s'agit de trouver l'arbre le plus petit possible ayant la plus grande performance possible. Toutefois, puisque les arbres que nous avons construits sont petits, l'élagage n'apparait pas nécessaire.

Sans l'élagage, la construction de l'arbre est simple. On sépare les données d'apprentissages en deux groupes en fonction du critère de segmentation et on recommence sur chacun de ces groupes jusqu'à obtenir des groupes homogènes.
Il vient ensuite une phase de validation de l'arbre. Pour le jeu de données de validation, l'arbre de décision nous permet de prédire les classes d'appartenance de chaque individu. On compare alors ces résultats avec la réalité et on évalue le taux d'erreur.

Notons qu'il aurait également été possible de construire un arbre de décision pour prédire la variable continue Y_1 directement. Le choix de considérer des classes nous a paru pertinent puisqu'il peut permettre de mieux mettre en valeur les comportements caractéristiques associés à chaque période, en diminuant l'influence du bruit résiduel lié à l'imprécision des mesures.

4.2 Résultats sur le premier jeu de données

L'apprentissage a été fait en utilisant la moitié des données disponibles, plus précisément sur les données correspondant aux semaines paires. Les variables qui définissent l'arbre obtenu

et représenté sur la figure 20 sont X_{203}, X_{212}, et X_{306}. On a ensuite fait la validation sur les données correspondant aux semaines impaires, ce qui a donné la matrice de confusion suivante :

Réel \Prédit	Avant	Sous peu	Pendant	Après
Avant	8	0	0	0
Sous peu	0	1	0	0
Pendant	0	1	4	1
Après	0	0	0	6

Le taux d'erreur est de 9.5%, ce qui est faible comparé aux taux d'erreur habituels des arbres de décision.

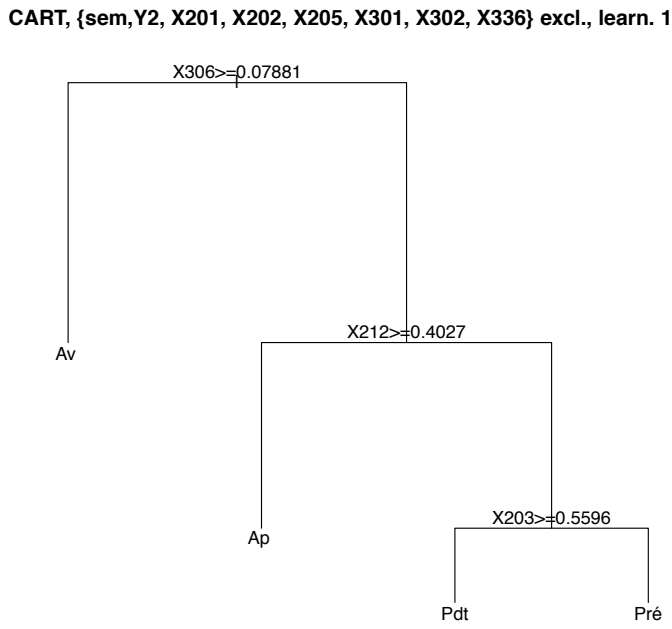


FIGURE 20 – Arbre de décision (méthode CART) créé en retirant les variables Y_2, X_{201}, X_{202}, X_{205}, X_{301}, X_{302} et X_{336}. L'apprentissage a été fait à partir des données correspondant aux semaines paires.

On a aussi fait un test de robustesse en bruitant légèrement les données de validation (bruit gaussien, dont la variance est égale à 10% de la valeur de la variable considérée). Le taux d'erreur était alors de 4.7%.

Toutefois, le nombre de données disponibles est trop faible pour en conclure que l'arbre est performant. Ainsi, la variable X_{306} est non significative selon André Augé. On a, par conséquent, recommencé l'expérience en retirant la variable X_{306} du premier jeu de données. L'arbre créé est représenté sur la figure 21. Les variables qui ressortent sont X_{203}, X_{212} et X_{343}. Le taux d'erreur est maintenant de 19% et la matrice de confusion est la suivante :

Réel \Prédit	Avant	Sous peu	Pendant	Après
Avant	6	2	0	0
Sous peu	0	1	0	0
Pendant	0	1	4	1
Après	0	0	0	6

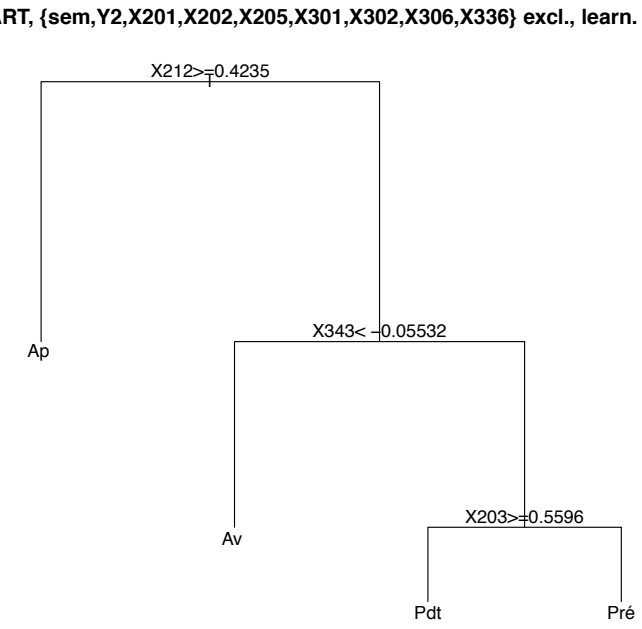


FIGURE 21 – Arbre de décision (méthode CART) créé en retirant les variables Y_2, X_{201}, X_{202}, X_{205}, X_{301}, X_{302}, X_{306} et X_{336}. L'apprentissage a été fait à partir des données correspondants aux semaines paires.

Si on fait l'apprentissage sur toutes les données du premier jeu, on construit un arbre légèrement différent (voir la figure 22). On ne peut malheureusement pas évaluer sa performance sur des données de validation puisqu'on ne dispose pas des données $X_{2\star\star}$ pour les années précédentes.

4.3 Résultats sur le deuxième jeu de données

On a également mis en œuvre la méthode CART sur le deuxième jeu de données. Plus précisément, on a construit l'arbre (voir la figure 23) avec la totalité des 43 données associées à l'année 2010 et on a fait la validation sur les autres données. Il apparaît que la seule période détectée est celle de 2010 sur laquelle l'apprentissage avait été réalisé alors qu'il y a eu 3 autres périodes entre 2007 et 2009. On peut fournir deux explications possibles :

- certaines des variables $X_{2\star\star}$ sont vraiment importantes et on peut difficilement s'en passer,
- la méthode d'arbre de décision est performante quand elle est utilisée sur la période sur laquelle s'est fait l'apprentissage (même légèrement bruitée), mais est incapable de détecter les autres périodes.

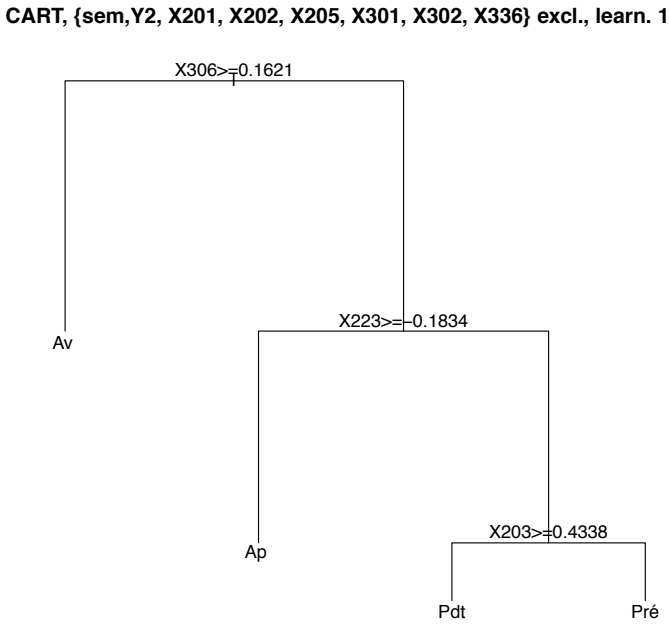


FIGURE 22 – Arbre de décision (méthode CART) créé en retirant les variables Y_2, X_{201}, X_{202}, X_{205}, X_{301}, X_{302} et X_{336}. L'apprentissage a été fait à partir de toutes les données (donc on ne peut pas faire de validation croisée).

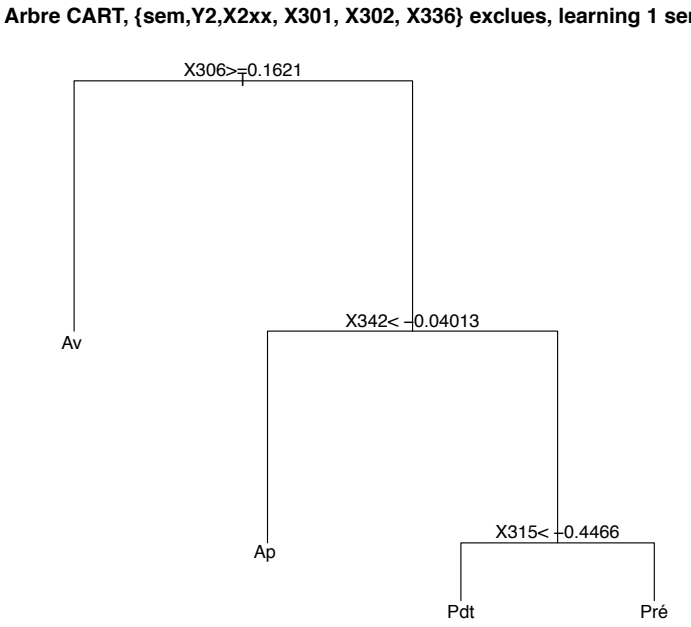


FIGURE 23 – Arbre de décision (méthode CART) créé en retirant les variables Y_2, $X_{2\star\star}$, X_{301}, X_{302} et X_{336}, dédié au test sur plusieurs années. L'apprentissage a été fait à partir de toutes les données (donc on ne peut pas faire de validation croisée).

D'autres essais ont été réalisés en changeant le critère de séparation de l'arbre (coefficient de Gini remplacé par la méthode ANOVA ou la divergence de Kullback-Leibler implémentées dans le package *rpart*). Ces autres arbres ne font pas apparaître d'autres variables que celles déjà présentées et présentent un taux d'erreur plus important.

4.4 Conclusion

Selon la méthode des arbres de décision, les variables X_{203}, X_{212}, X_{223} et X_{306} semblent jouer un rôle déterminant dans les fluctuations de 2010. La méthode est particulièrement performante lorsque la validation a lieu sur les données (éventuellement bruitées) issues de la même période. Toutefois, il faut se méfier des résultats obtenus puisque le nombre de données à notre disposition est trop faible. L'apparition de X_{306} dans l'arbre alors qu'elle n'est pas une variable pertinente, est ainsi symptomatique.

D'autre part, l'arbre de décision construit sur les données de 2010 échoue à identifier les périodes de fluctuations sur les données des années précédentes. La diminution du nombre de variables considérées sur cette période peut expliquer, en partie, cette contre-performance. Pour améliorer ces résultats, il pourrait être intéressant de construire un arbre de décision à partir des données sur 2 ou 3 périodes et de le tester sur autres.

Pour conclure, nous donnons quelques pistes que nous n'avons pas eu le temps d'explorer : travailler sur la variable cible Y_1 ; redéfinir les classes d'une manière sensiblement différente (par exemple : loin de la période, juste avant, pendant et juste après) ; exploiter d'autres critères de segmentation, le lien du χ^2 par exemple.

5 Autres méthodes explorées mais non abouties

5.1 Régression linéaire

Une des premières méthodes que nous avons étudiée a été la régression linéaire par les moindres carrés. L'idée est de construire un estimateur de la variable Y_1 comme une combinaison linéaire de certaines variables explicatives. Plus précisément, pour n variables $(X_{a_1}, \dots, X_{a_n})$ choisies parmi $(X_{101}, X_{2\star\star}, X_{3\star\star})$, on cherche $(\hat{\alpha}_1, \dots, \hat{\alpha}_n)$ tel que

$$\left\| Y_1 - \sum_{i=1}^{n} \hat{\alpha}_i X_{a_i} \right\|_2 = \min_{\alpha \in \mathbb{R}^n} \left\| Y_1 - \sum_{i=1}^{n} \alpha_i X_{a_i} \right\|_2 .$$

On estime alors la qualité de l'estimateur $\hat{Y}_1 := \sum_{i=1}^{n} \hat{\alpha}_i X_{a_i}$ en calculant le coefficient de détermination R^2 défini par

$$R^2 := \frac{\sum_{k=1}^{43} \left(y_1^k - \hat{y}_1^k\right)^2}{\sum_{k=1}^{43} \left(y_1^k - \bar{y}_1\right)^2}$$

où $\bar{y}_1 = \frac{1}{43} \sum_{k=1}^{43} y_1^k$. Plus le coefficient R^2 est proche de 1, plus l'estimateur $\hat{Y}$ est performant. Toutefois, ce n'est pas parce qu'il est performant que le modèle est pertinent. Il convient de s'assurer également visuellement de la pertinence du modèle linéaire. Inversement, plus le coefficient R^2 est proche de 0, moins l'estimateur est performant. Toutefois, ce n'est pas parce que l'approximation linéaire n'est pas bonne qu'il n'existe pas un autre lien (par exemple quadratique) entre les variables $(X_{a_1}, \dots, X_{a_n})$ et Y_1.

Nous avons étudié la méthode de la régression linéaire en privilégiant un nombre restreint de variables. Plus précisément, on s'est démandé, étant donné un entier n petit, quel est le choix de variables qui optimise la performance de l'estimateur, ou de manière équivalente qui maximise le coefficient de détermination. Nous souhaitons ainsi déterminer des variables significatives et tester la validité du modèle linéaire. Nous présentons, dans la suite, les résultats obtenus pour $n = 1$, $n = 2$ et $n \geq 3$.

Si on prend une variable explicative, le coefficient de détermination est maximum pour la variable X_{315} et atteint $0,59$. Viennent ensuite les variables X_{304} et X_{309} avec des coefficients de détermination supérieurs à $0,5$. Ces valeurs étant relativement petites, le modèle linéaire à une variable explicative n'apparait pas très performant. En complément de cette analyse, nous

avons représenté les nuages de points et les droites de régression associées pour la régression linéaire de Y_1 par X_{315} (voir la figure 24) et par X_{309} (voir la figure 25).

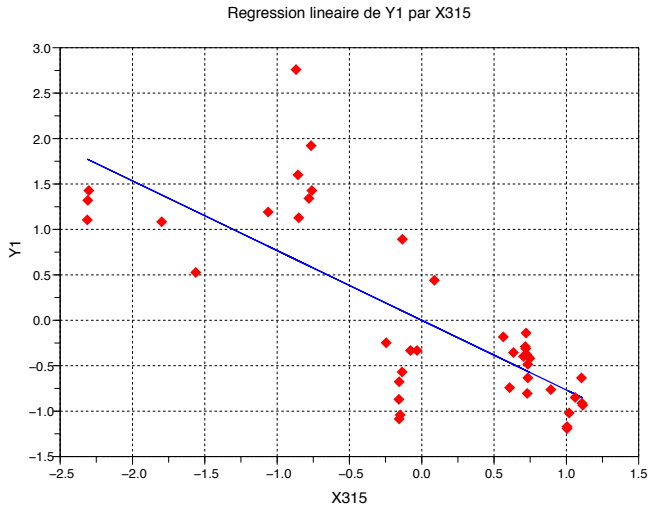


FIGURE 24 – Régression linéaire de Y_1 par X_{315}. Le coefficient de détermination vaut 0.59.

Si on prend deux variables explicatives, le coefficient de détermination est maximum pour le couple (X_{309}, X_{315}) et atteint 0.85. Viennent ensuite les couples (X_{223}, X_{315}) et (X_{218}, X_{315}) dont les coefficients de détermination sont inférieurs à 0.78. Bien que mécaniquement le coefficient de détermination est amené à croître avec le nombre de variables explicatives choisi, le gain du passage à deux variables est très marqué dans ce cas. Le coefficient de détermination semble indiquer que l'estimateur $\hat{Y}_1 = 0.53X_{309} - 0.61X_{315}$ est performant. En complément de cette analyse, nous avons représenté le nuage de points et le plan de régression pour la régression linéaire de Y_1 par le couple (X_{309}, X_{315}) dans la figure 26.

Si on considère plus de 2 variables explicatives, la performance du modèle augmente très peu et il ne nous semble pas pertinent de poursuivre au-delà de la dimension 2. Pour donner une idée, si on prend quatre variables explicatives, le coefficient de détermination n'excède pas 0.89 et, une nouvelle fois, les variables X_{309} et X_{315} apparaissent dans le quadruplet optimal. Ces dernières semblent être les variables les plus caractéristiques par cette méthode.

Nous n'avons pas plus approfondi la régression linéaire pour plusieurs raisons. D'une part, il nous a semblé avoir atteint ses limites au sens où ajouter des variables explicatives ne semble pas améliorer de manière significative sa performance. D'autre part, il est difficile de valider les résultats de la régression linéaire car notre échantillon de données n'est pas issu de tirages indépendants. Une méthode que nous n'avons pas eu le temps de mettre en œuvre et qui pourrait s'avérer particulièrement adaptée à la nature des données (dépendance des mesures, nombre de variables élevé), est la régression PLS (Partial Least Square).

5.2 Clusters

Une des premières méthodes utilisées a été le clustering qui consiste à regrouper les points des données en un nombre fixé de paquets. Un algorithme classique pour cette méthode est le k-means. Il est en général plus intéressant d'appliquer cet algorithme aux variables transformées par l'analyse en composantes principales (les premières variables ont le plus d'importance, les suivantes peuvent être négligées) qu'aux données brutes.

Parmi les variables fournies, on a retiré Y_2, X_{201}, X_{202}, X_{205} et X_{336} et on a seulement conservé les 7 premières composantes de l'analyse en composantes principales (de manière à

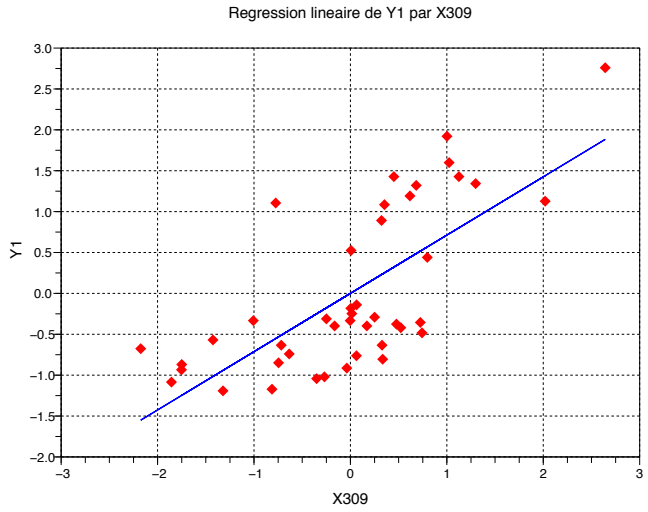


FIGURE 25 – Régression linéaire de Y_1 par X_{309}. Le coefficient de détermination vaut 0.51.

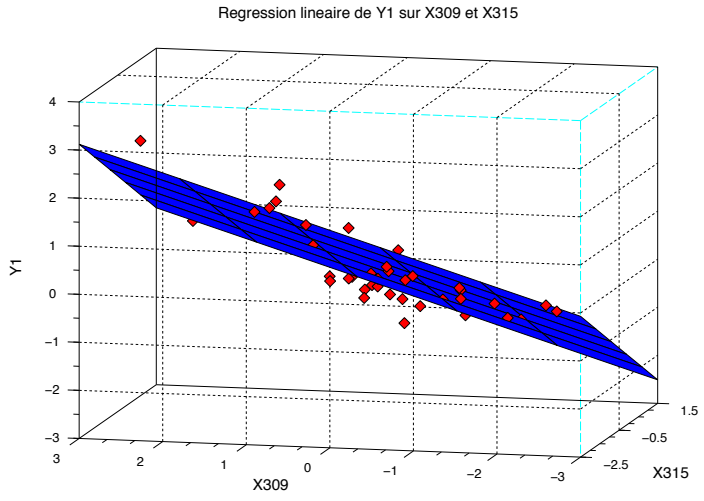


FIGURE 26 – Régression linéaire de Y_1 par X_{309} et X_{315}. Le coefficient de détermination vaut 0.85.

avoir 75% de la variance dans les composantes retenues). Si l'on fixe le nombre de paquets à 3, on constate que les 3 classes obtenues semblent bien différencier les périodes avant, pendant et après (voir Figure 27).

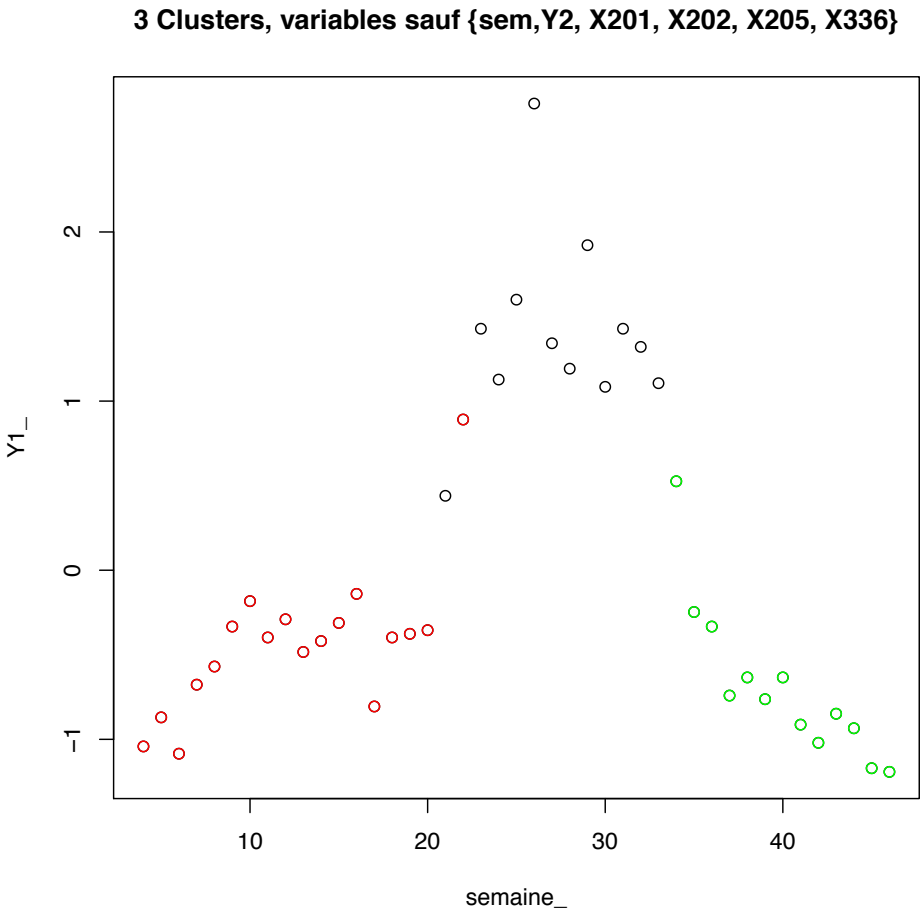


FIGURE 27 – Représentation des 3 clusters : Sur le graphe de la variable à expliquer pour chaque semaine, on a attribué à chaque point la couleur qui correspond au cluster auquel il appartient.

Si l'on fixe le nombre de paquets à 2, on pourrait s'attendre à ce qu'un des clusters corresponde à la période pendant et l'autre aux périodes avant ou après, mais on a seulement une séparation entre les 20 premières semaines et les 23 suivantes (voir Figure 28).

La méthode des clusters, si elle semble être une première approche intéressante, dépend de beaucoup de paramètres arbitraires (nombre de variables utilisées, nombre de classes). Cela risque d'avoir pour conséquence un pouvoir prédictif assez faible. Néanmoins, il convient de préciser que l'algorithme du k-means utilisé ici considère une distance euclidienne pour former les clusters, et ce n'est pas forcément le meilleur choix. D'autres choix de distance ou d'autres algorithmes de clustering pourraient s'avérer plus performants.

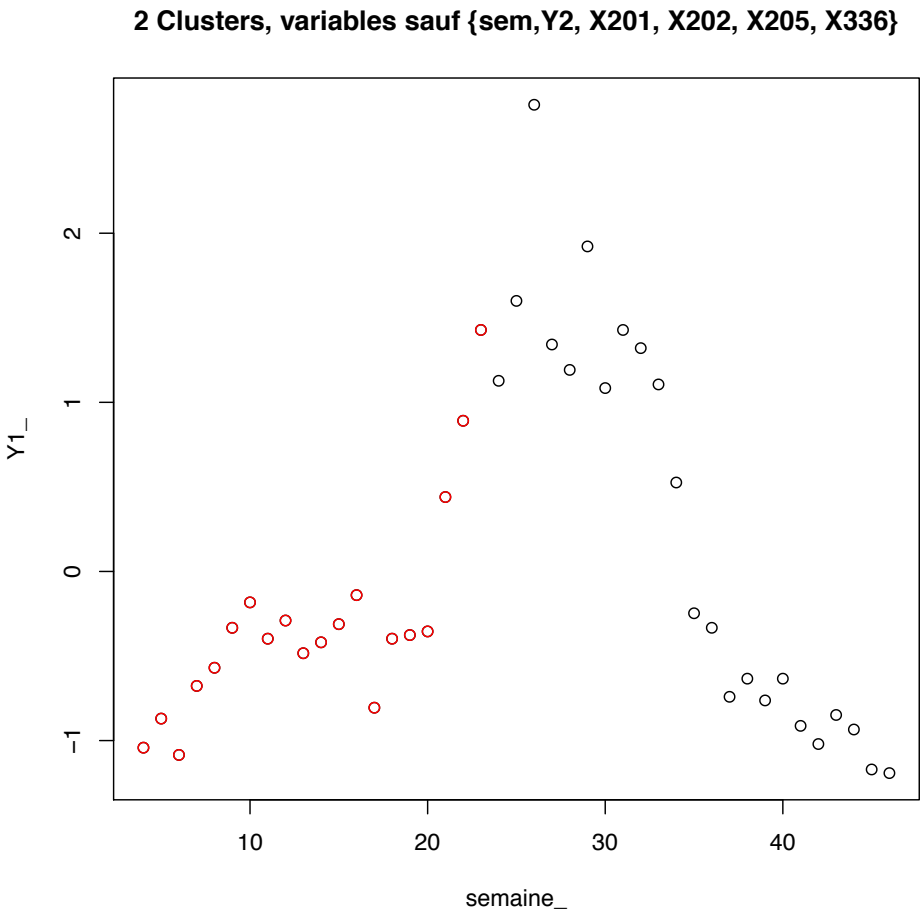


FIGURE 28 – Représentation des 2 clusters : Sur le graphe de la variable à expliquer pour chaque semaine, on a attribué à chaque point la couleur correspondant au cluster auquel il appartient.

6 Conclusion

Le problème rencontré par Rio Tinto Alcan et proposé par André Augé à l'occasion de la Semaine d'Étude Maths–Entreprises a été étudié sous l'angle des analyses multivariées. À partir des deux jeux de données correspondant aux mesures effectuées dans une usine du groupe, notre objectif était de mettre en évidence les variables pouvant être causes des fluctuations observées, afin que l'entreprise puisse les prédire. Pour y parvenir, nous avons appliqué plusieurs méthodes.

La première consiste à étudier la corrélation entre la variable à expliquer Y_1 et les autres variables, en fonction d'un déphasage (c'est-à-dire d'un décalage des courbes de quelques semaines). Certaines variables ont ainsi pu être identifiées comme étant plutôt des conséquences et d'autres des causes. Cependant, l'étude menée sur le jeu de données étendu aux quatre année ne confirme pas forcément les résultats obtenus sur le premier jeu de données. Cela pourrait s'expliquer par différentes natures de fluctuations.

L'analyse en composante principale, quant à elle, permet de séparer notre ensemble de semaines en quatre périodes et de mettre en évidence les variables qui semblent distinguer les différentes périodes. Nos tentatives d'explications sont à considérer avec prudence et à relier à la signification réelle des variables. Par ailleurs, elle apparaît comme un outil intéressant pour mettre en œuvre d'autres méthodes, comme les arbres de décision.

Ceux-ci permettent de trouver des variables significatives, dont la valeur détermine l'appartenance d'un point à différentes catégories. Les arbres créés ont un taux d'erreur plutôt faible, mais le manque de mesures nous empêche là encore d'apporter des conclusions claires.

D'autres méthodes ont été testées, mais n'ont malheureusement pas abouti. Nous espérons avoir quand même donné des pistes et des perspectives à André Augé, que notre étude en aveugle et notre regard de mathématiciens aient eu un intérêt, et souhaitons que son œil d'expert, que l'interprétation des variables et que la connaissance de leurs dépendances lui apporteront de meilleures conclusions.

Remerciements

Pour nous avoir donné la possibilité de participer à une telle expérience et pour leur accueil, leurs conseils et leur disponibilité, nous tenons tout d'abord à remercier les organisateurs de la Semaine d'Étude Maths–Entreprises : Simon Masnou, Bertrand Maury, Sylvain Faure et Thierry Dumont, ainsi que Régis Monneau, Didier Auroux et Magali Raynaud.

Pour nous avoir proposé un sujet très intéressant, pour ses explications enrichissantes et le temps qu'il a accordé à ce projet, nous remercions André Augé de l'entreprise Rio Tinto Alcan.

Pour nous avoir guidés dans notre découverte des statistiques et pour ses précieux avis, merci également à Gabriela Ciuperca.

Références

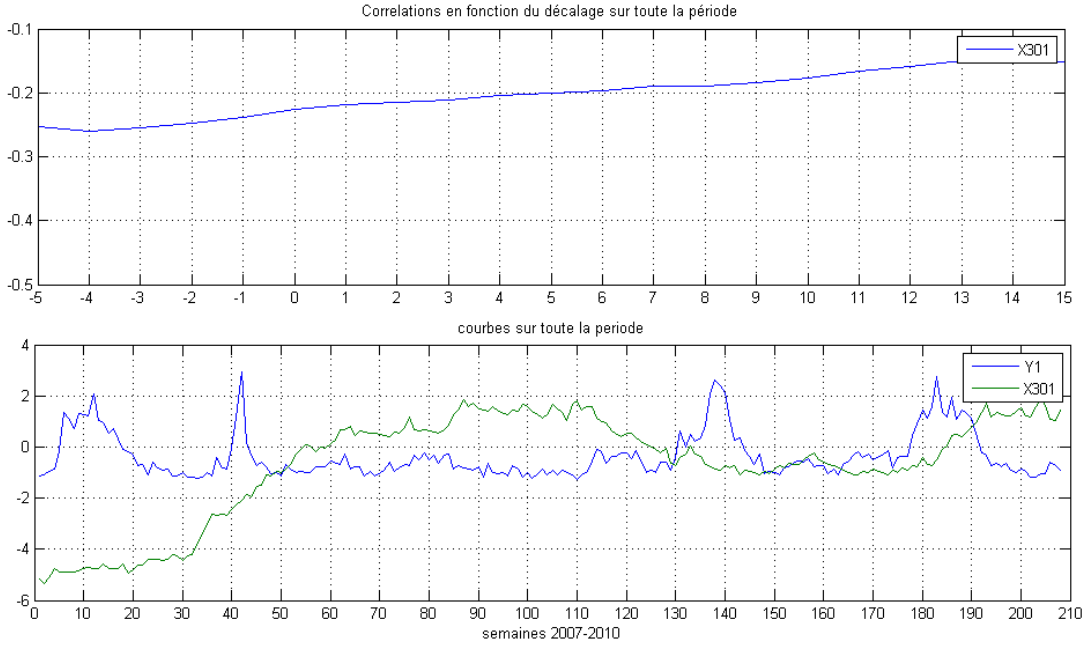
[1] C. Duby, S. Robin, Analyse en Composantes Principales, Institut National Agronomique Paris - Grignon, 2006

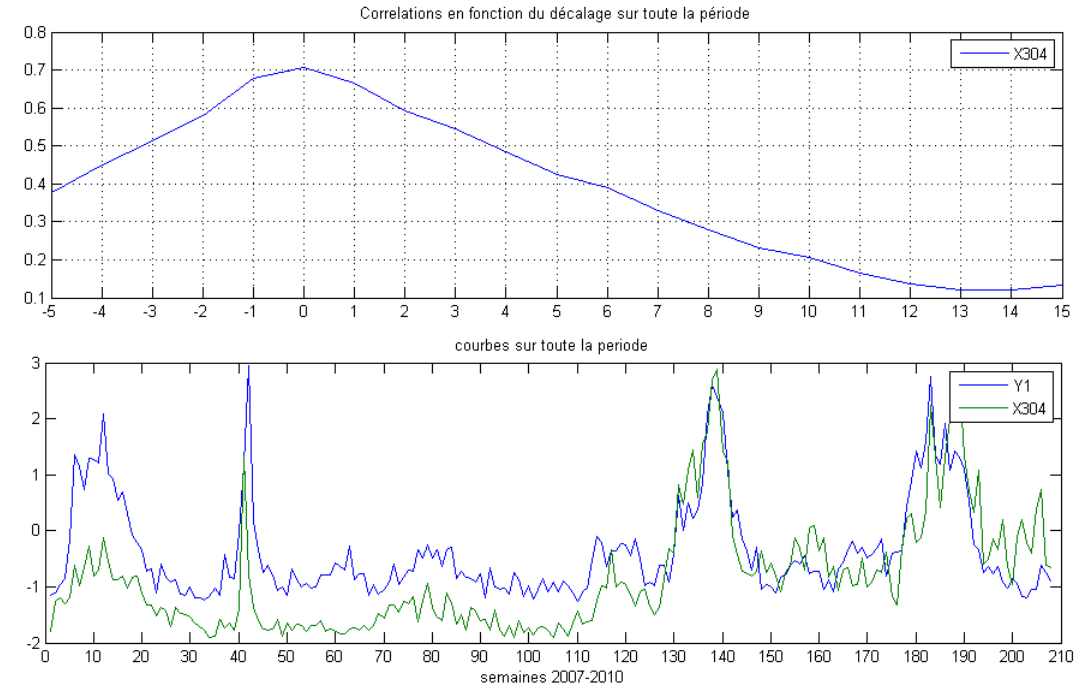
[2] Ricco Rakotomalala, Tutoriels Tanagra pour le Data Mining, 2008

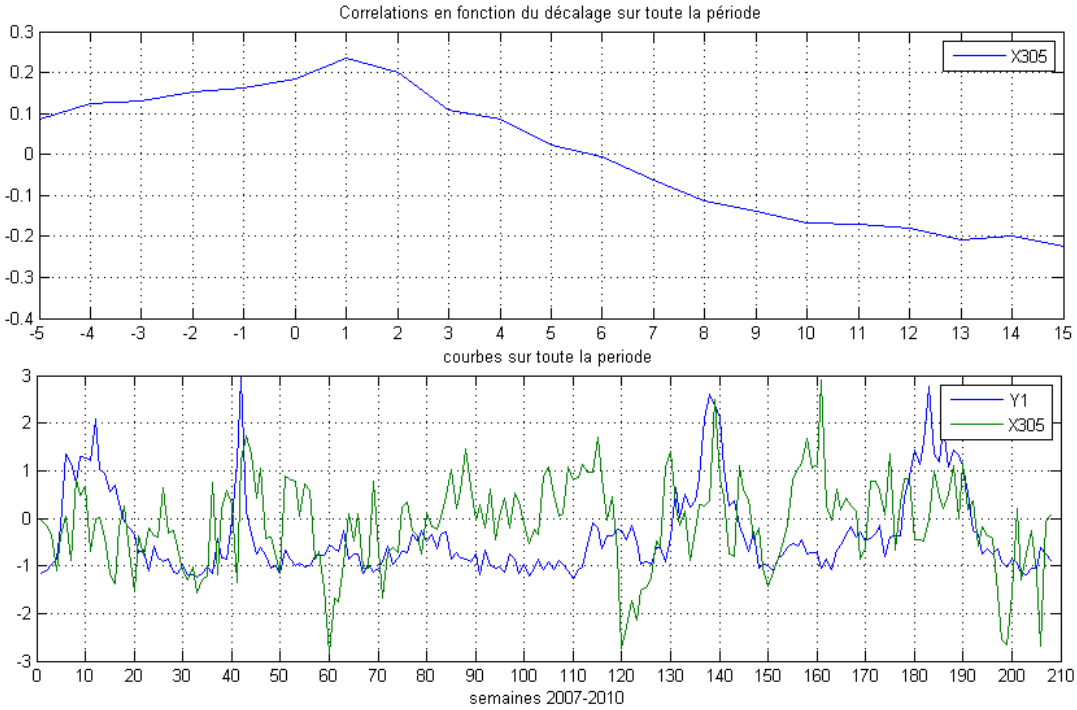
[3] `http://www.riotintoalcan.com/`

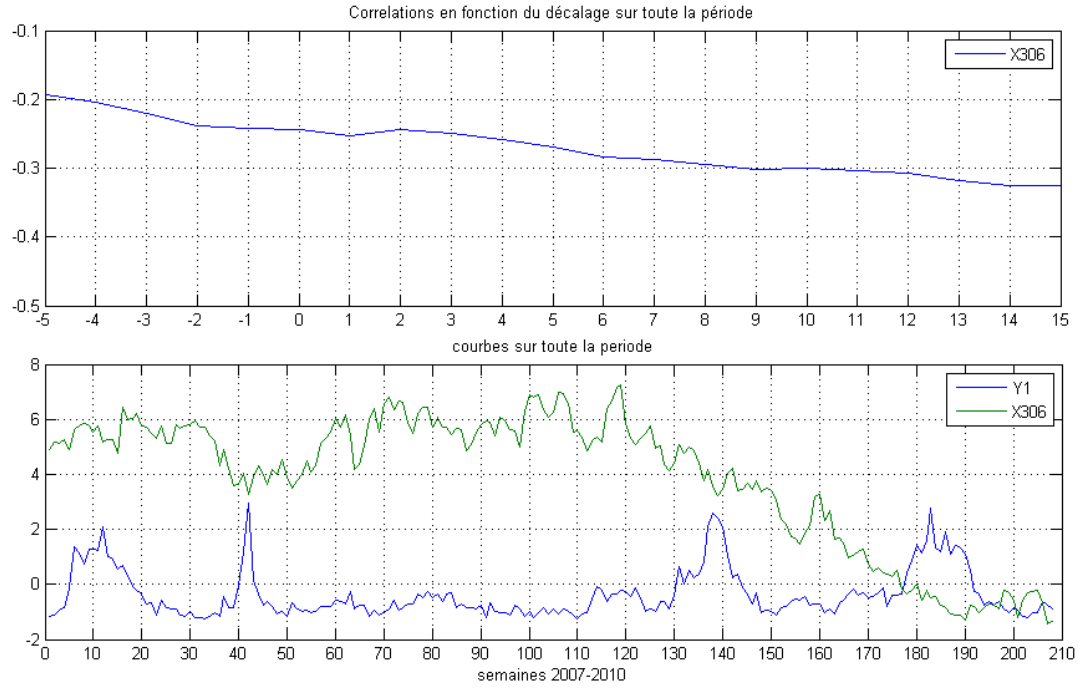
A Annexe sur la méthode de corrélation-déphasage

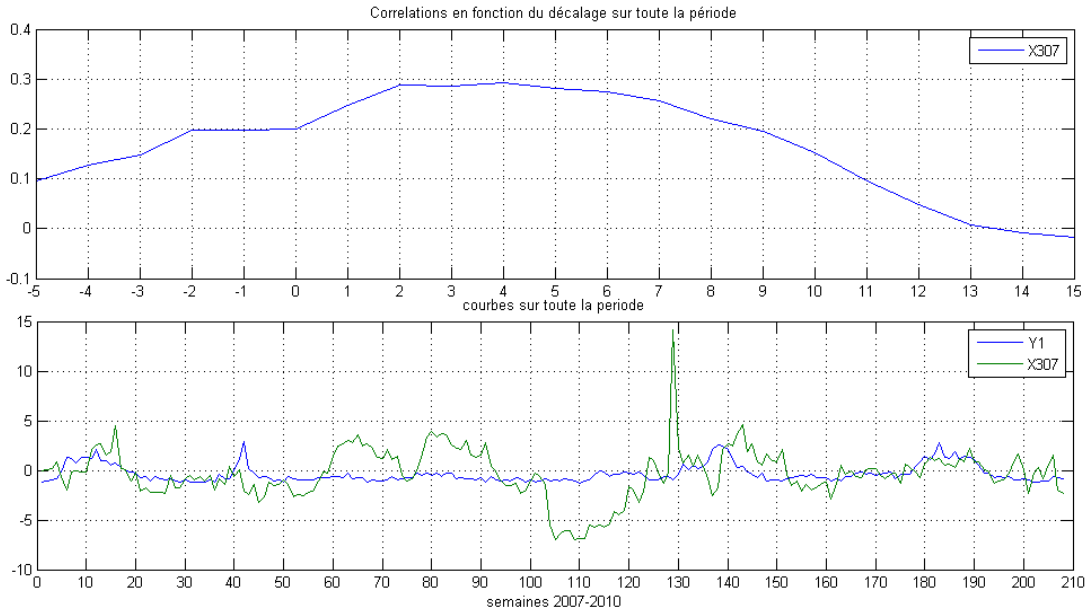
Nous donnons ici un ensemble de courbes représentant l'évolution de la corrélation entre Y_1 et $X_{3\star\star}$ lorsqu'on introduit un décalage de -5 à 15 semaines sur la donnée $X_{3\star\star}$, pour les données étendues X_{301} à X_{349}.

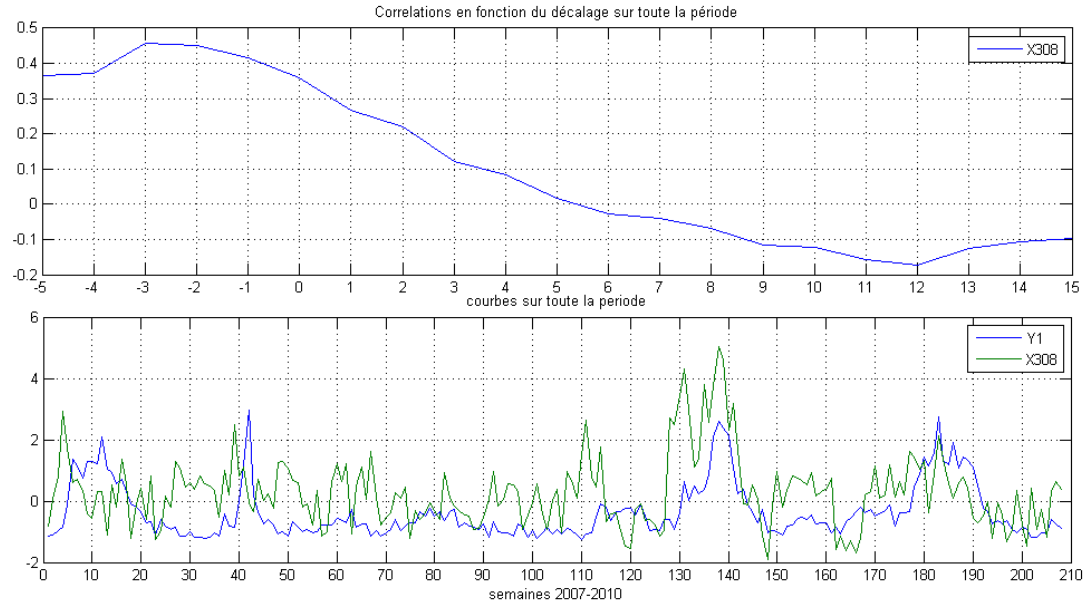


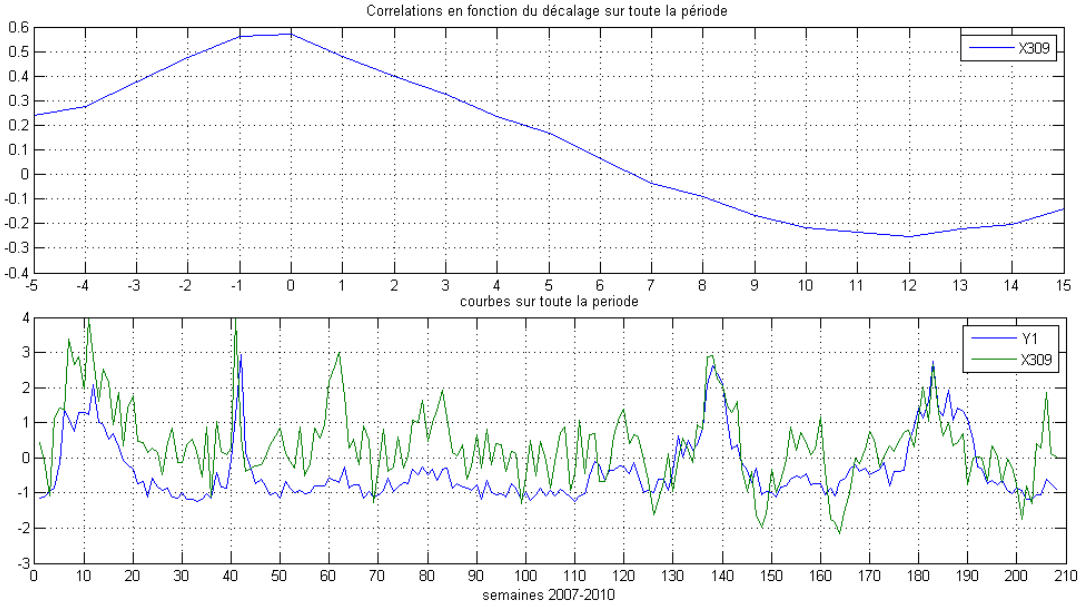


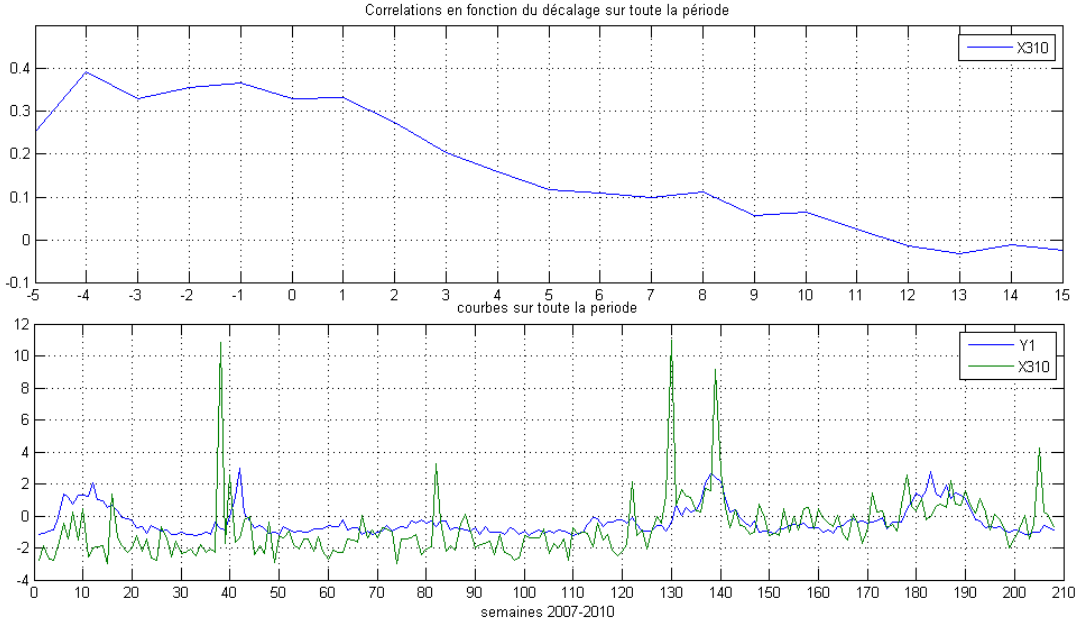

Correlations en fonction du décalage sur toute la période
X305
courbes sur toute la periode
Y1
X305
semaines 2007-2010

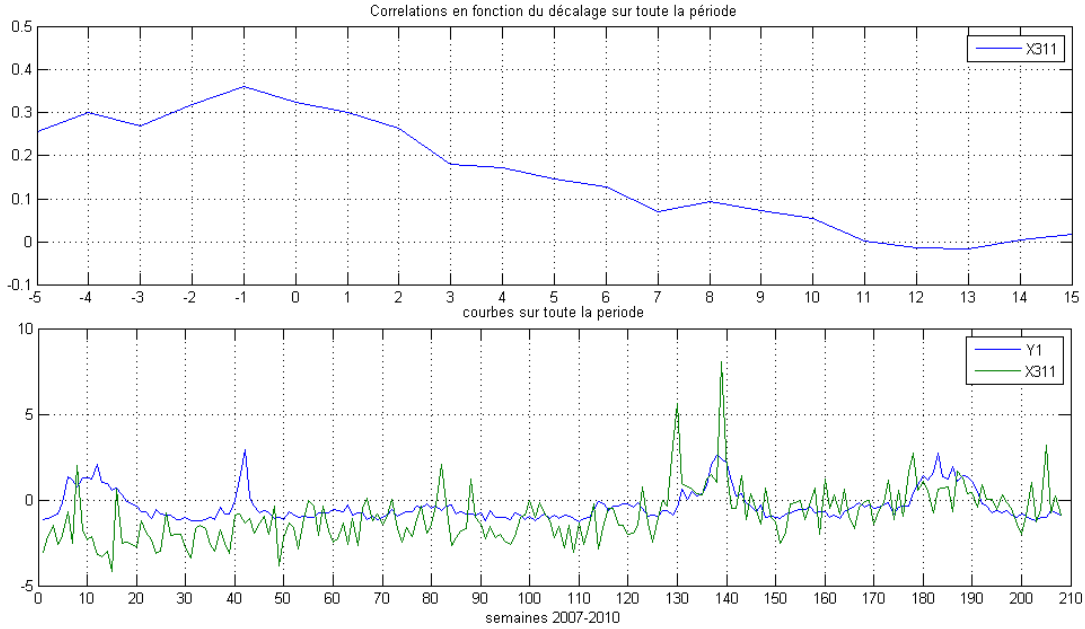

Correlations en fonction du décalage sur toute la période
X306
courbes sur toute la periode
Y1
X306
semaines 2007-2010

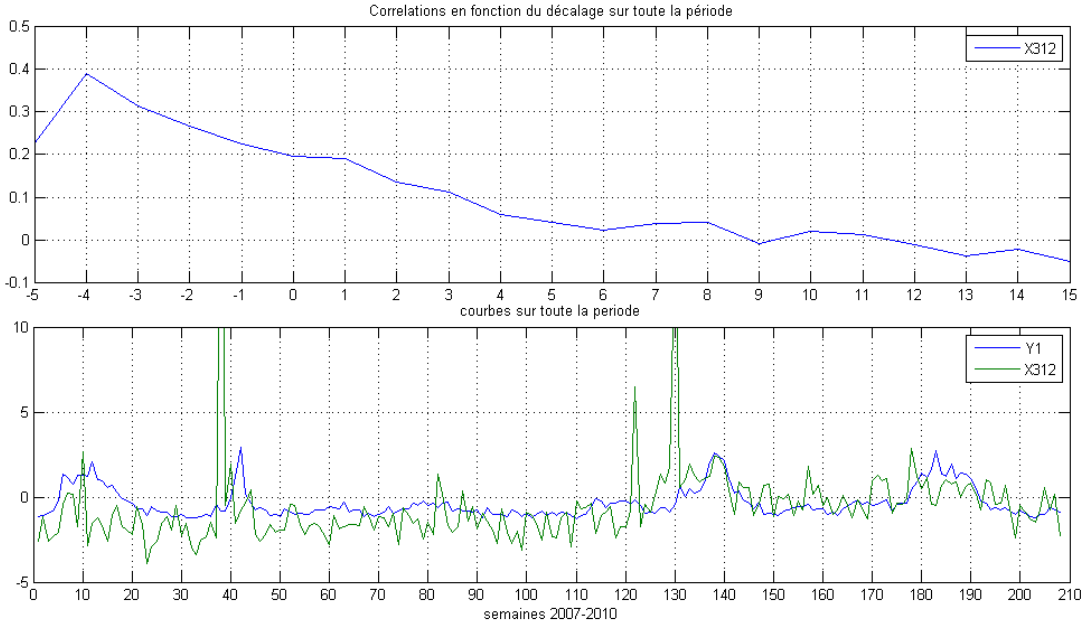

Correlations en fonction du décalage sur toute la période
X307
courbes sur toute la periode
Y1
X307
semaines 2007-2010

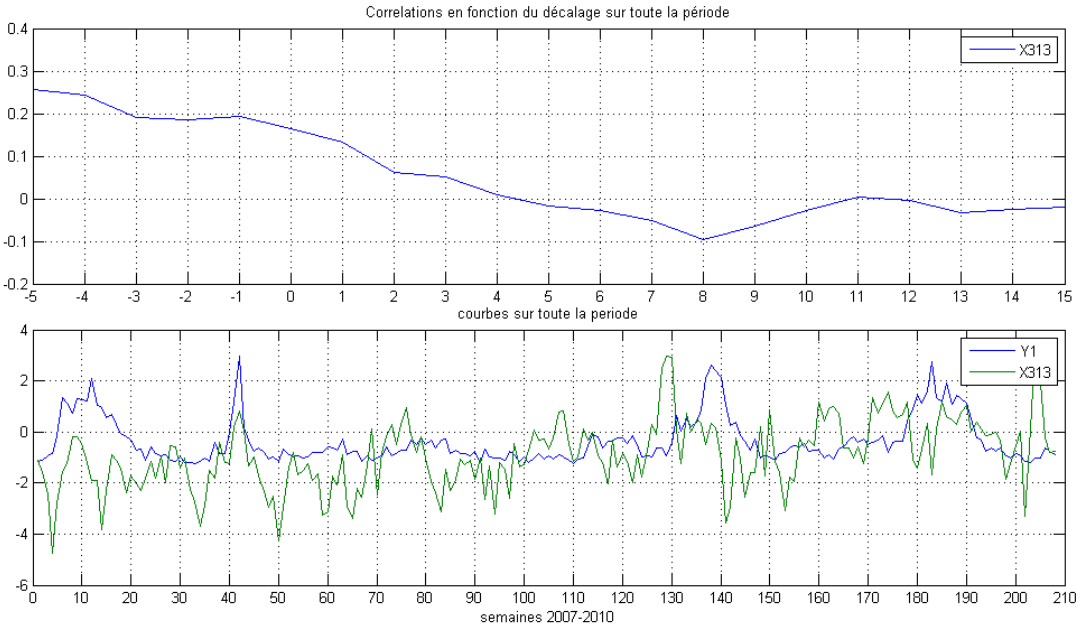

Correlations en fonction du décalage sur toute la période
X308
courbes sur toute la periode
Y1
X308
semaines 2007-2010

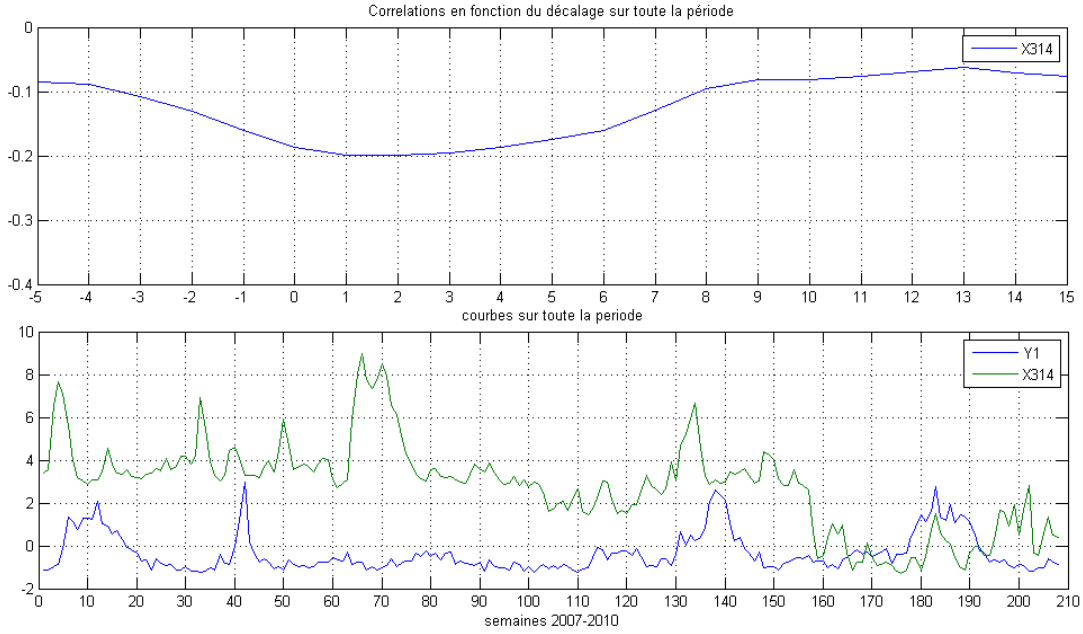

Correlations en fonction du décalage sur toute la période
X309
courbes sur toute la periode
Y1
X309
semaines 2007-2010

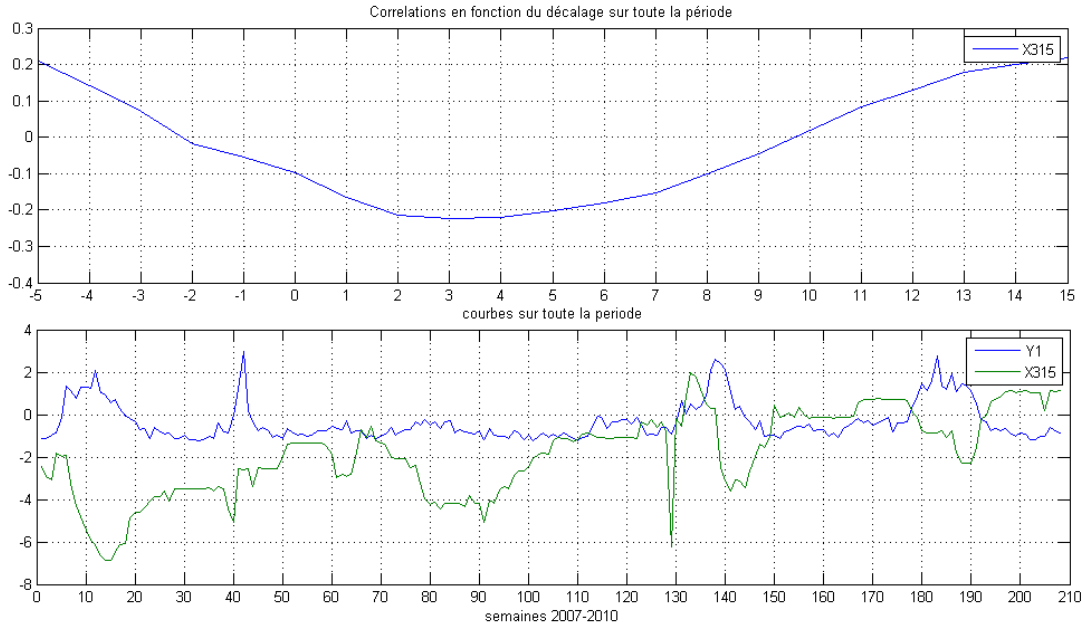

Correlations en fonction du décalage sur toute la période
X310
courbes sur toute la periode
Y1
X310
semaines 2007-2010

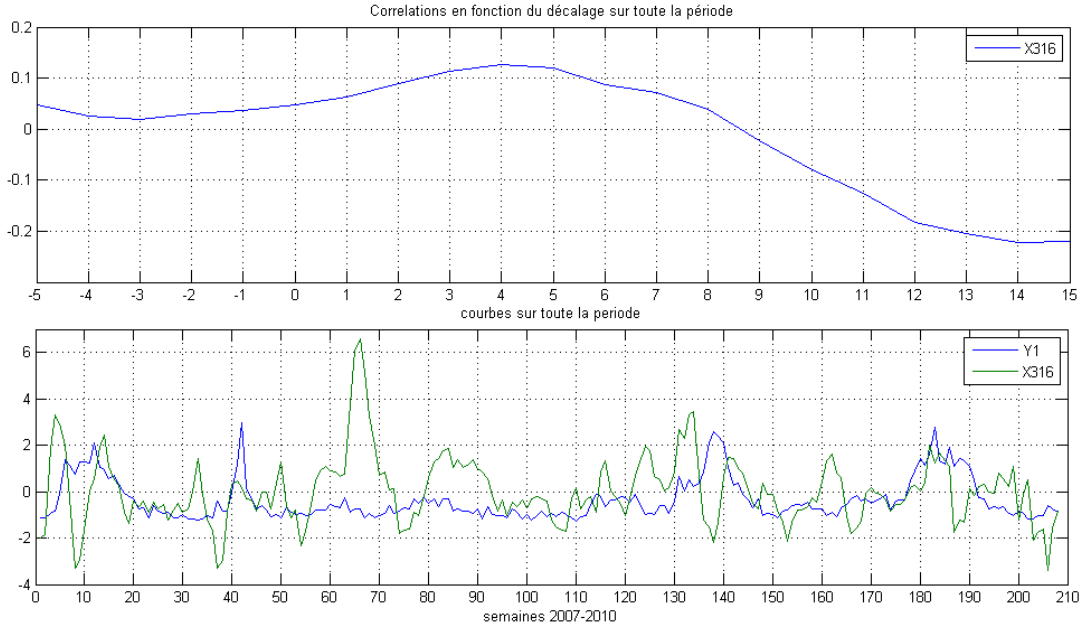

Correlations en fonction du décalage sur toute la période
X311
courbes sur toute la periode
Y1
X311
semaines 2007-2010

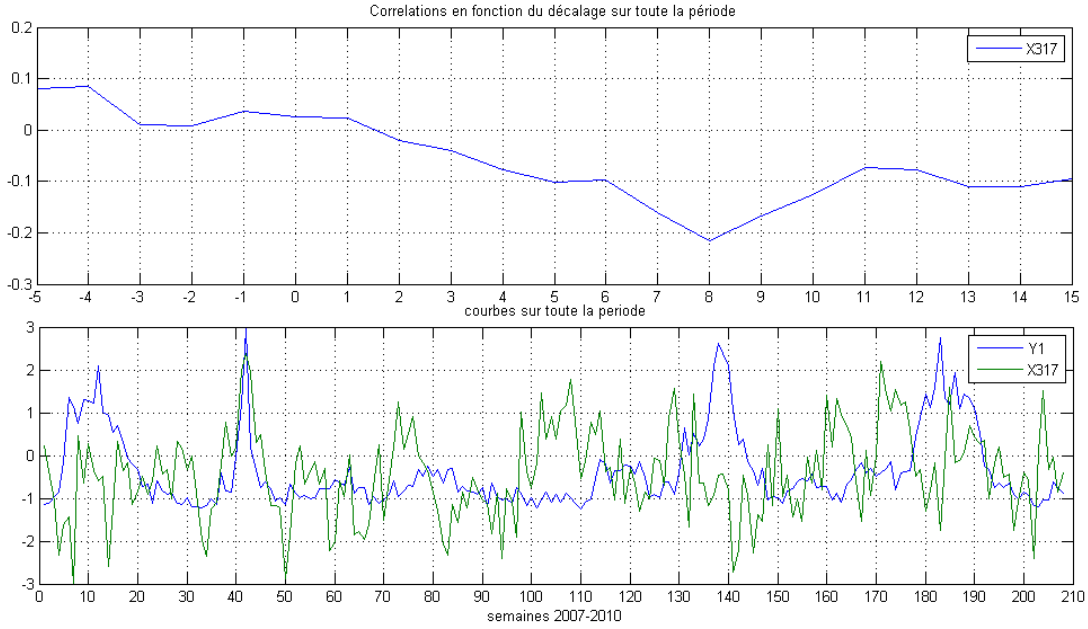

Correlations en fonction du décalage sur toute la période
X312
courbes sur toute la periode
Y1
X312
semaines 2007-2010

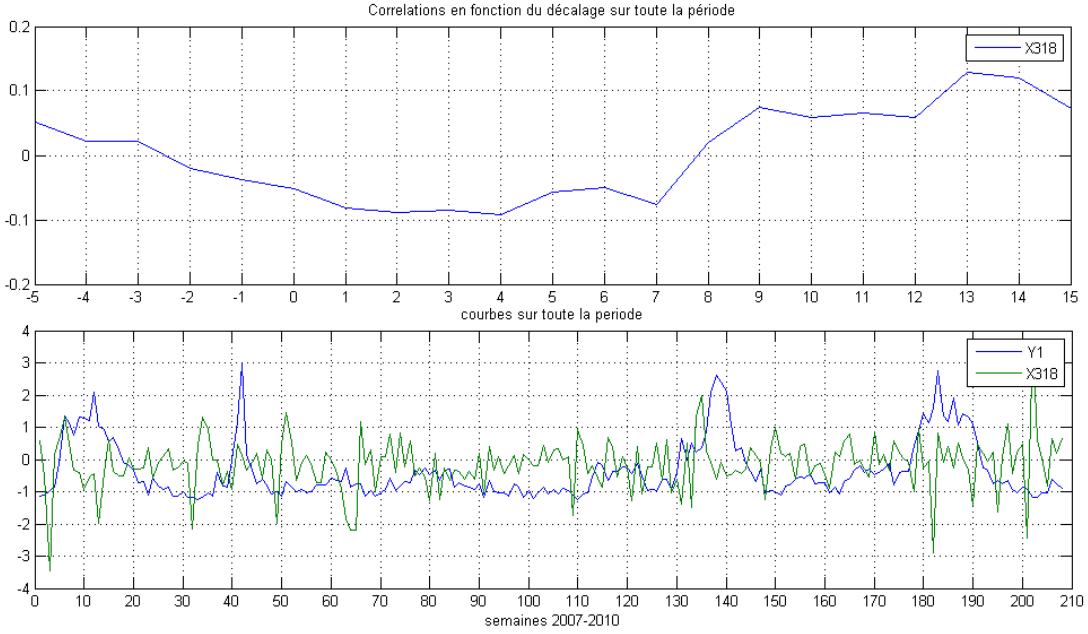

Correlations en fonction du décalage sur toute la période
X313
courbes sur toute la periode
Y1
X313
semaines 2007-2010

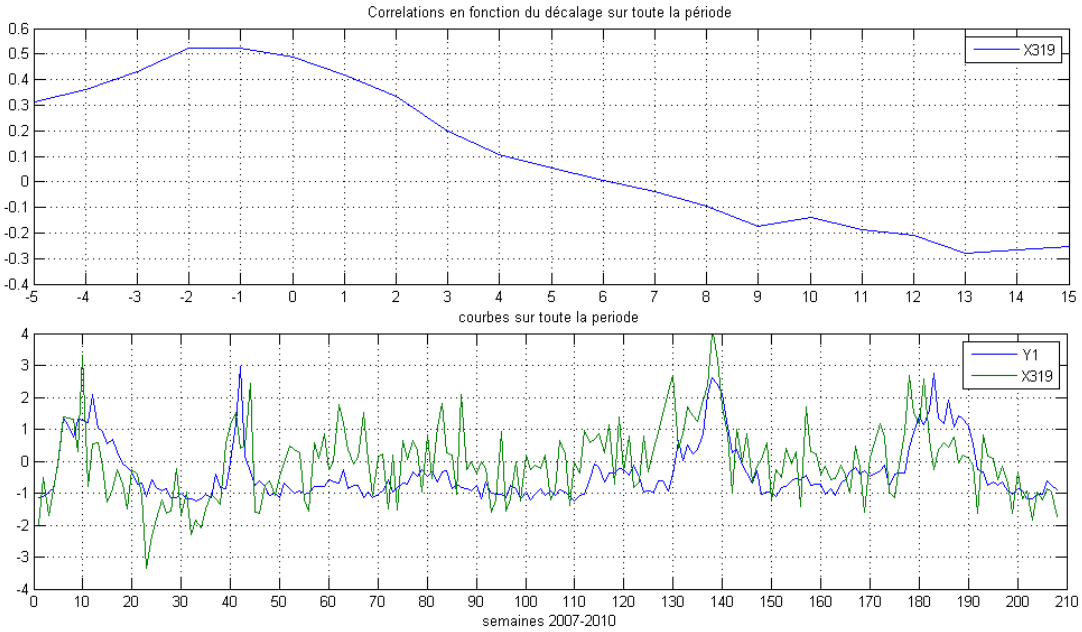

Correlations en fonction du décalage sur toute la période
X314
courbes sur toute la periode
Y1
X314
semaines 2007-2010

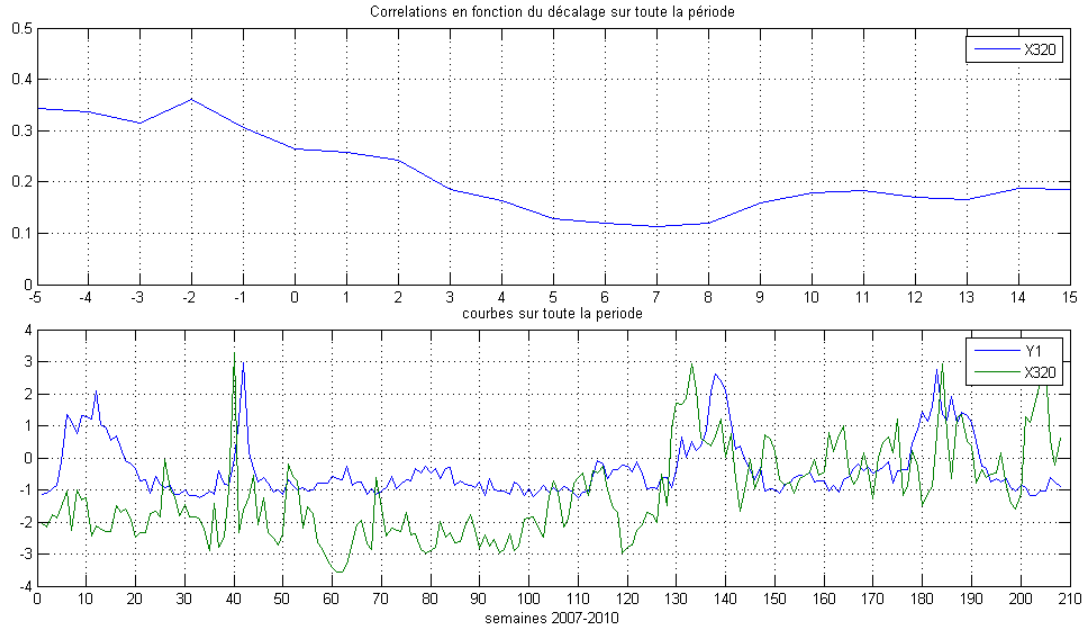

Correlations en fonction du décalage sur toute la période
X315
courbes sur toute la periode
Y1
X315
semaines 2007-2010

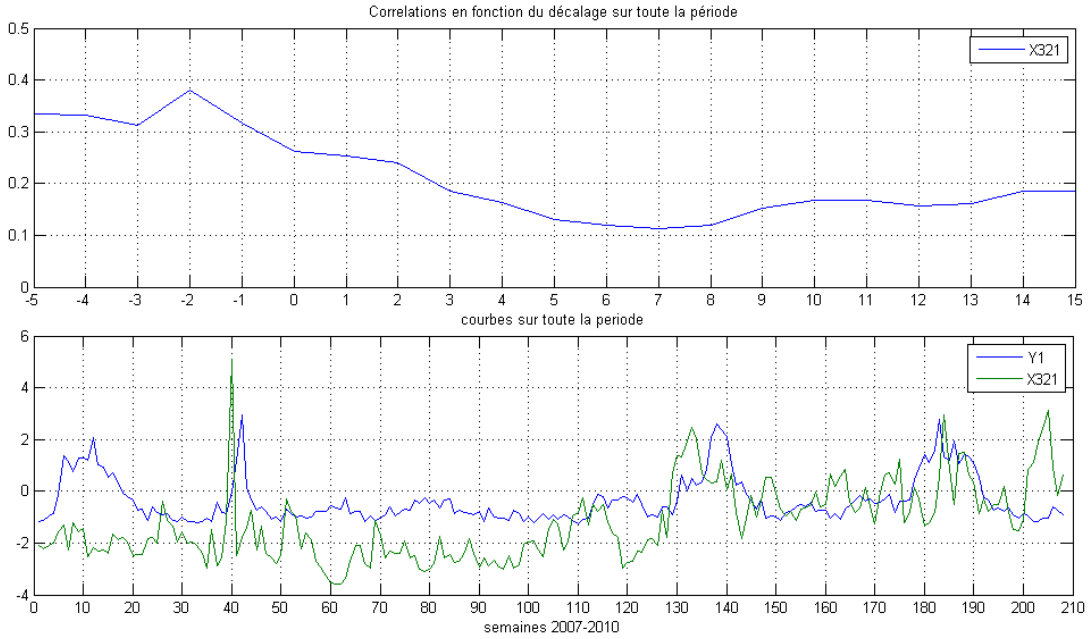

Correlations en fonction du décalage sur toute la période
X316
courbes sur toute la periode
Y1
X316
semaines 2007-2010

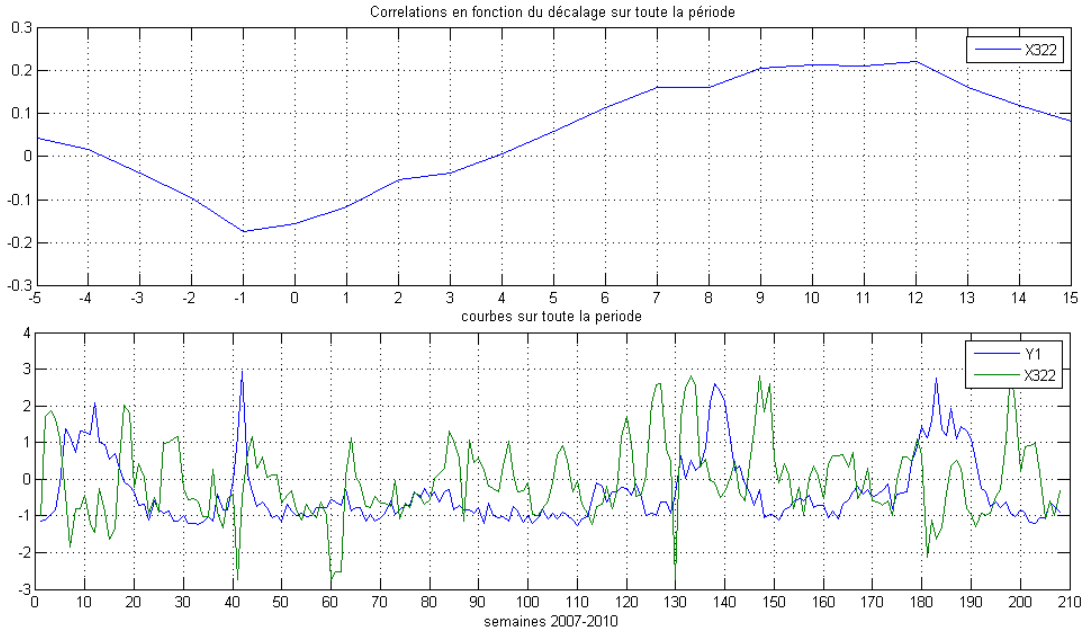

Correlations en fonction du décalage sur toute la période
X317
courbes sur toute la periode
Y1
X317
semaines 2007-2010

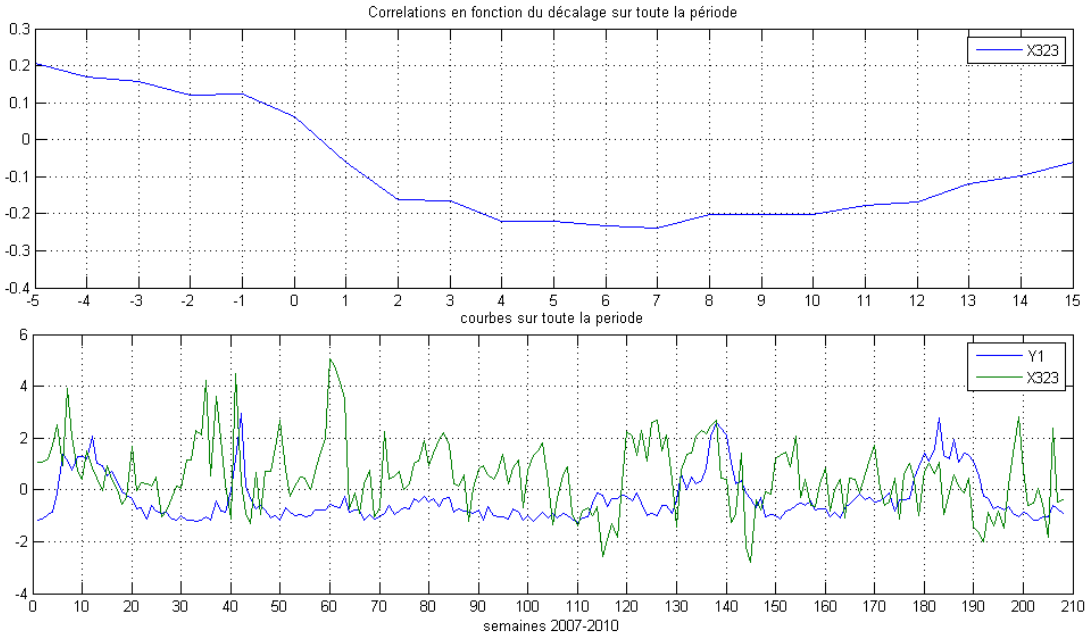

Correlations en fonction du décalage sur toute la période
X318
courbes sur toute la periode
Y1
X318
semaines 2007-2010

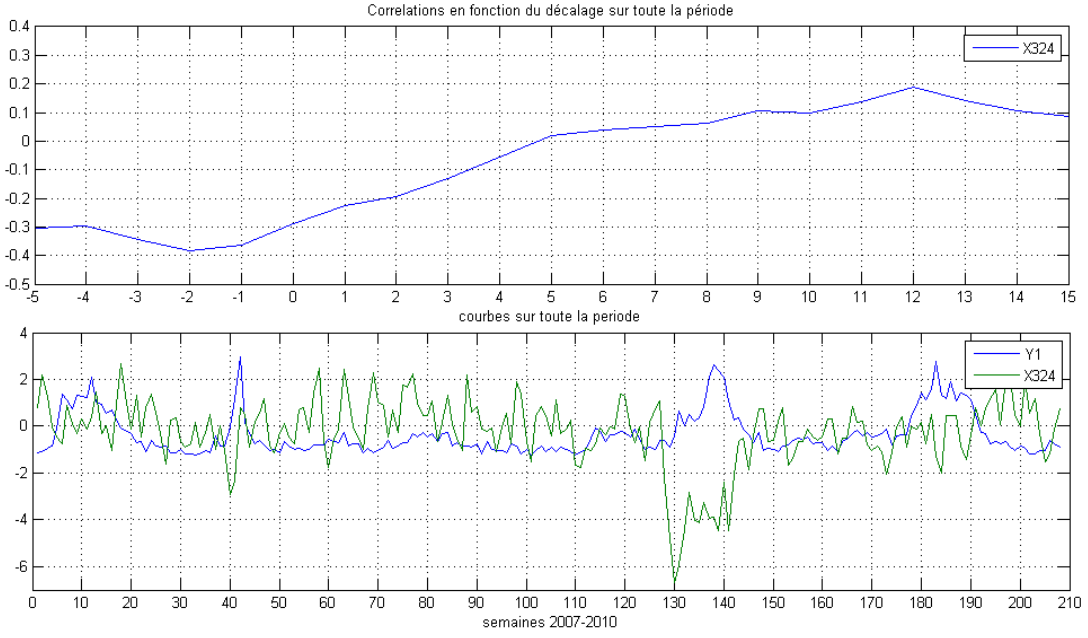

Correlations en fonction du décalage sur toute la période
X319
courbes sur toute la periode
Y1
X319
semaines 2007-2010

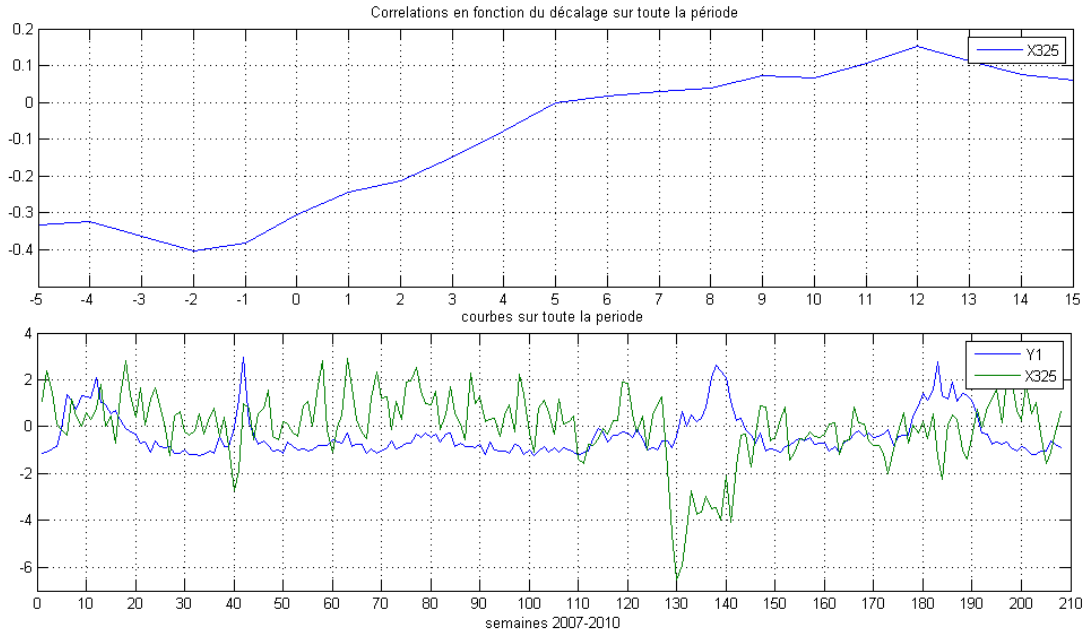

Correlations en fonction du décalage sur toute la période
X320
courbes sur toute la periode
Y1
X320
semaines 2007-2010

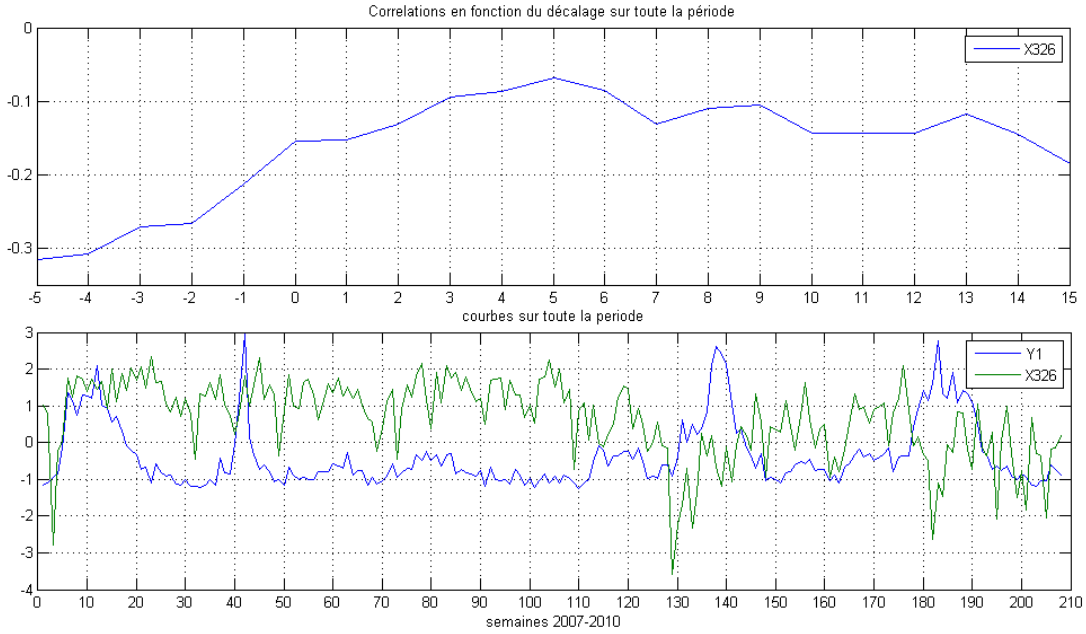

Correlations en fonction du décalage sur toute la période
X321
courbes sur toute la periode
Y1
X321
semaines 2007-2010

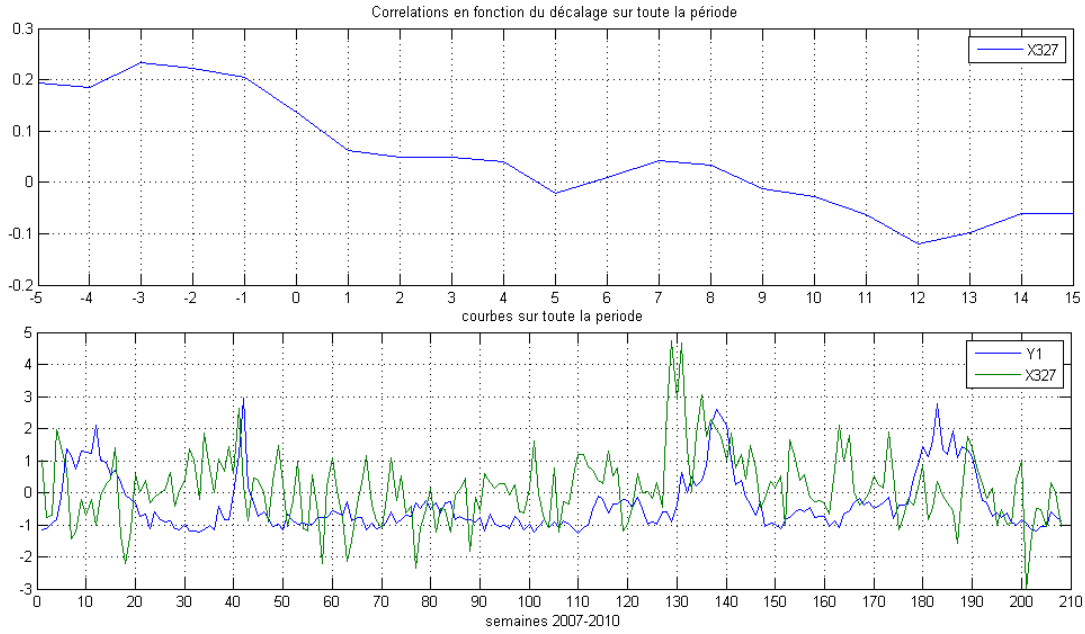

Correlations en fonction du décalage sur toute la période
X322
courbes sur toute la periode
Y1
X322
semaines 2007-2010

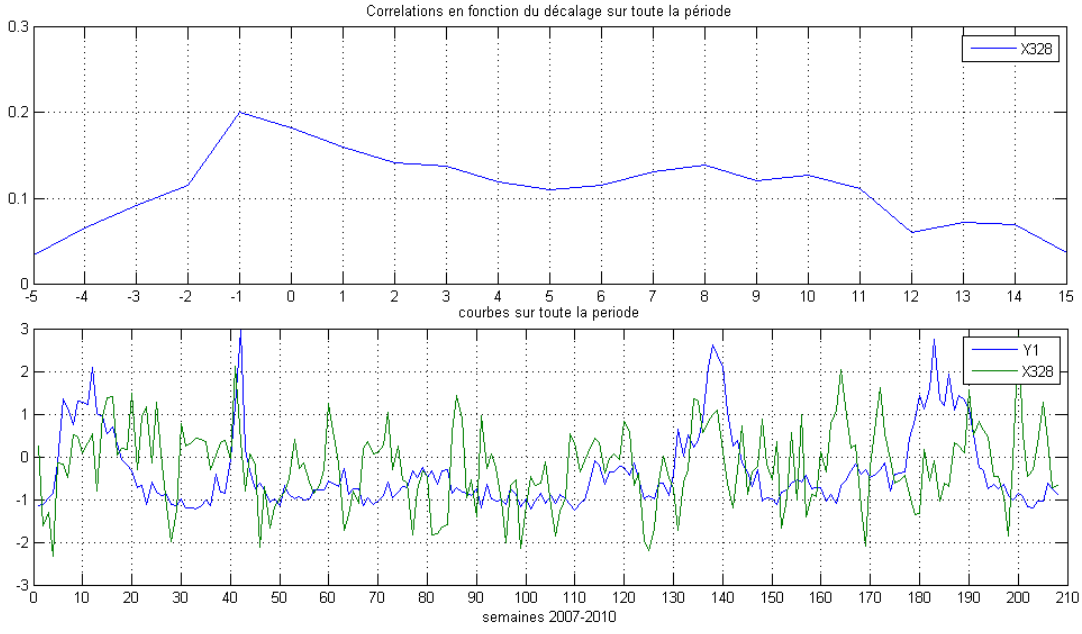

Correlations en fonction du décalage sur toute la période
X323
courbes sur toute la periode
Y1
X323
semaines 2007-2010

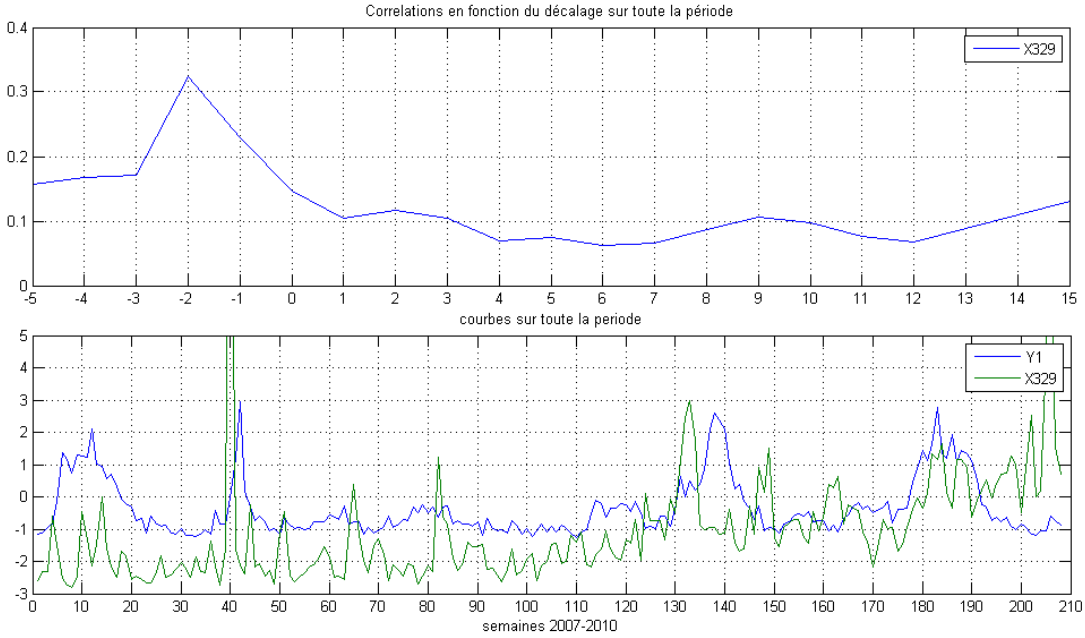

Correlations en fonction du décalage sur toute la période
X324
courbes sur toute la periode
Y1
X324
semaines 2007-2010

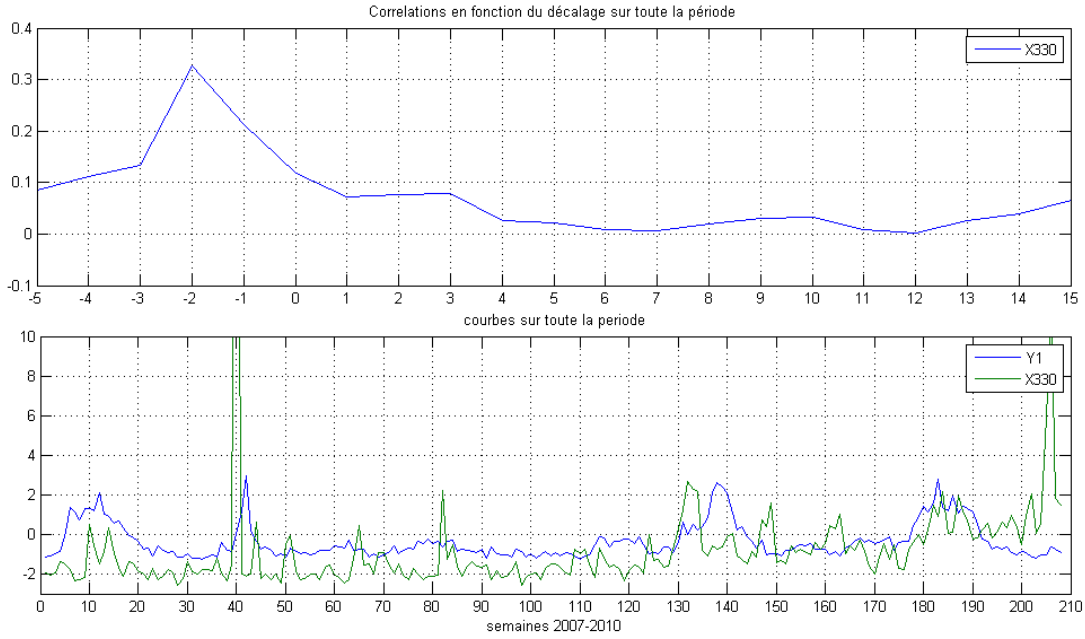

Correlations en fonction du décalage sur toute la période
X325
courbes sur toute la periode
Y1
X325
semaines 2007-2010

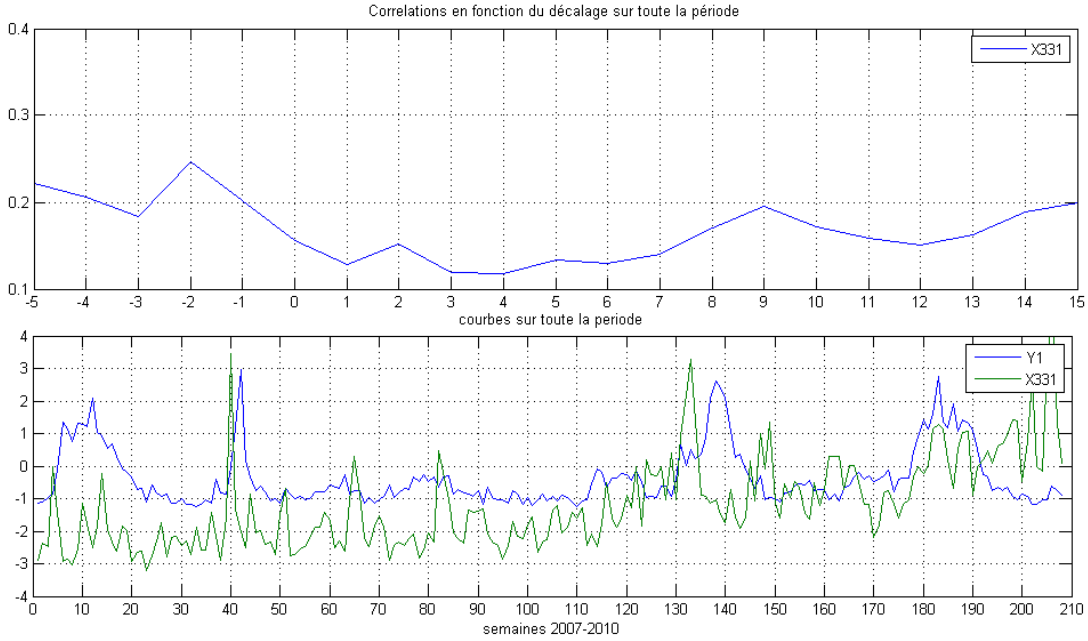

Correlations en fonction du décalage sur toute la période
X326
courbes sur toute la periode
Y1
X326
semaines 2007-2010

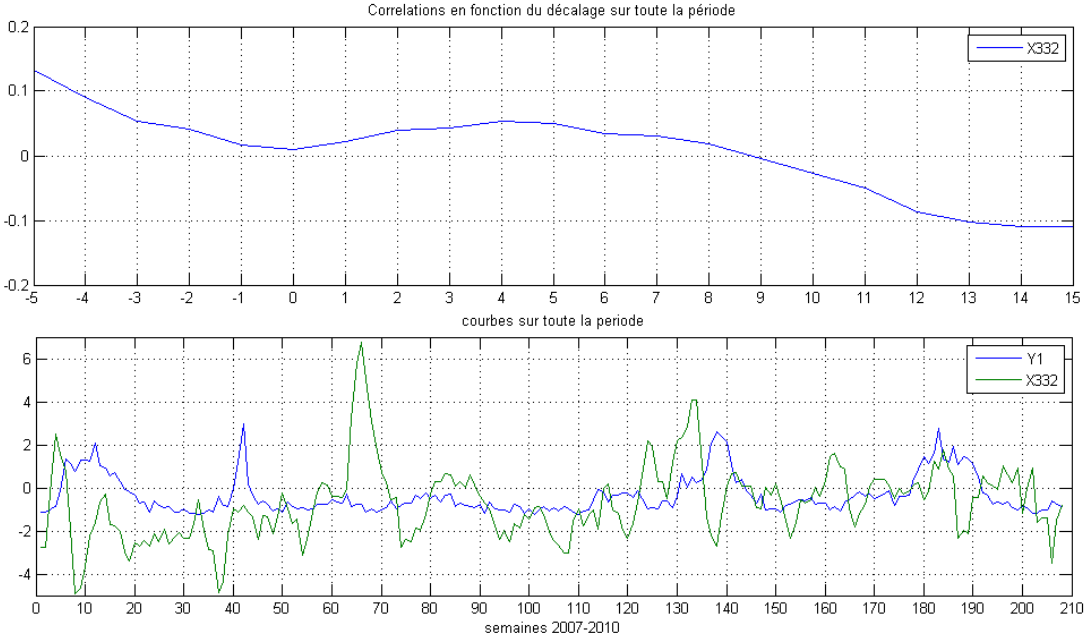

Correlations en fonction du décalage sur toute la période
X327
courbes sur toute la periode
Y1
X327
semaines 2007-2010

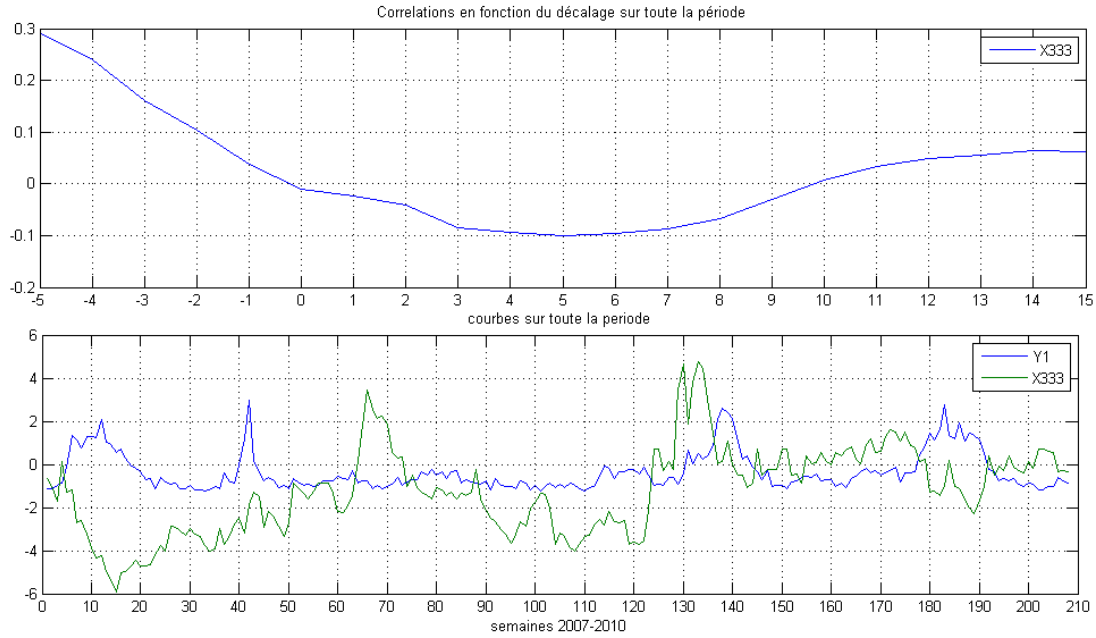

Correlations en fonction du décalage sur toute la période
X328
courbes sur toute la periode
Y1
X328
semaines 2007-2010

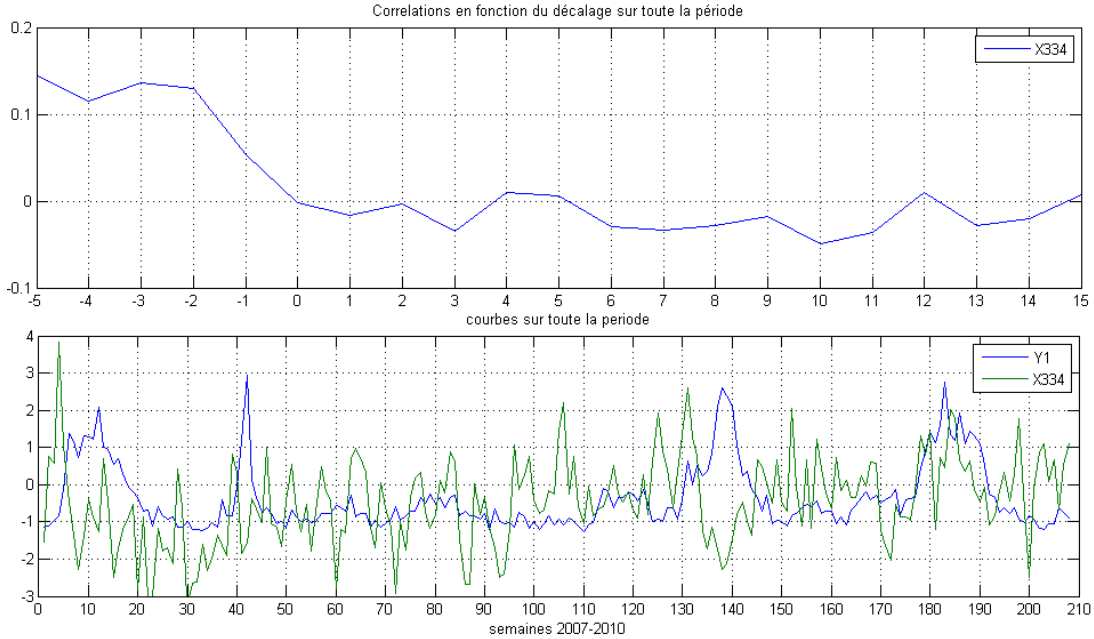

Correlations en fonction du décalage sur toute la période
X329
courbes sur toute la periode
Y1
X329
semaines 2007-2010

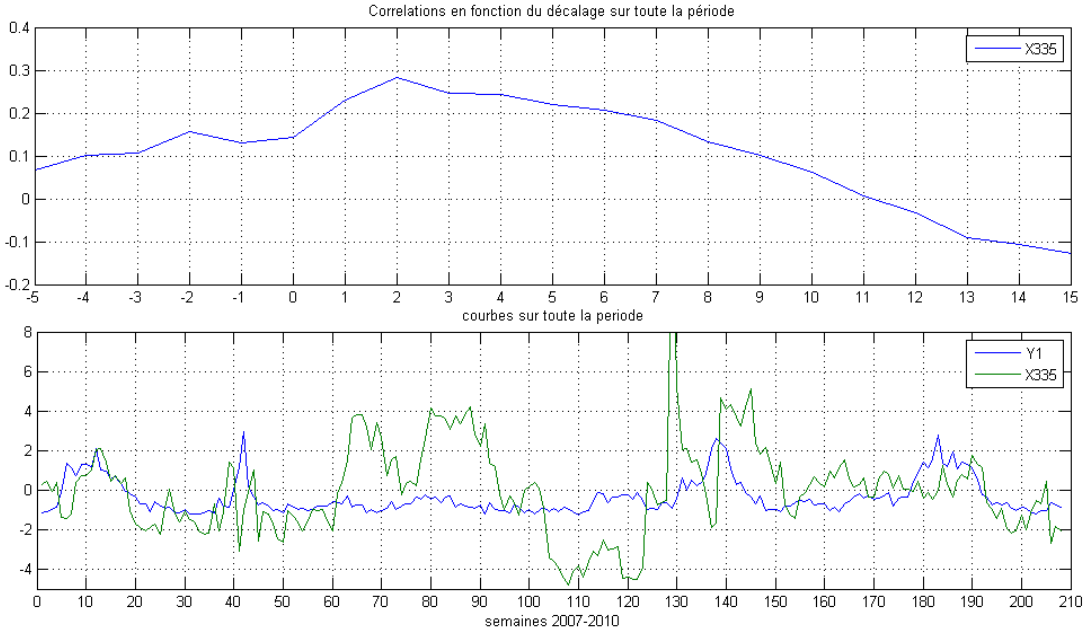

Correlations en fonction du décalage sur toute la période
X330
courbes sur toute la periode
Y1
X330
semaines 2007-2010

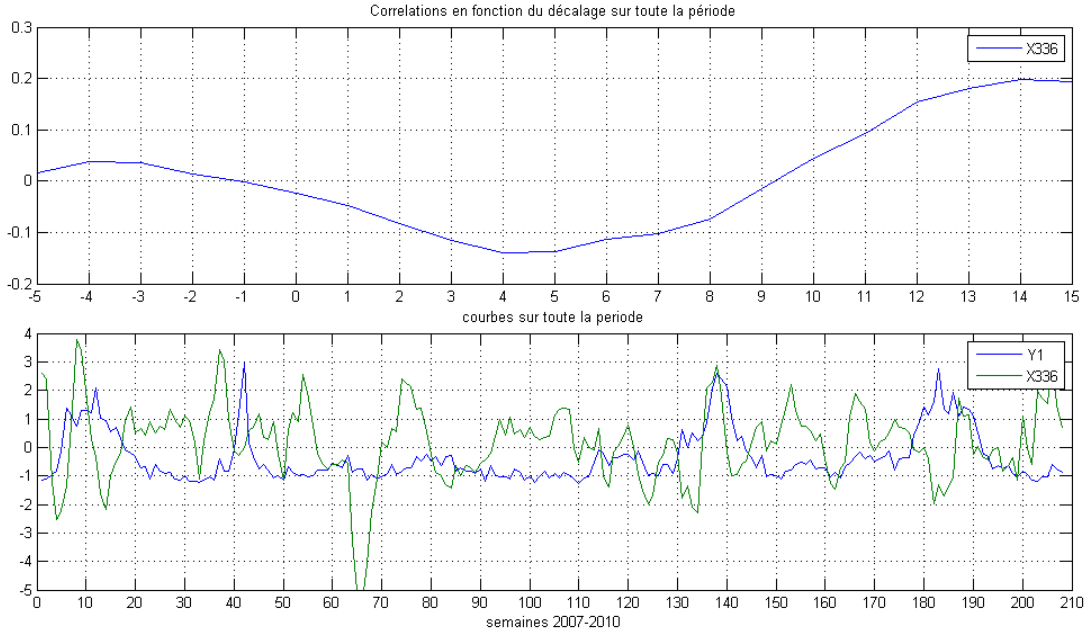

Correlations en fonction du décalage sur toute la période
X331
courbes sur toute la periode
Y1
X331
semaines 2007-2010

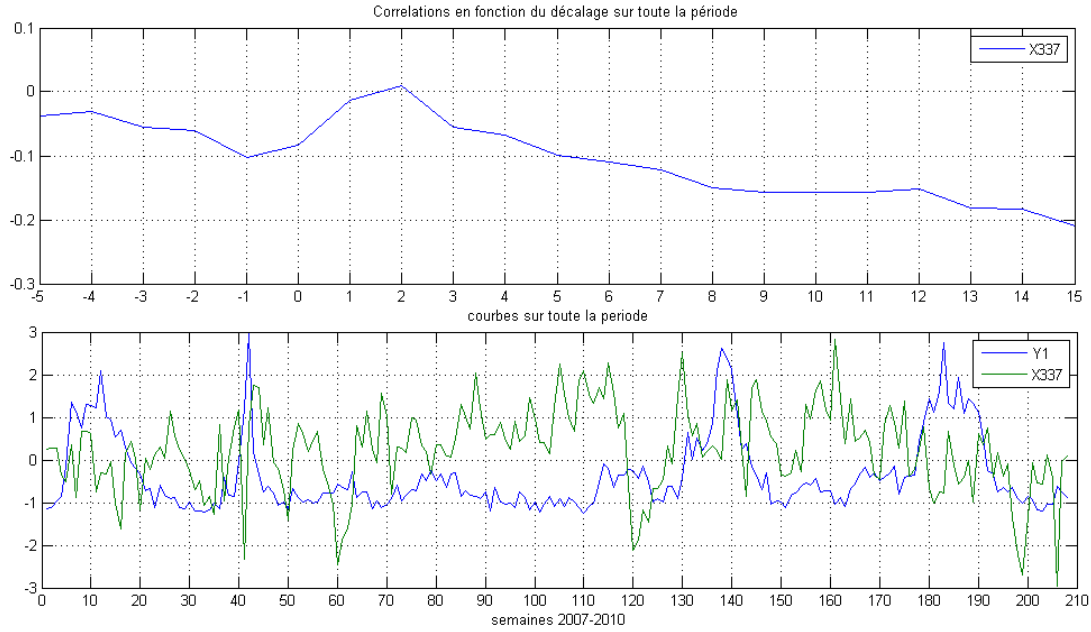

Correlations en fonction du décalage sur toute la période
X332
courbes sur toute la periode
Y1
X332
semaines 2007-2010

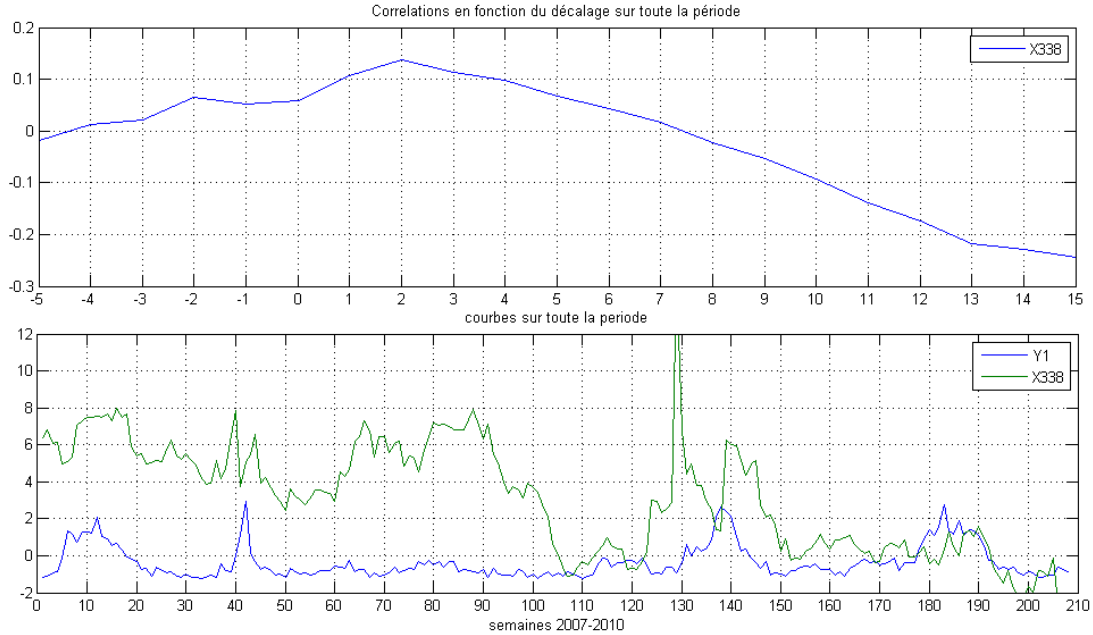

Correlations en fonction du décalage sur toute la période
X333
courbes sur toute la periode
Y1
X333
semaines 2007-2010

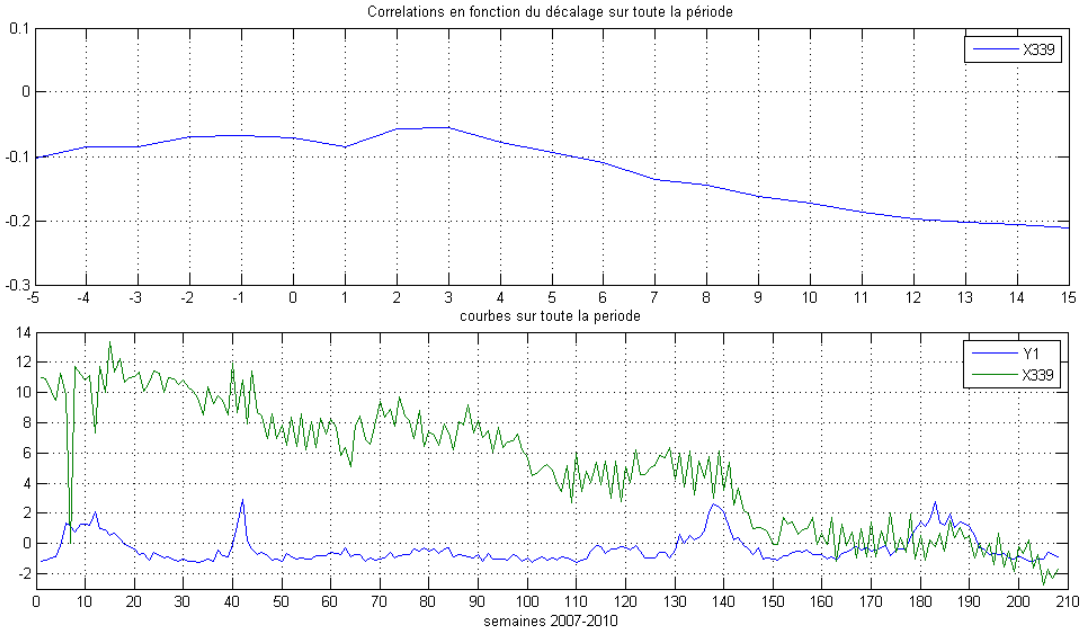

Correlations en fonction du décalage sur toute la période
X334
courbes sur toute la periode
Y1
X334
semaines 2007-2010

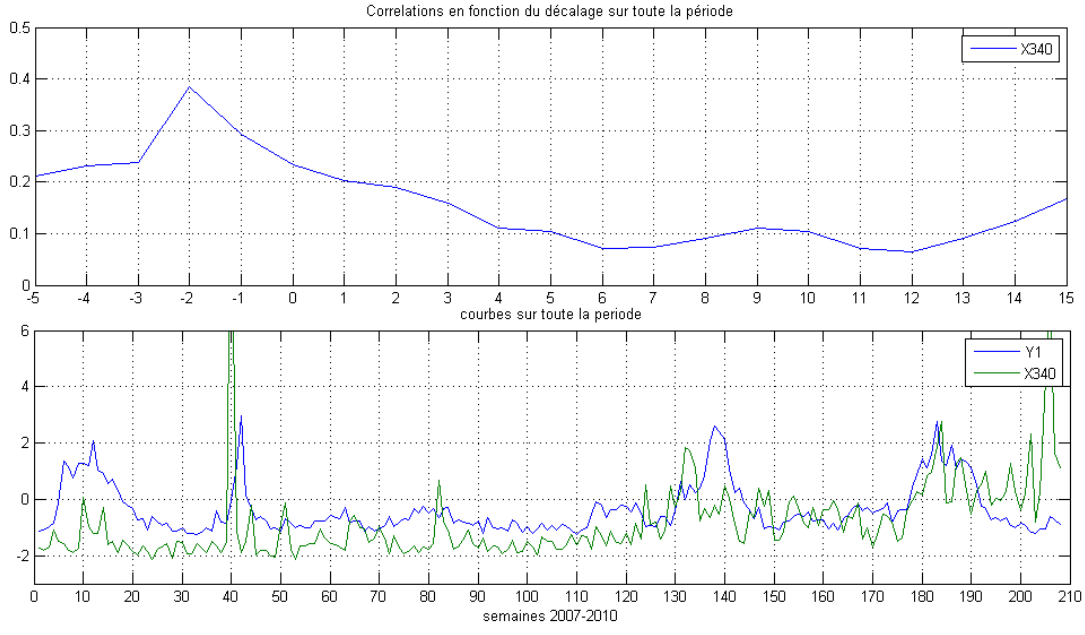

Correlations en fonction du décalage sur toute la période
X335
courbes sur toute la periode
Y1
X335
semaines 2007-2010

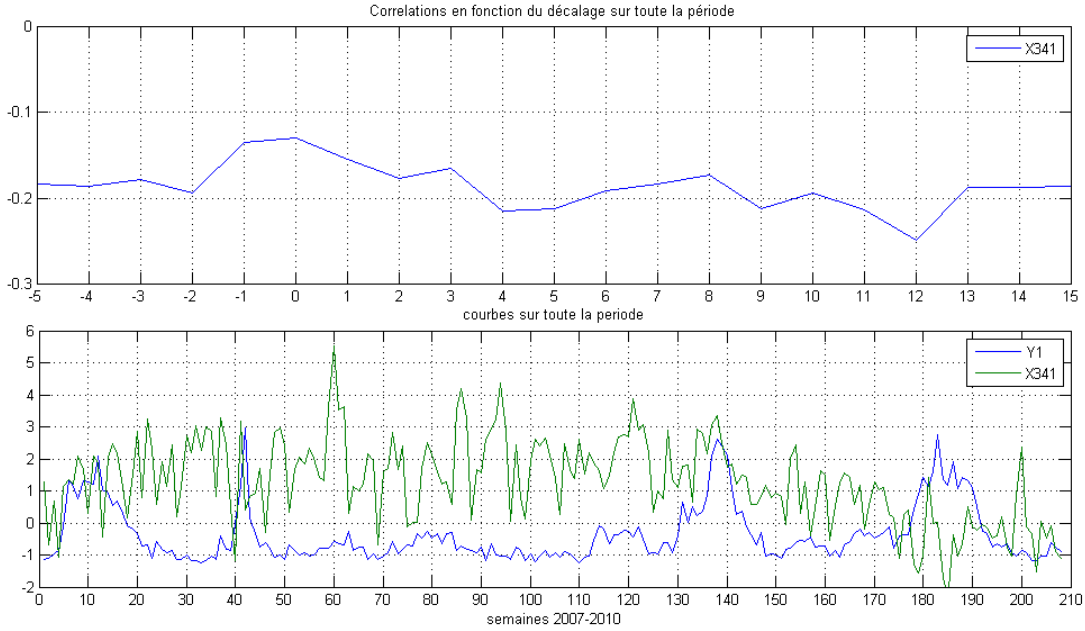

Correlations en fonction du décalage sur toute la période
X336
courbes sur toute la periode
Y1
X336
semaines 2007-2010

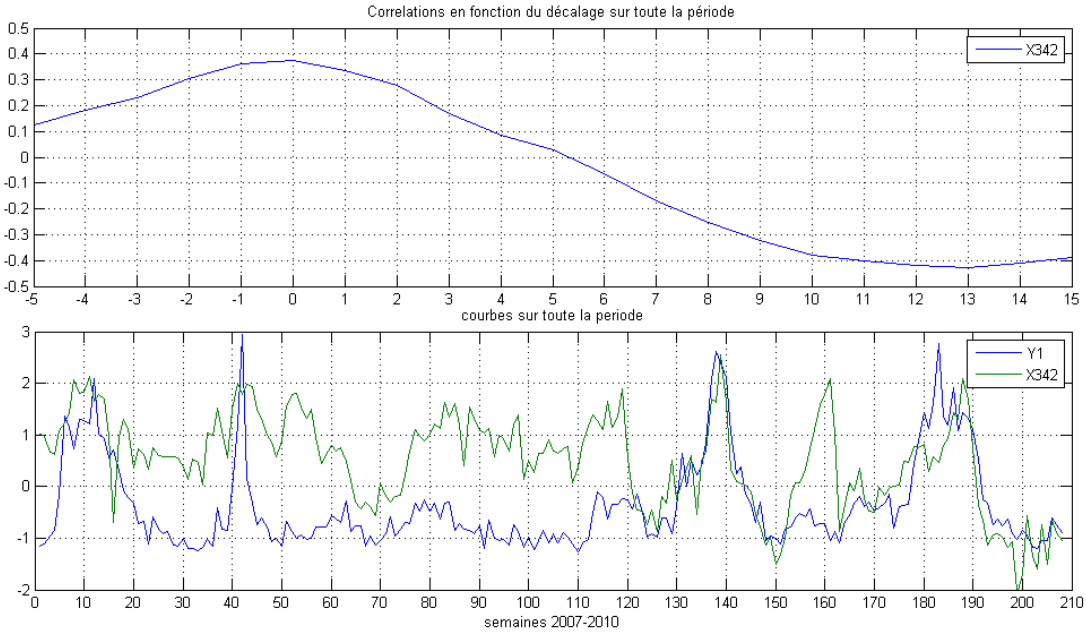

Correlations en fonction du décalage sur toute la période
X337
courbes sur toute la periode
Y1
X337
semaines 2007-2010

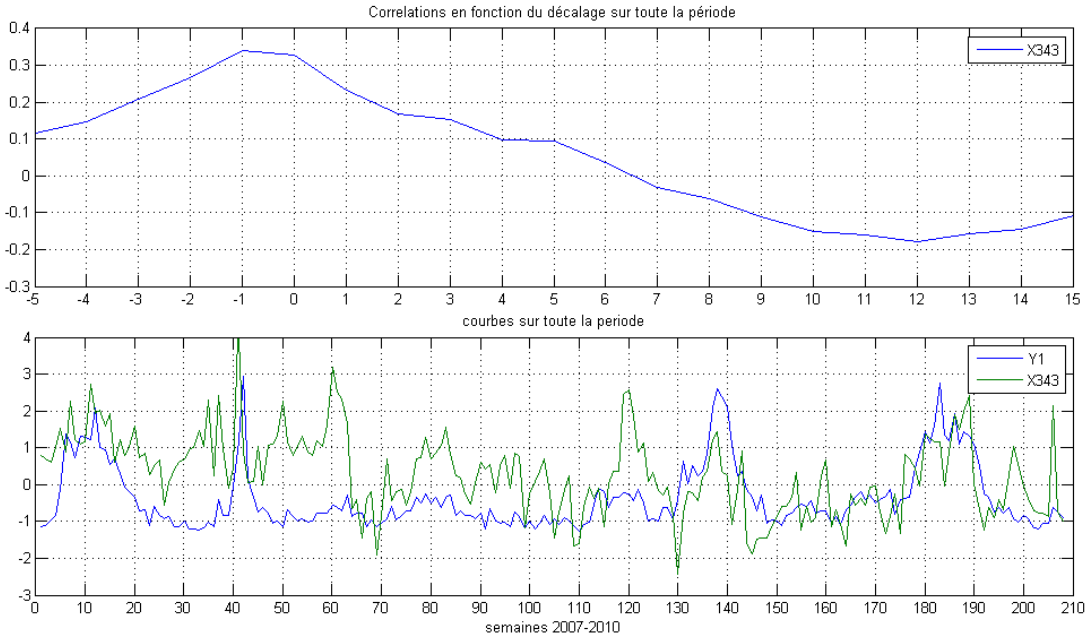

Correlations en fonction du décalage sur toute la période
X338
courbes sur toute la periode
Y1
X338
semaines 2007-2010

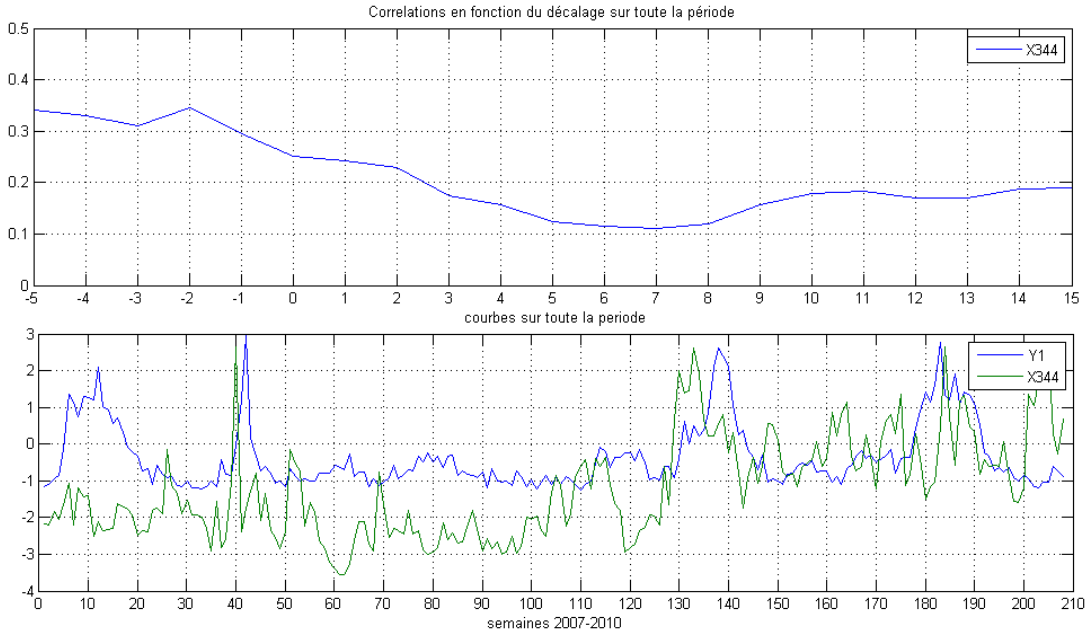

Correlations en fonction du décalage sur toute la période
X339
courbes sur toute la periode
Y1
X339
semaines 2007-2010

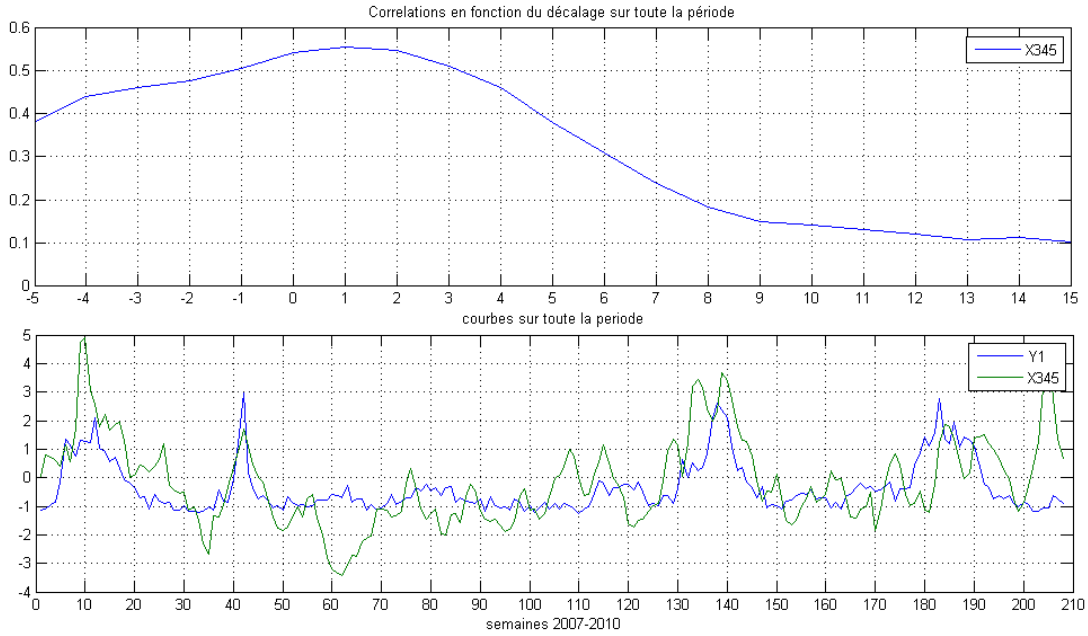

Correlations en fonction du décalage sur toute la période
X340
courbes sur toute la periode
Y1
X340
semaines 2007-2010

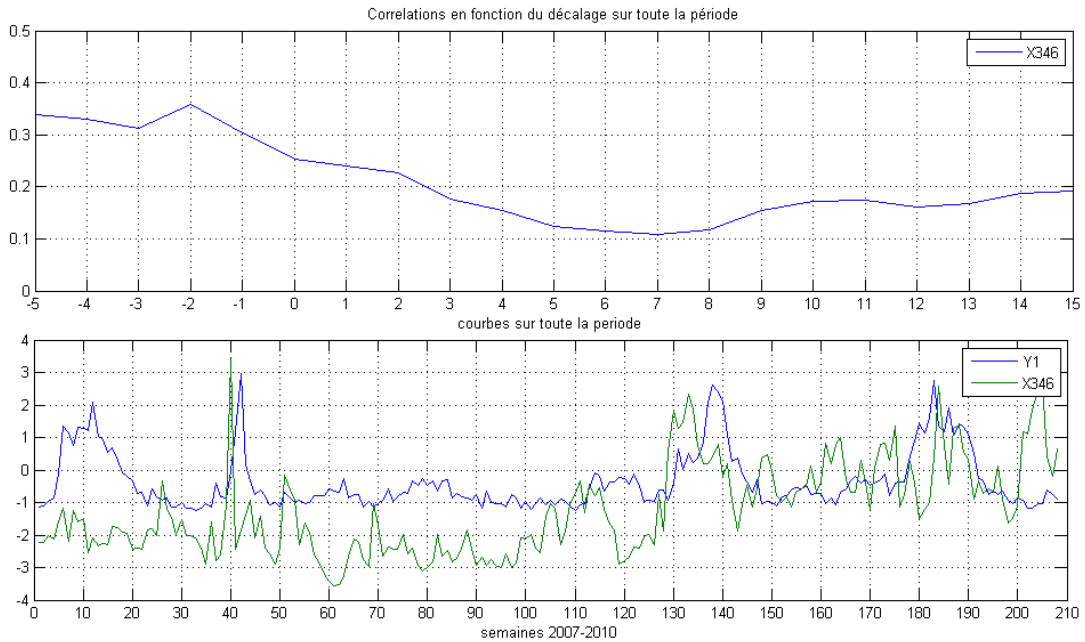

Correlations en fonction du décalage sur toute la période
X341
courbes sur toute la periode
Y1
X341
semaines 2007-2010

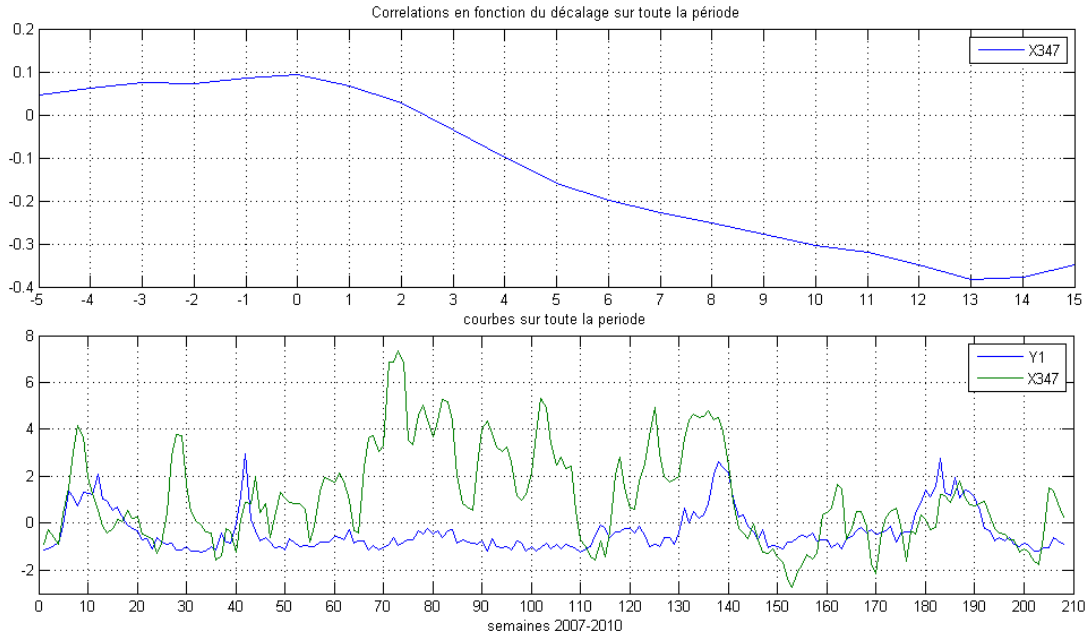

Correlations en fonction du décalage sur toute la période
X342
courbes sur toute la periode
Y1
X342
semaines 2007-2010

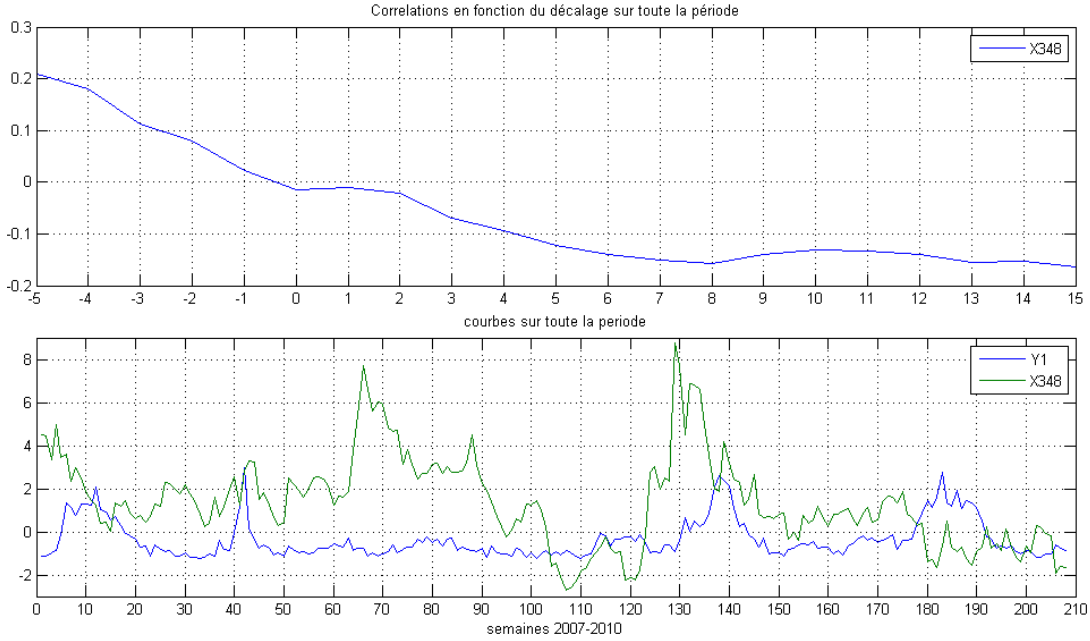

Correlations en fonction du décalage sur toute la période
X343
courbes sur toute la periode
Y1
X343
semaines 2007-2010

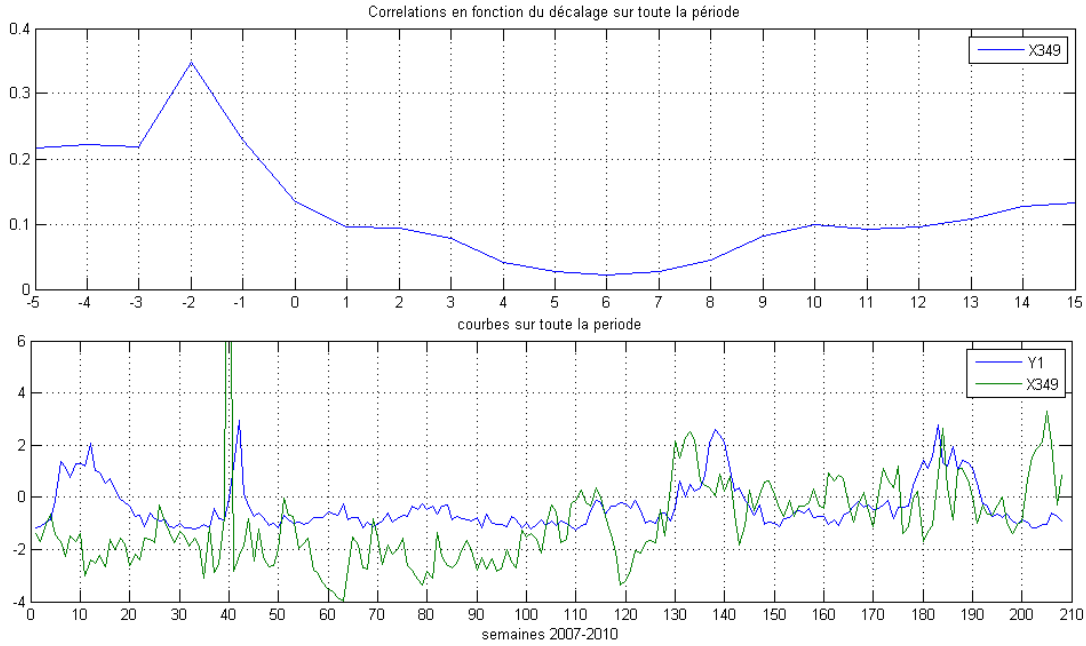

Correlations en fonction du décalage sur toute la période
X344
courbes sur toute la periode
Y1
X344
semaines 2007-2010


Correlations en fonction du décalage sur toute la période
X345
courbes sur toute la periode
Y1
X345
semaines 2007-2010


Correlations en fonction du décalage sur toute la période
X346
courbes sur toute la periode
Y1
X346
semaines 2007-2010


Correlations en fonction du décalage sur toute la période
X347
courbes sur toute la periode
Y1
X347
semaines 2007-2010


Correlations en fonction du décalage sur toute la période
X348
courbes sur toute la periode
Y1
X348
semaines 2007-2010


Correlations en fonction du décalage sur toute la période
X349
courbes sur toute la periode
Y1
X349
semaines 2007-2010

B Annexe sur l'ACP

Cette annexe comprend toutes les courbes relatives à l'ACP.

B.1 Expérience 1 : ACP sur toutes les variables ainsi que toutes les 43 observations labellisées en 4 périodes.

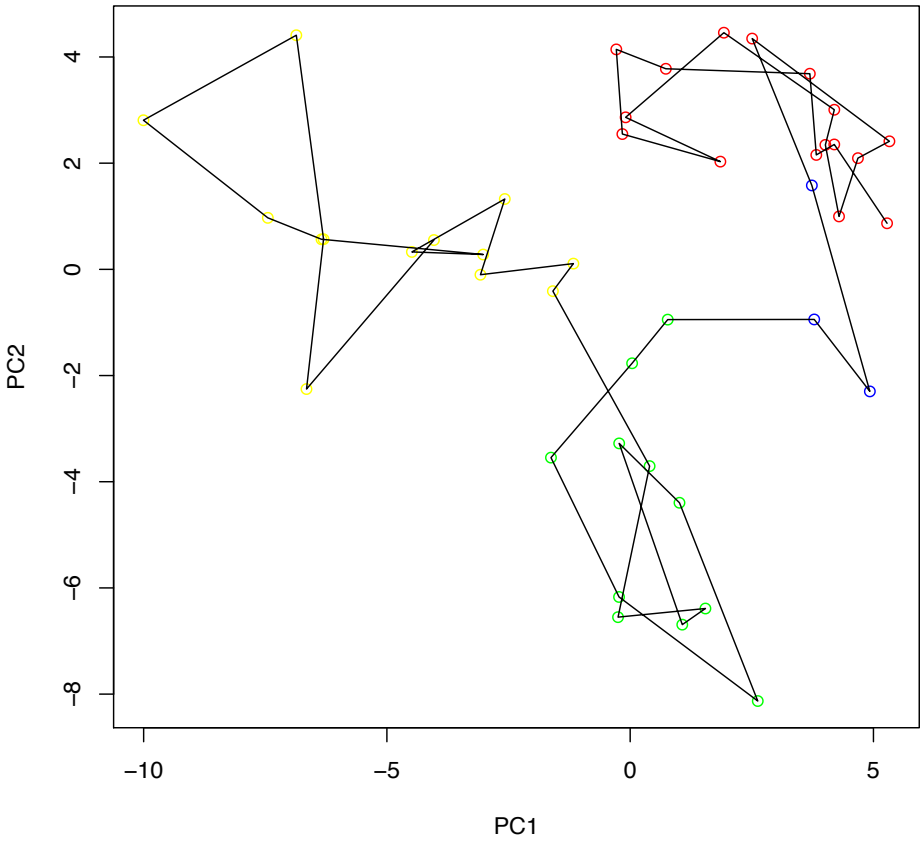


FIGURE 29 – Expérience 1 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

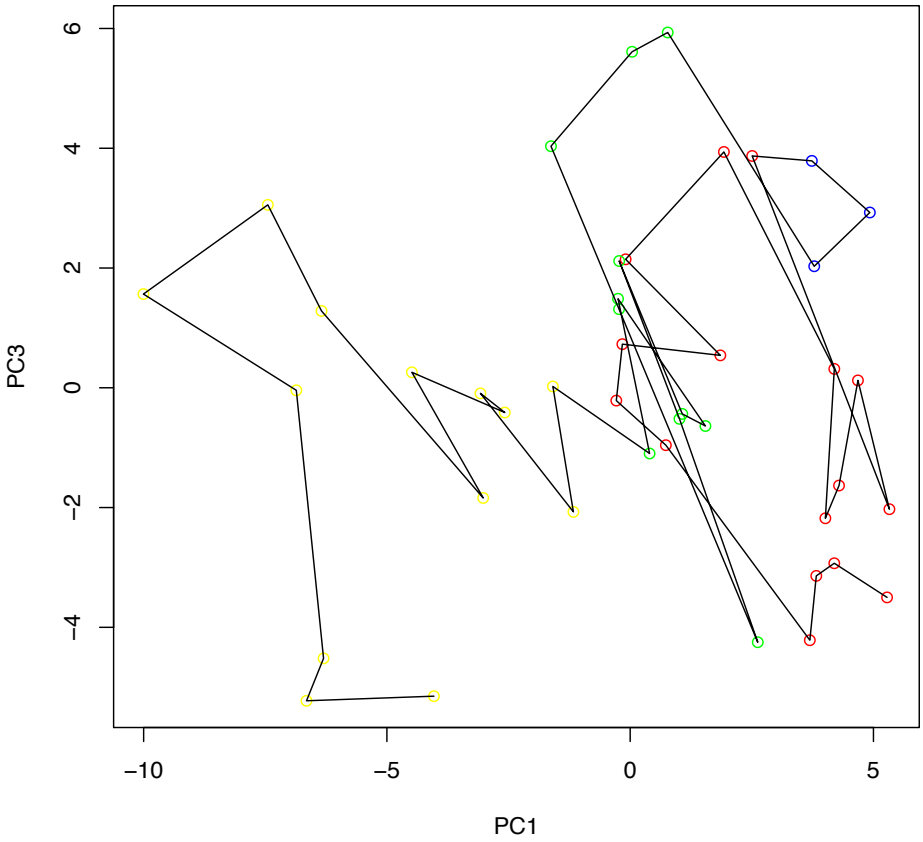


FIGURE 30 – Expérience 1 : Projection des individus sur le plan factoriel PC1× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

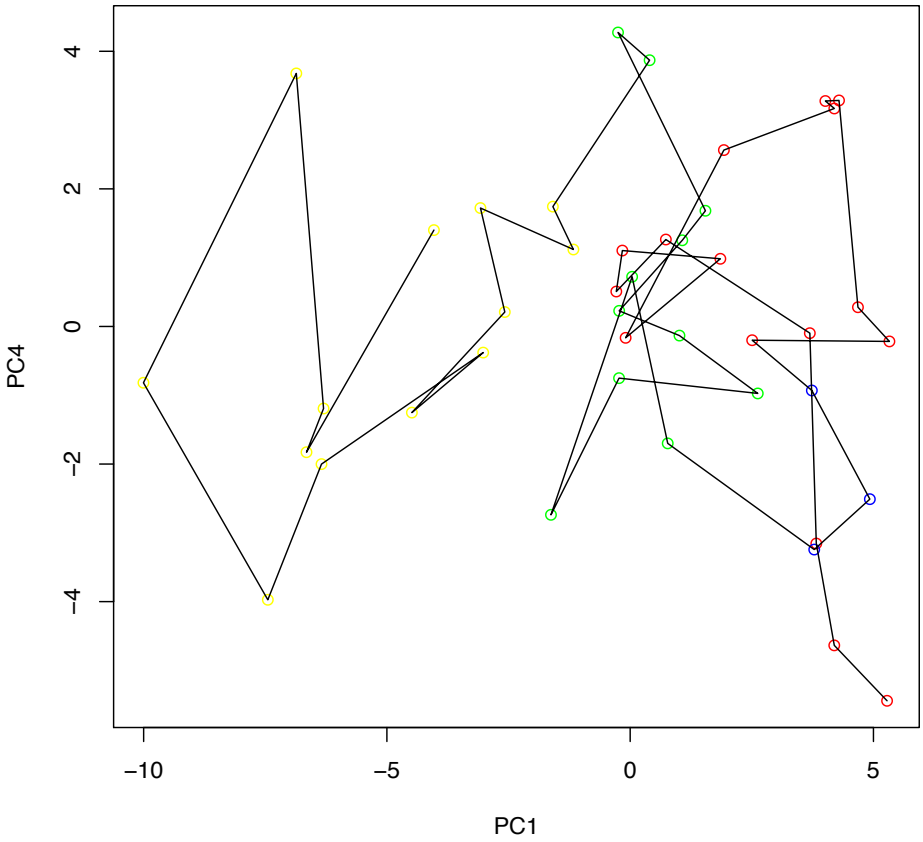


FIGURE 31 – Expérience 1 : Projection des individus sur le plan factoriel PC1× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

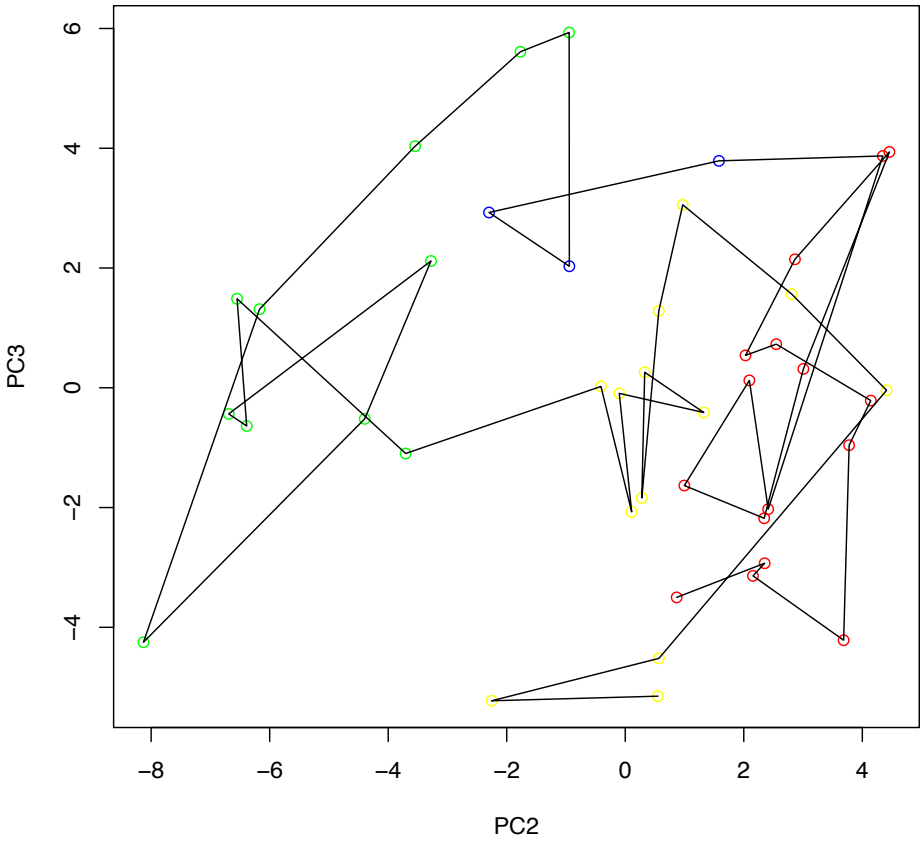


FIGURE 32 – Expérience 1 : Projection des individus sur le plan factoriel PC2× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

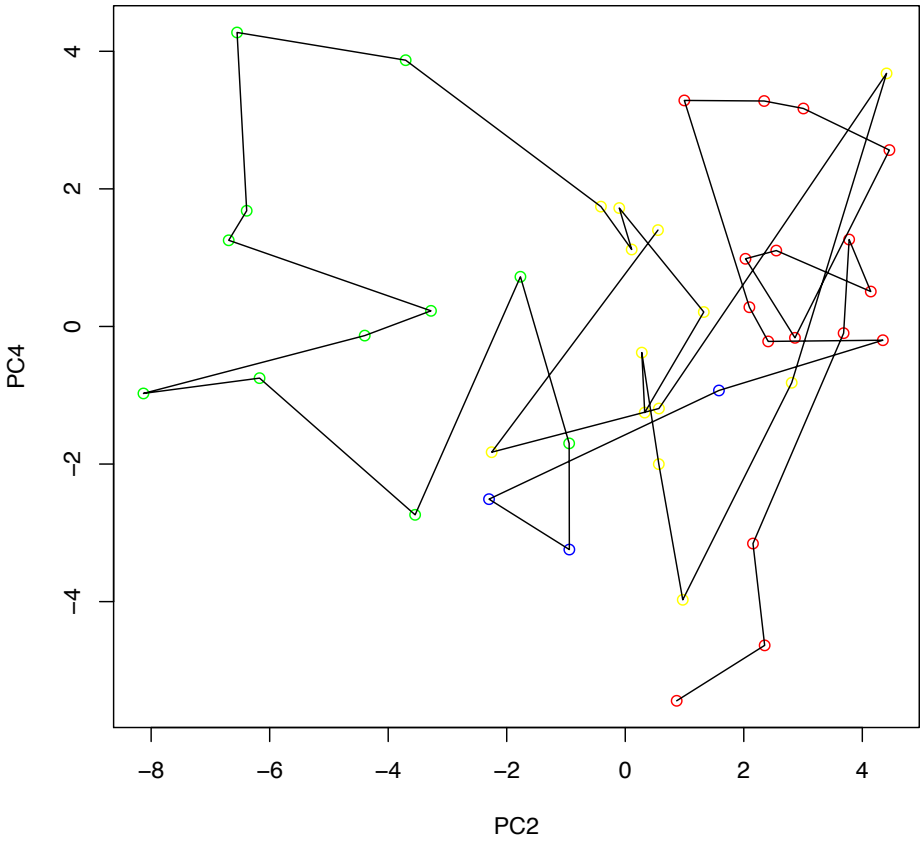


FIGURE 33 – Expérience 1 : Projection des individus sur le plan factoriel PC2× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

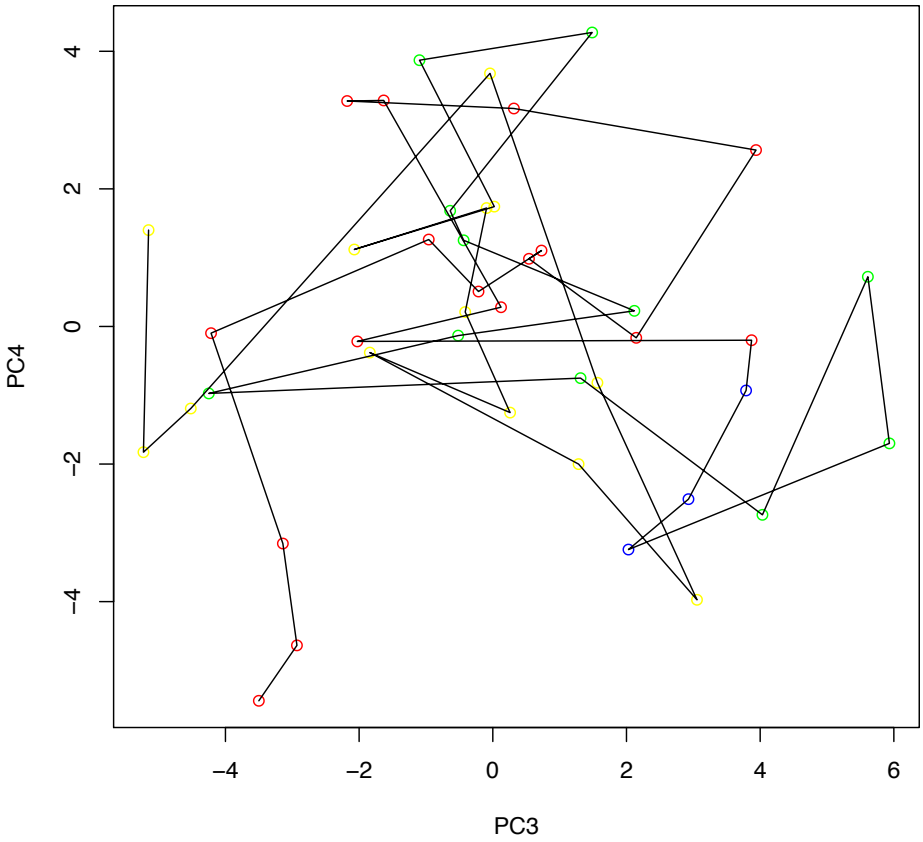


FIGURE 34 – Expérience 1 : Projection des individus sur le plan factoriel PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

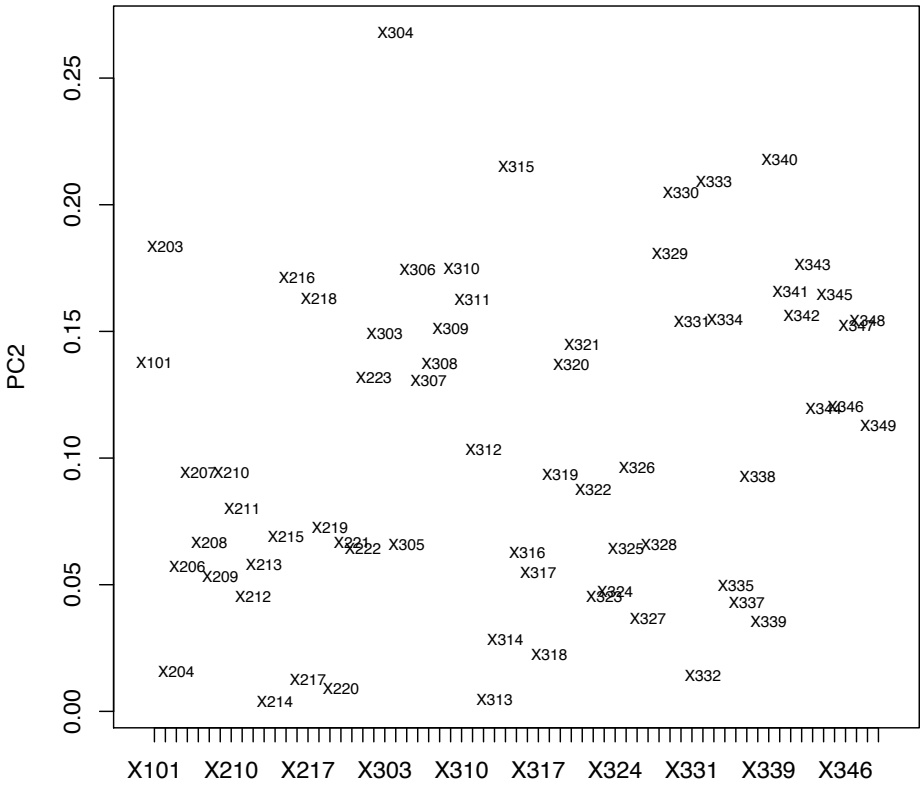


FIGURE 35 – Expérience 1 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2.

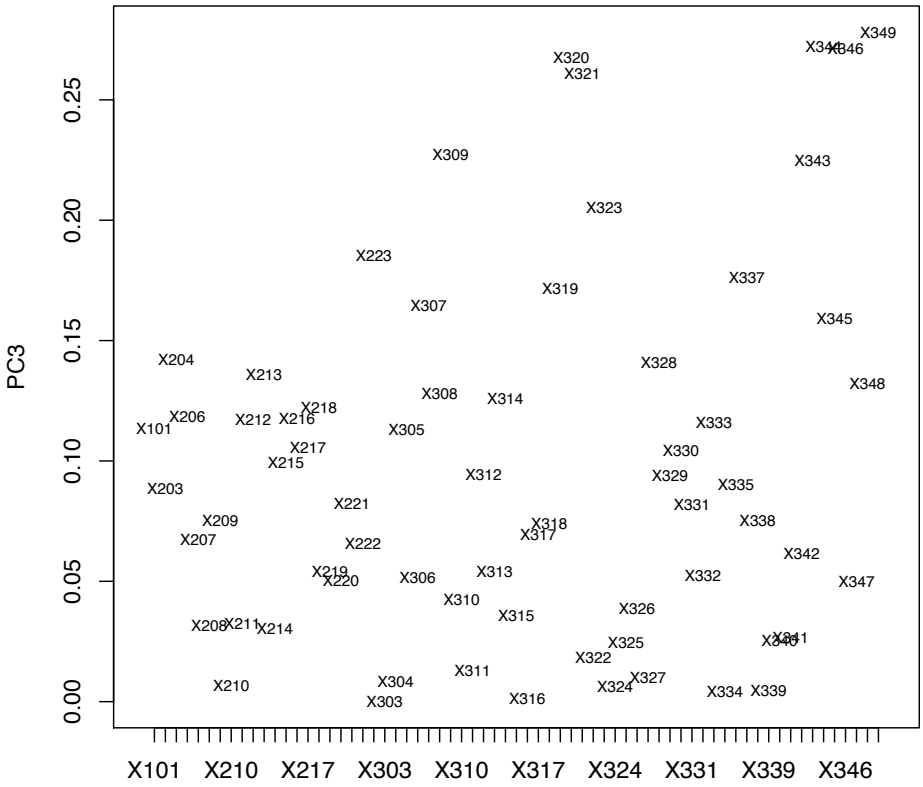


FIGURE 36 – Expérience 1 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3.

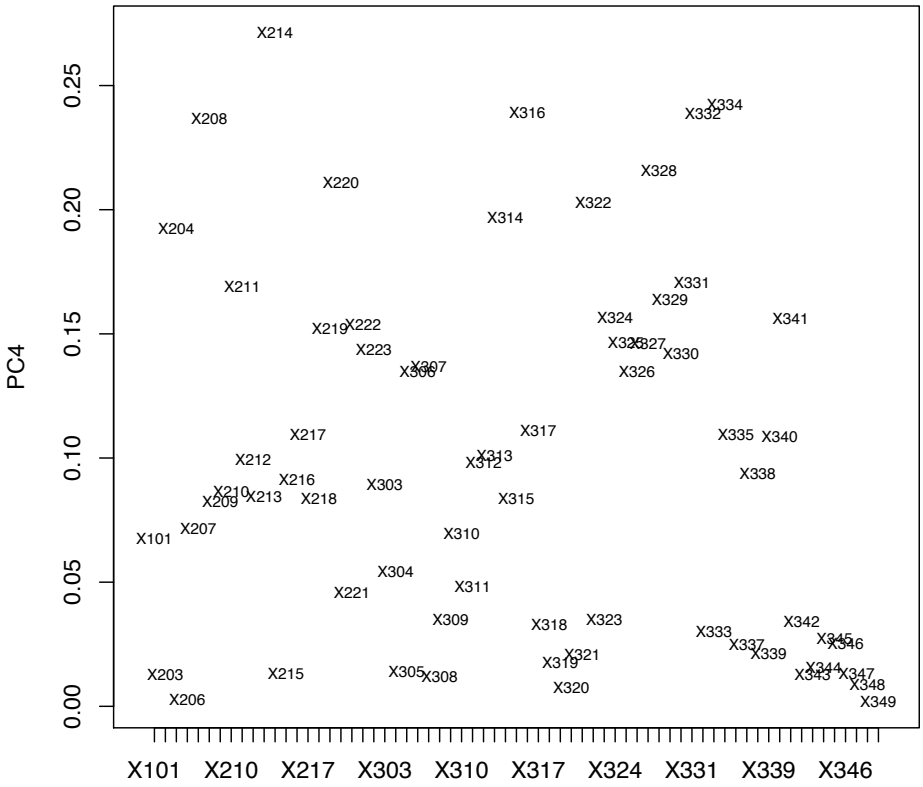


FIGURE 37 – Expérience 1 : Projection des variables sur l'axe factoriel PC4 montrant leur poids en valeur absolue dans la composante principale PC4.

B.2 Expérience 2 : ACP sur 12 variables, X_{304}, X_{340}, X_{315}, X_{333}, X_{330}, X_{203}, X_{329} et X_{343}, X_{217}, X_{212}, X_{215} et X_{210}, sélectionnées dans l'expérience 1 ainsi que toutes les 43 observations labellisées en 4 périodes.

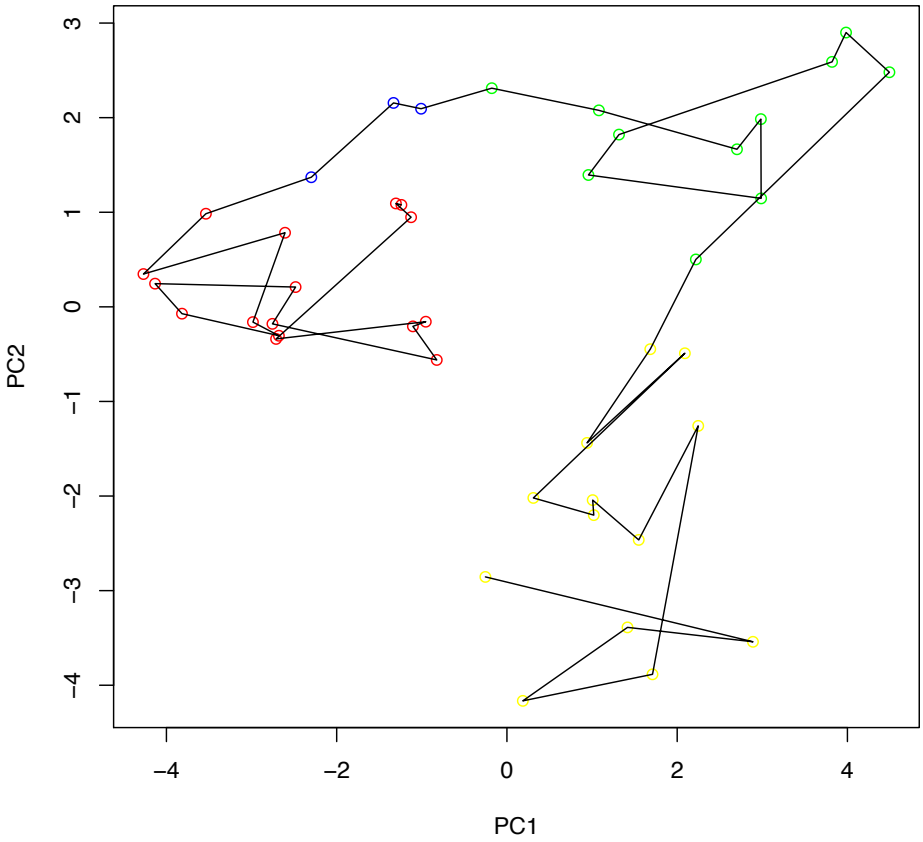


FIGURE 38 – Expérience 2 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps. L'axe PC1 permet de mettre en évidence le passage d'une période stable à une période avec des fluctuations.

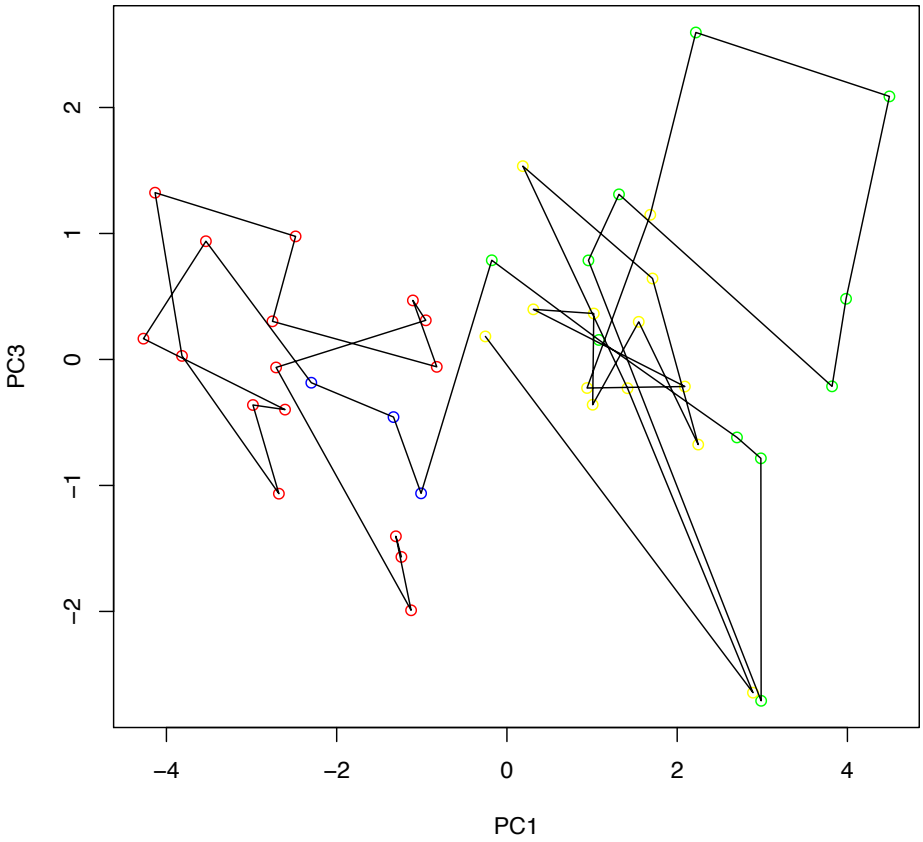


FIGURE 39 – Expérience 2 : Projection des individus sur le plan factoriel PC1× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

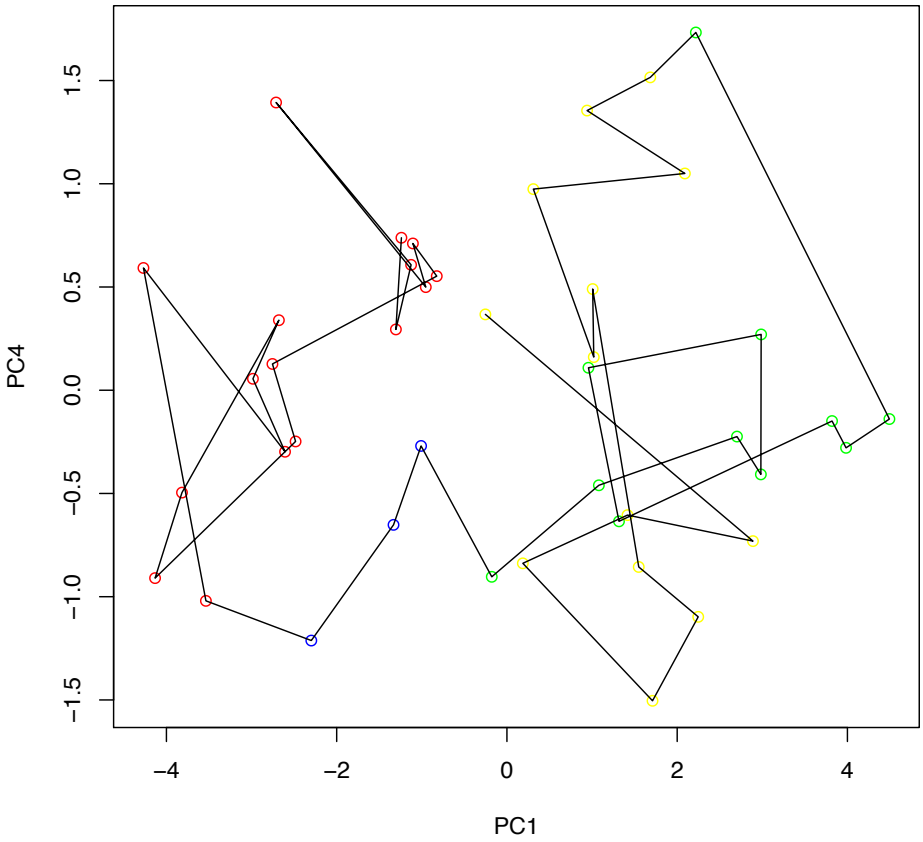


FIGURE 40 – Expérience 2 : Projection des individus sur le plan factoriel PC1× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

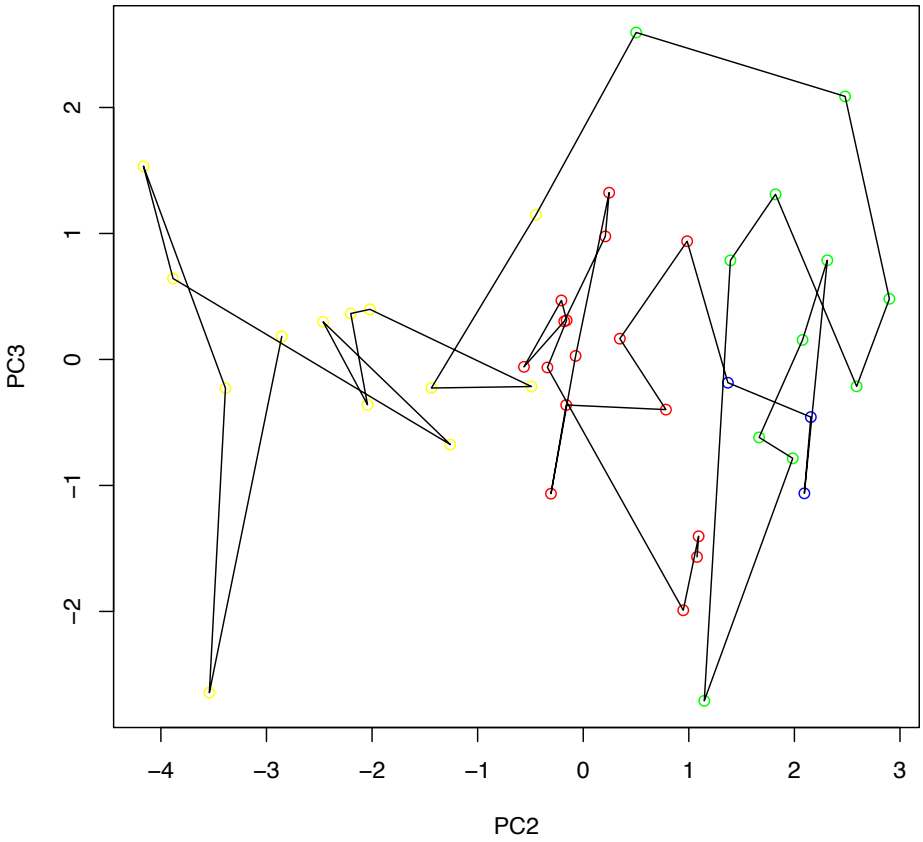


FIGURE 41 – Expérience 2 : Projection des individus sur le plan factoriel PC2× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

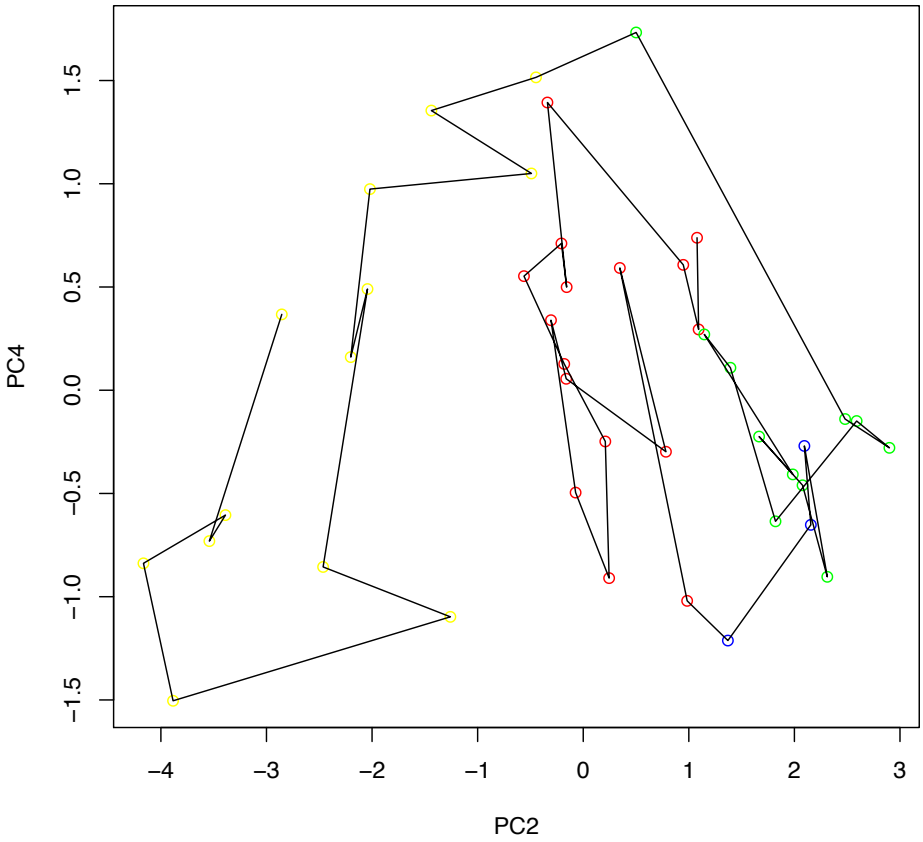


FIGURE 42 – Expérience 2 : Projection des individus sur le plan factoriel PC2× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

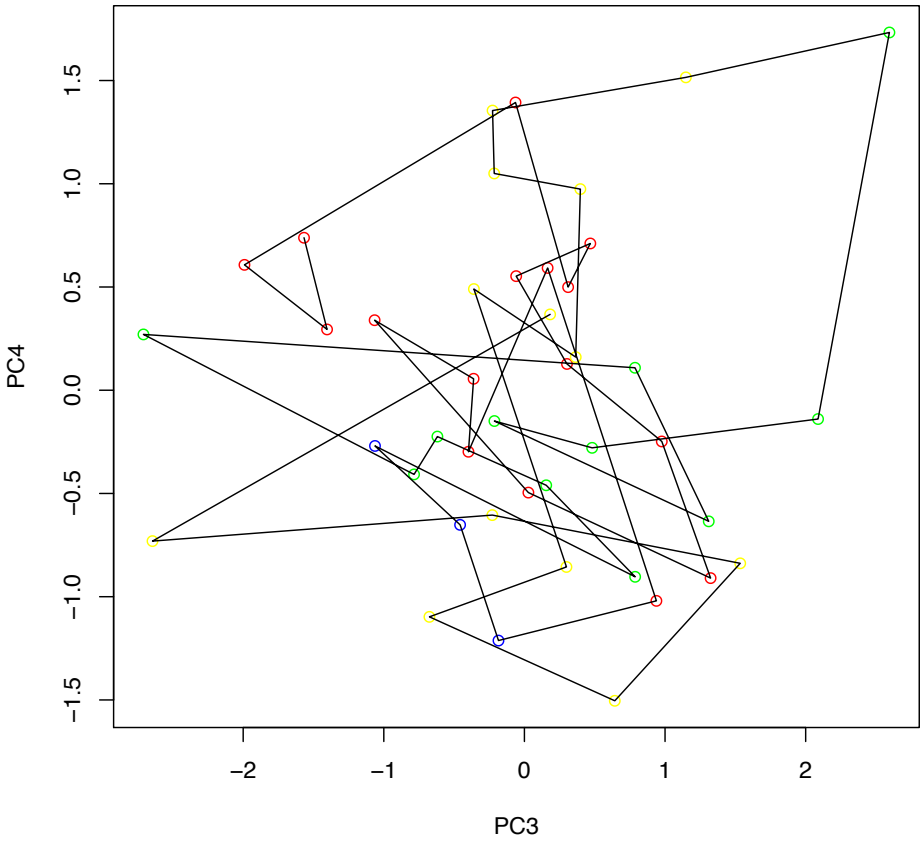


FIGURE 43 – Expérience 2 : Projection des individus sur le plan factoriel PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

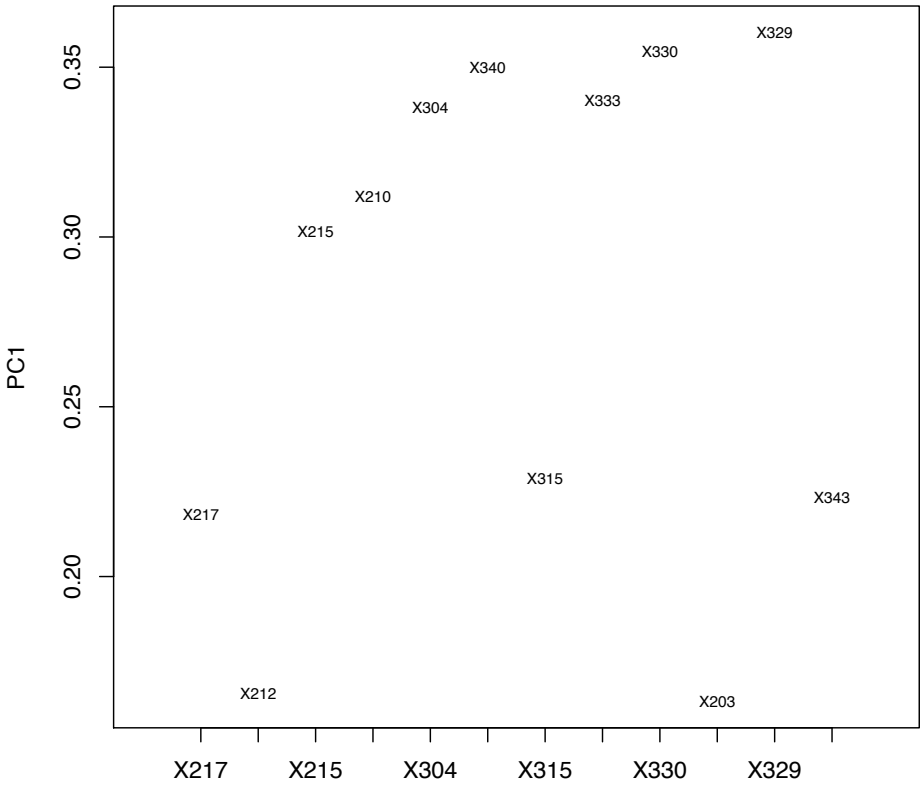


FIGURE 44 – Expérience 2 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1.

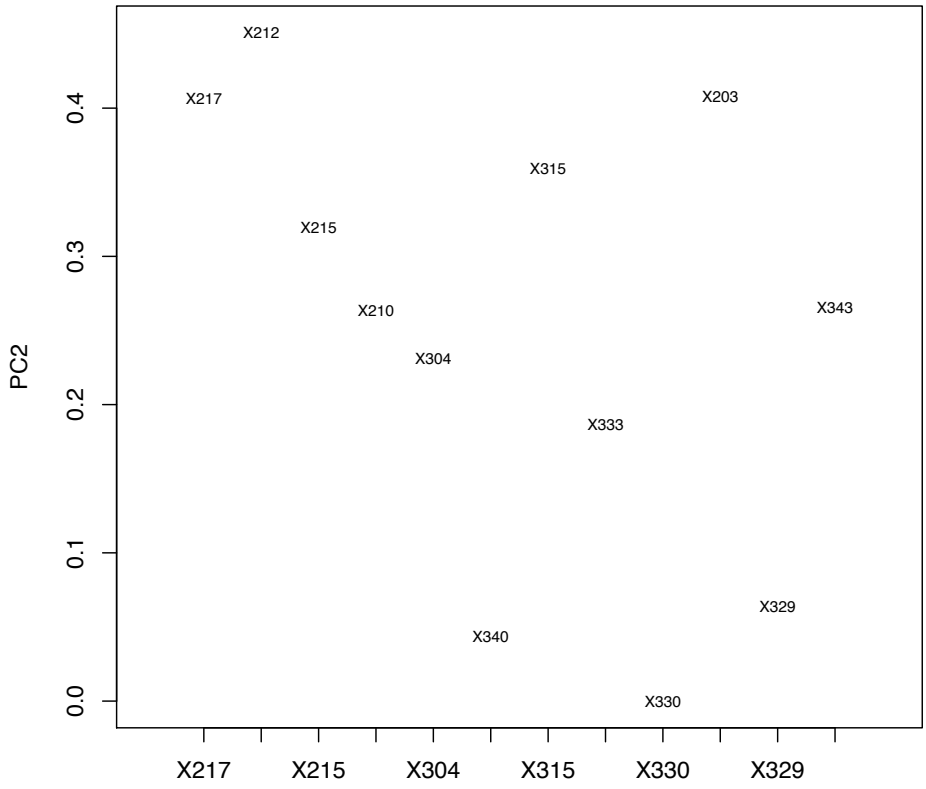


FIGURE 45 – Expérience 2 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2.

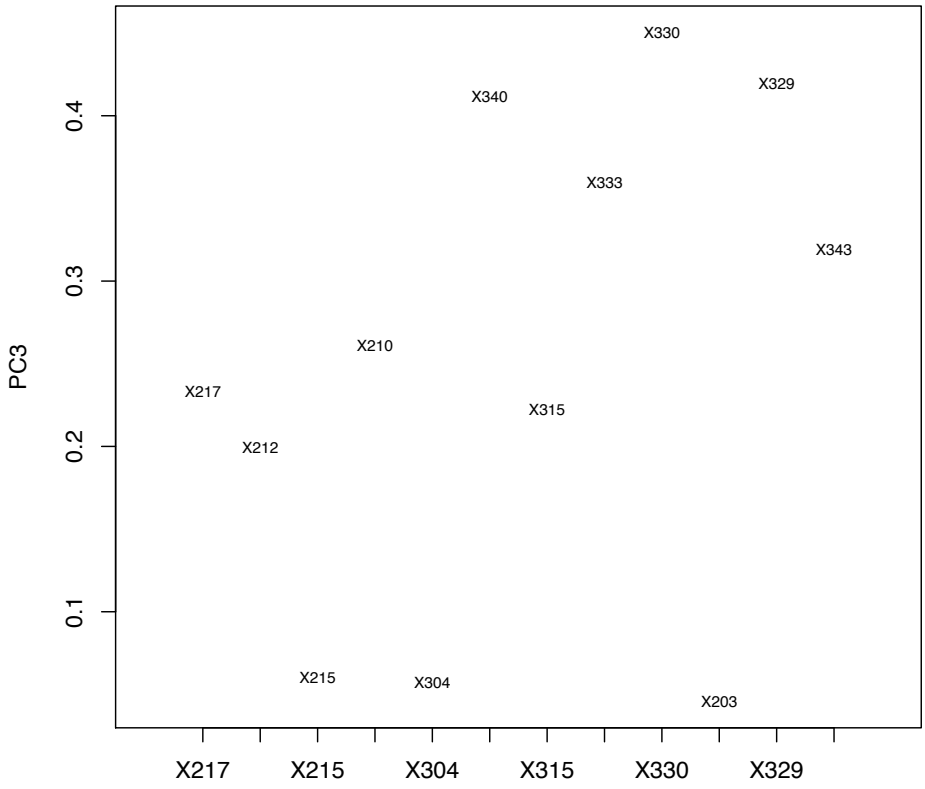


FIGURE 46 – Expérience 2 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3.

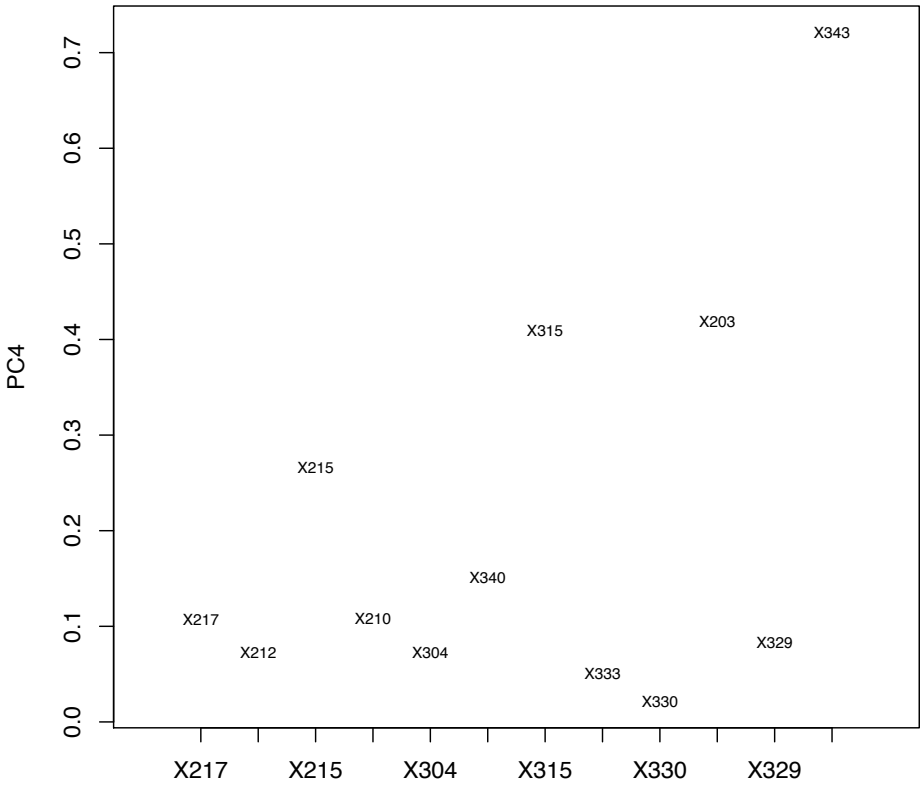


FIGURE 47 – Expérience 2 : Projection des variables sur l'axe factoriel PC4 montrant leur poids en valeur absolue dans la composante principale PC4.

B.3 Expérience 3 : 19 premières semaines qui correspondant aux semaines avant la période de fluctuations en considérant toutes les 43 variables.

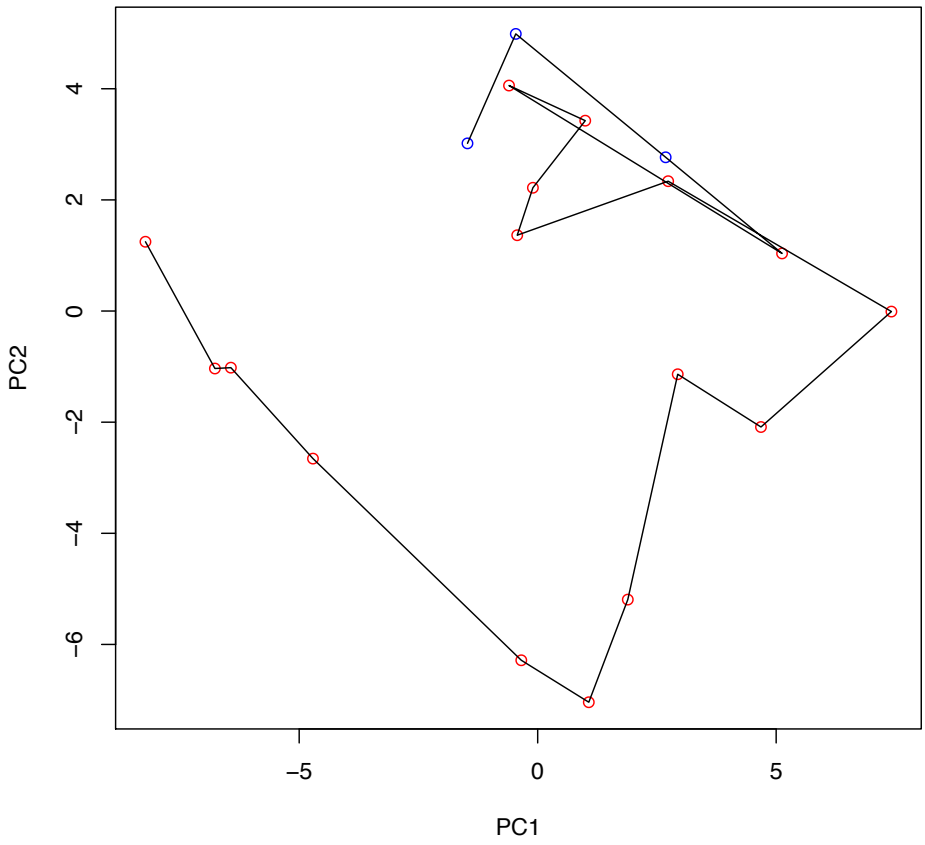


FIGURE 48 – Expérience 3 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

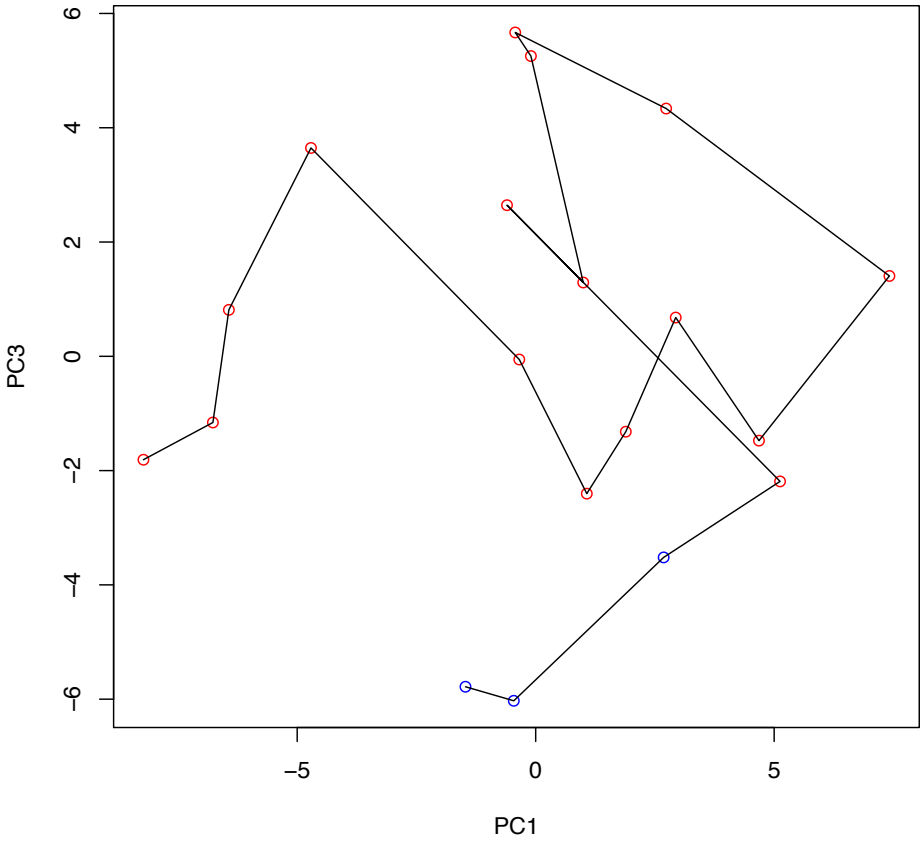


FIGURE 49 – Expérience 3 : Projection des individus sur le plan factoriel PC1× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

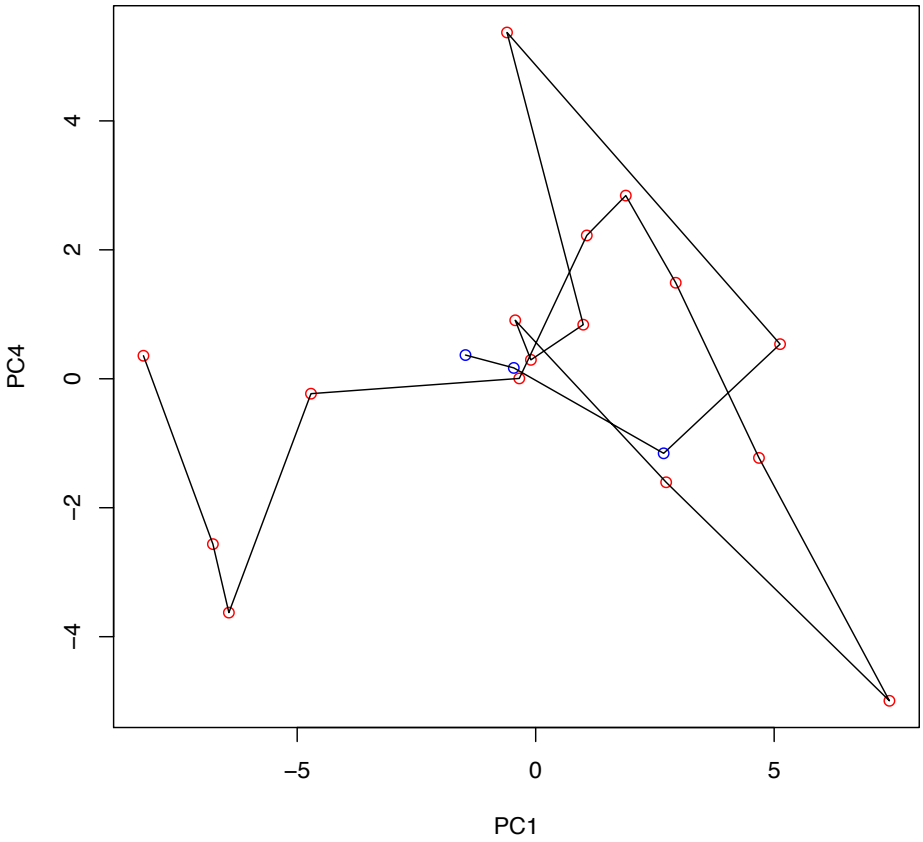


FIGURE 50 – Expérience 3 : Projection des individus sur le plan factoriel PC1× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

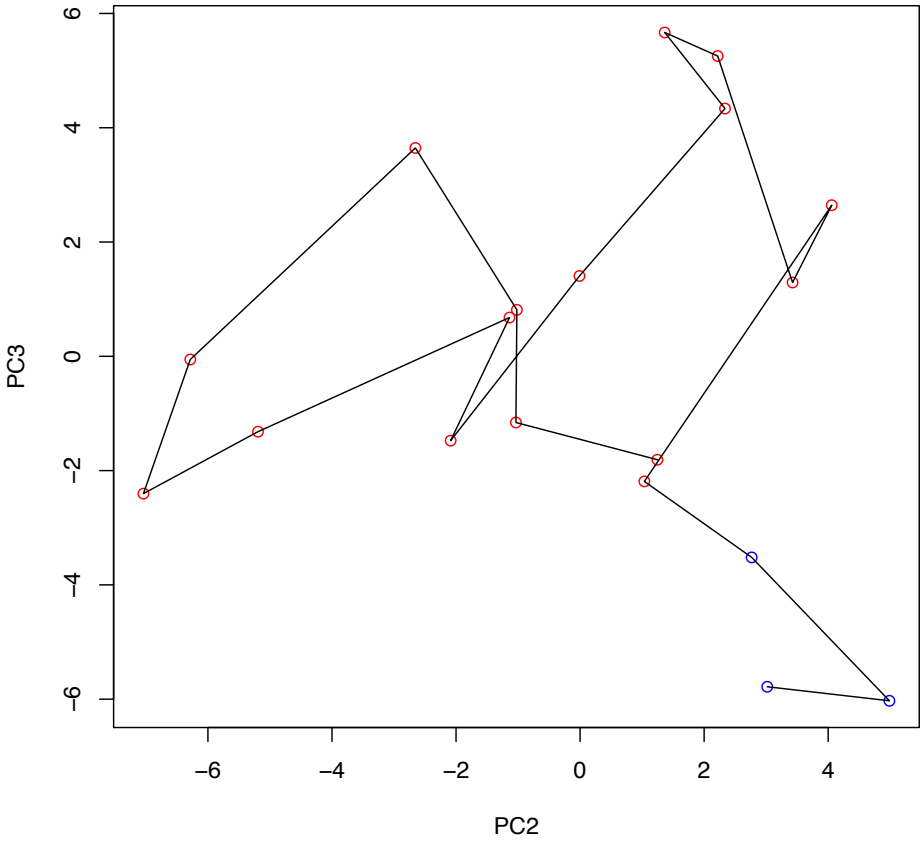


FIGURE 51 – Expérience 3 : Projection des individus sur le plan factoriel PC2× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

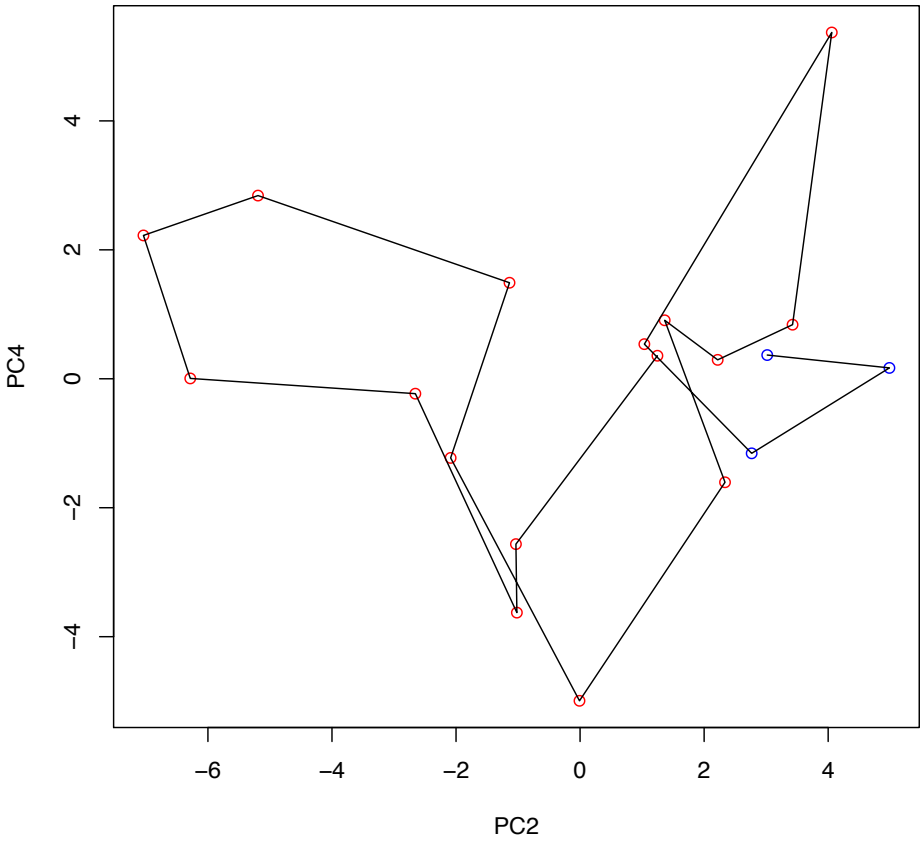


FIGURE 52 – Expérience 3 : Projection des individus sur le plan factoriel PC2× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

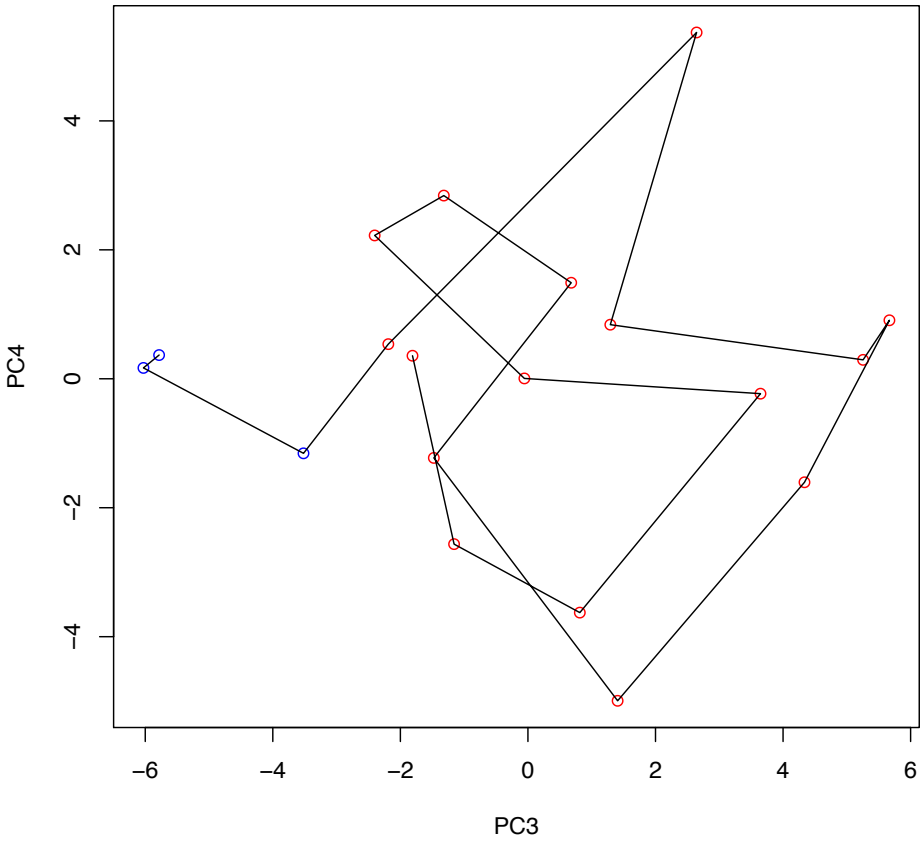


FIGURE 53 – Expérience 3 : Projection des individus sur le plan factoriel PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

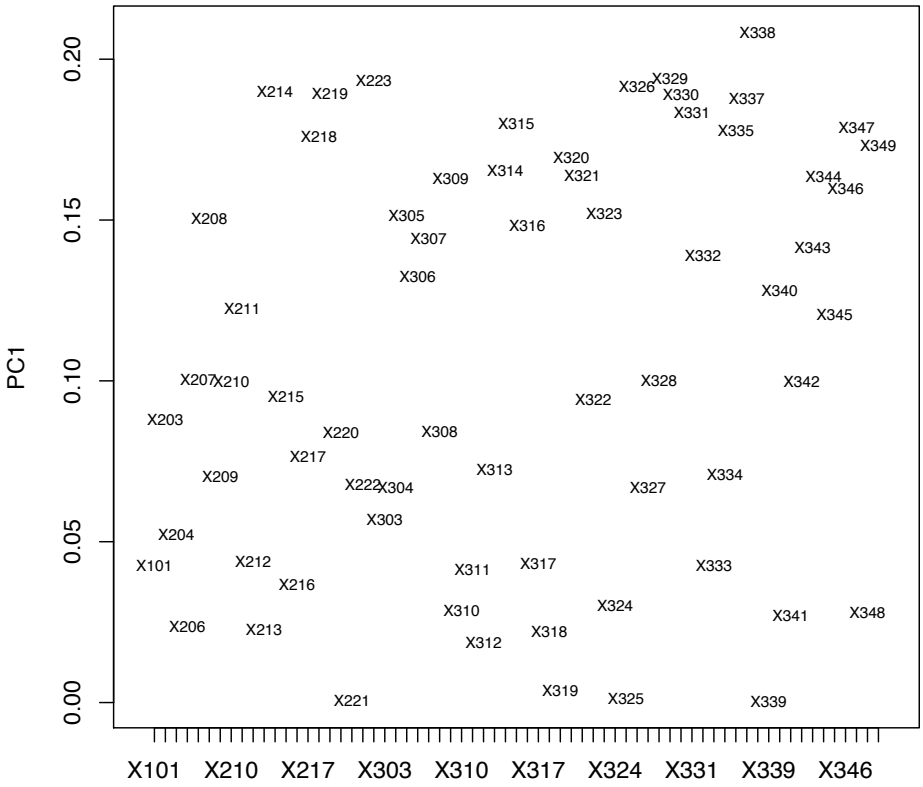


FIGURE 54 – Expérience 3 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1.

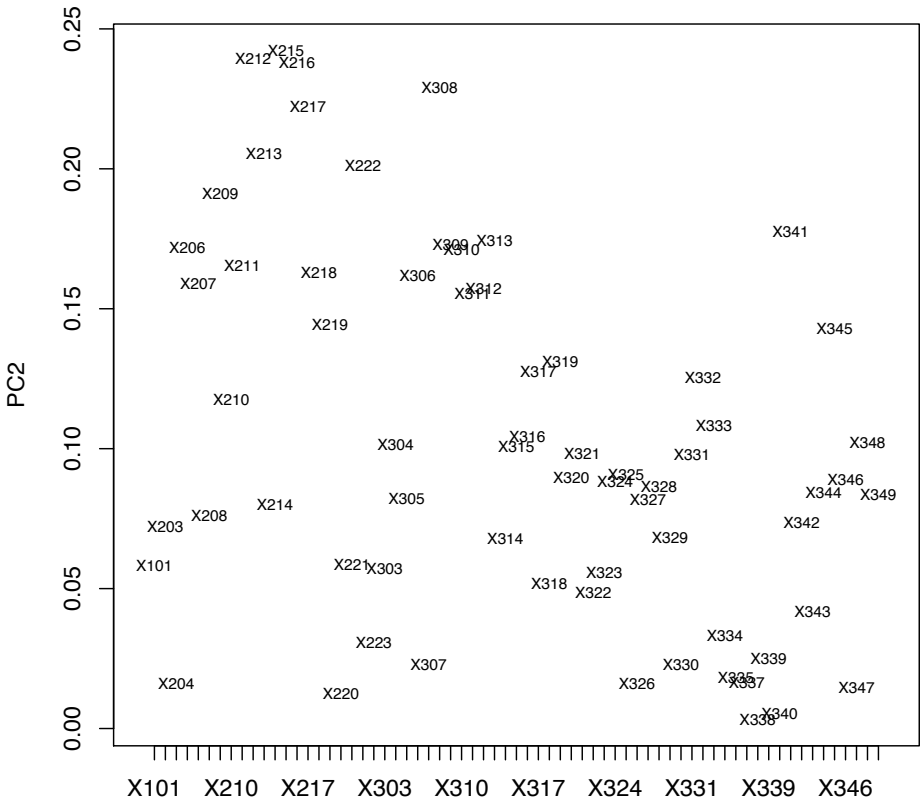


FIGURE 55 – Expérience 3 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2.

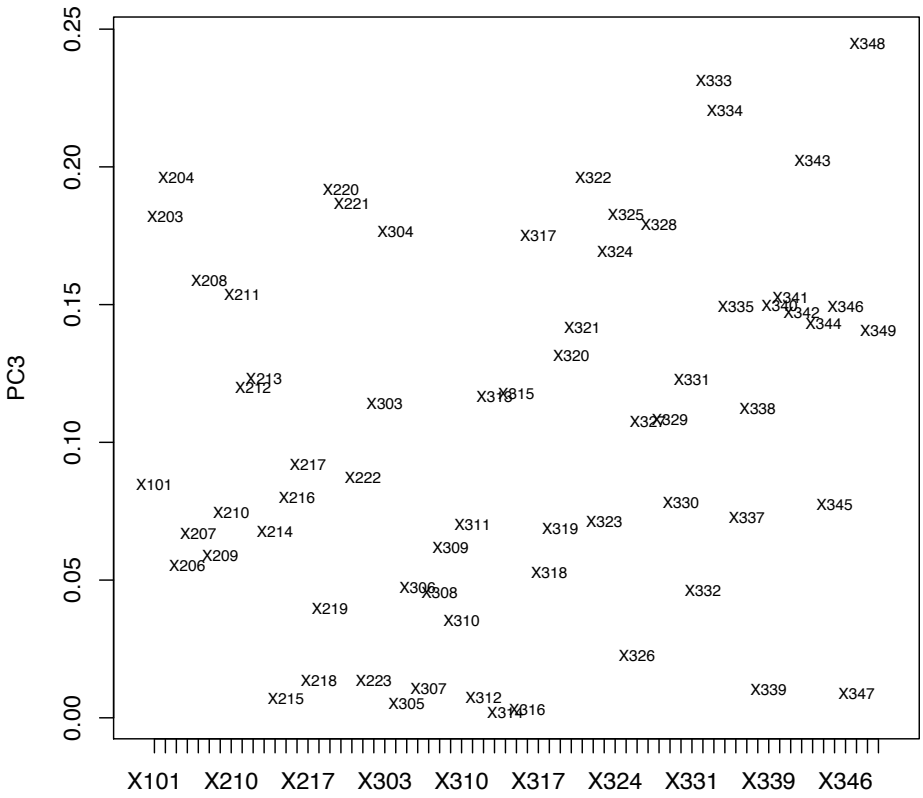


FIGURE 56 – Expérience 3 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3.

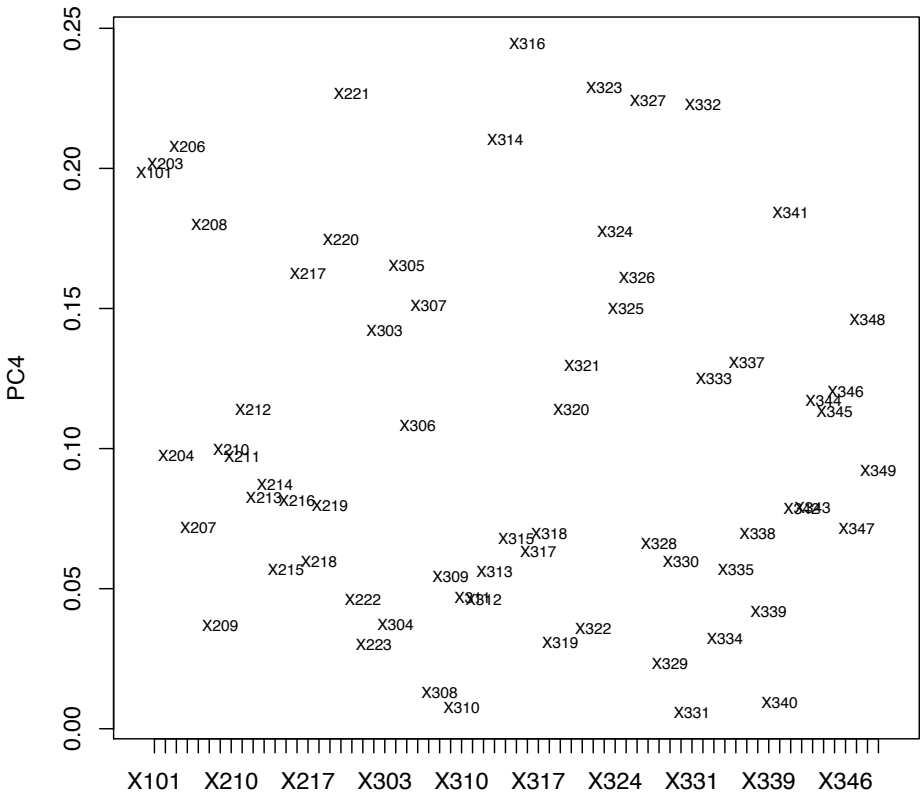


FIGURE 57 – Expérience 3 : Projection des variables sur l'axe factoriel PC4 montrant leur poids en valeur absolue dans la composante principale PC4.

B.4 Expérience 4 : ACP sur 12 variables, X_{348}, X_{333}, X_{334}, X_{343}, X_{204}, X_{322}, X_{215}, X_{212}, X_{216}, X_{308}, X_{338}, X_{217}, sélectionnées dans l'expérience 3 ainsi que les 19 premières observations labellisées en périodes.

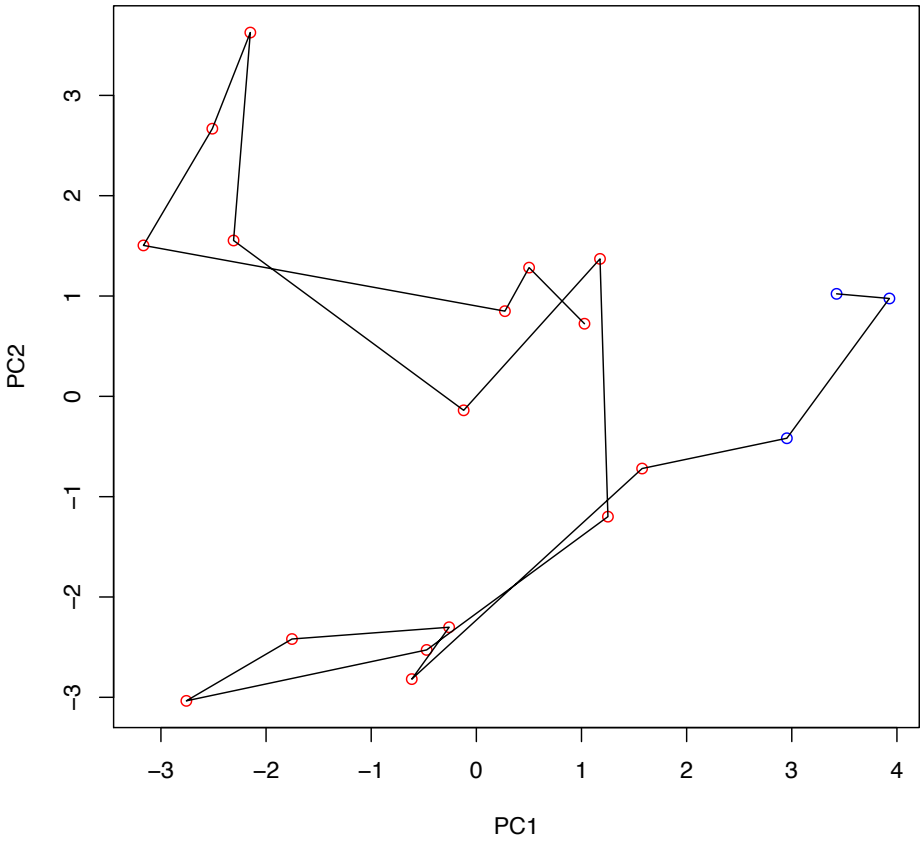


FIGURE 58 – Expérience 4 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

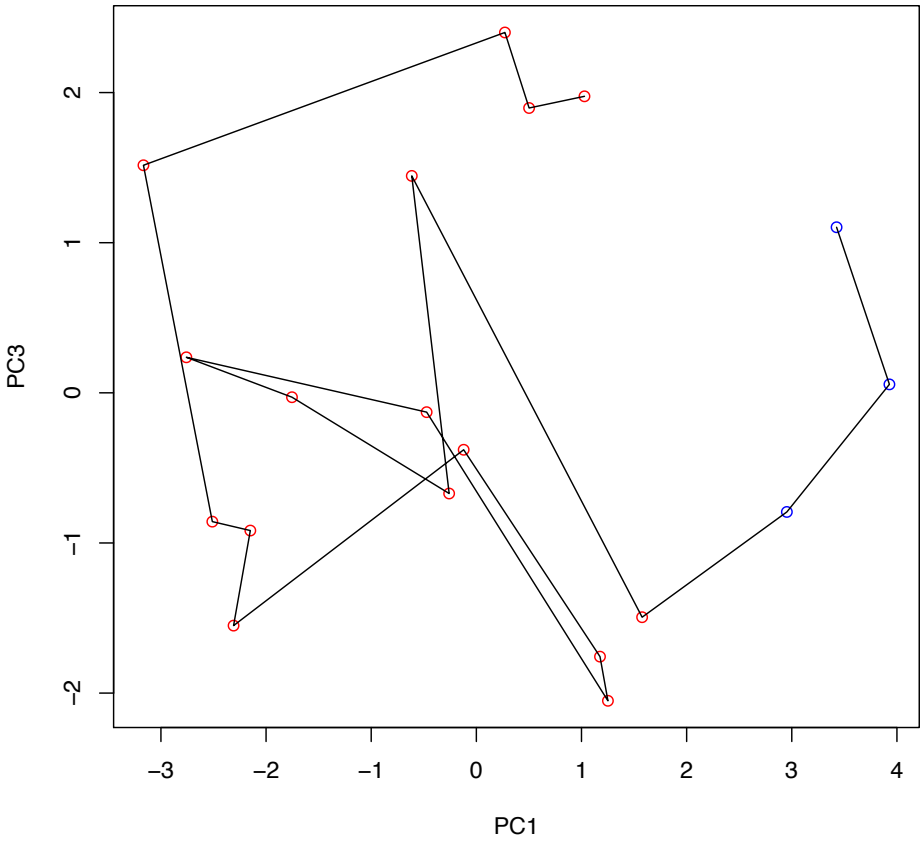


FIGURE 59 – Expérience 4 : Projection des individus sur le plan factoriel PC1× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

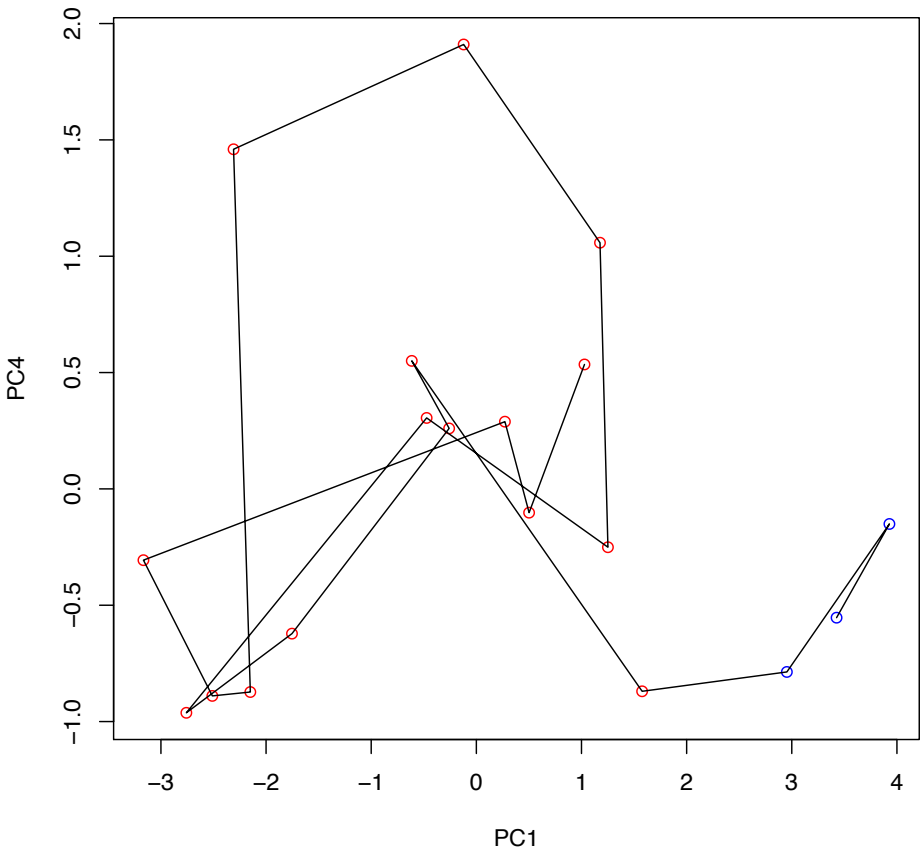


FIGURE 60 – Expérience 4 : Projection des individus sur le plan factoriel PC1× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

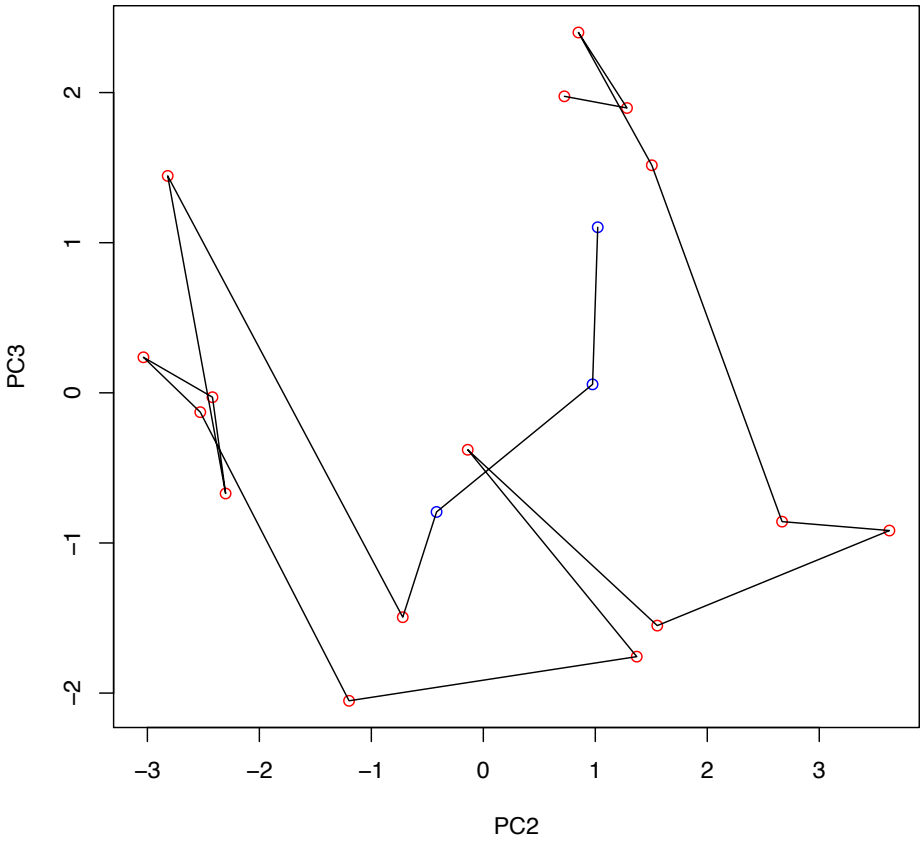


FIGURE 61 – Expérience 4 : Projection des individus sur le plan factoriel PC2× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

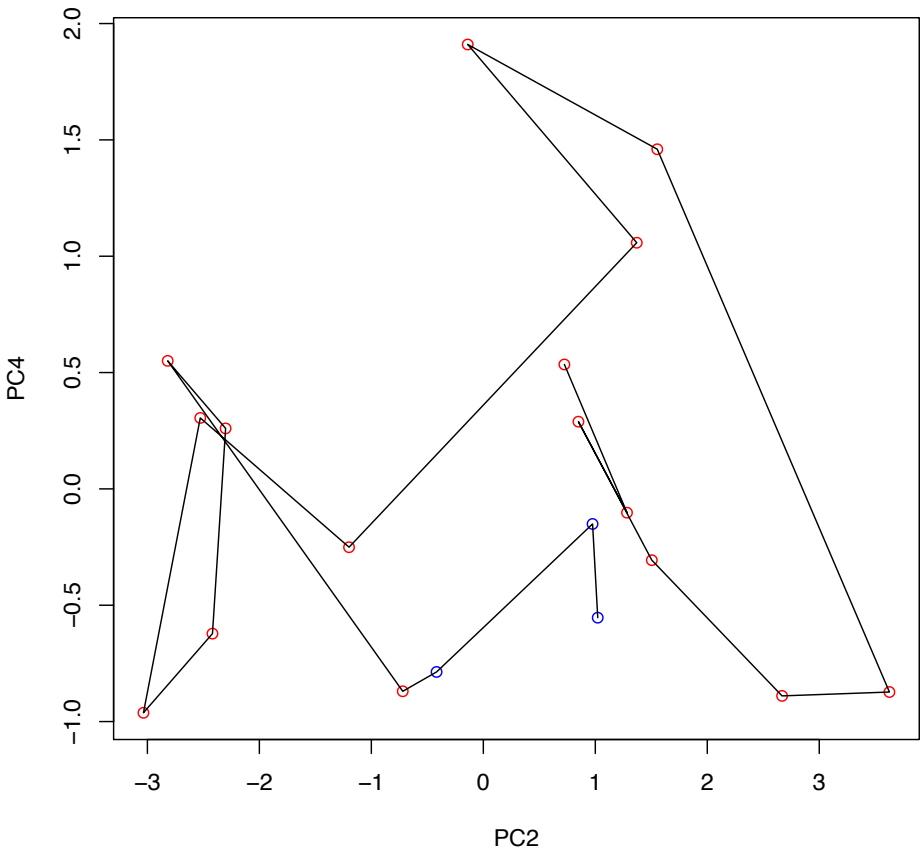


FIGURE 62 – Expérience 4 : Projection des individus sur le plan factoriel PC2× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

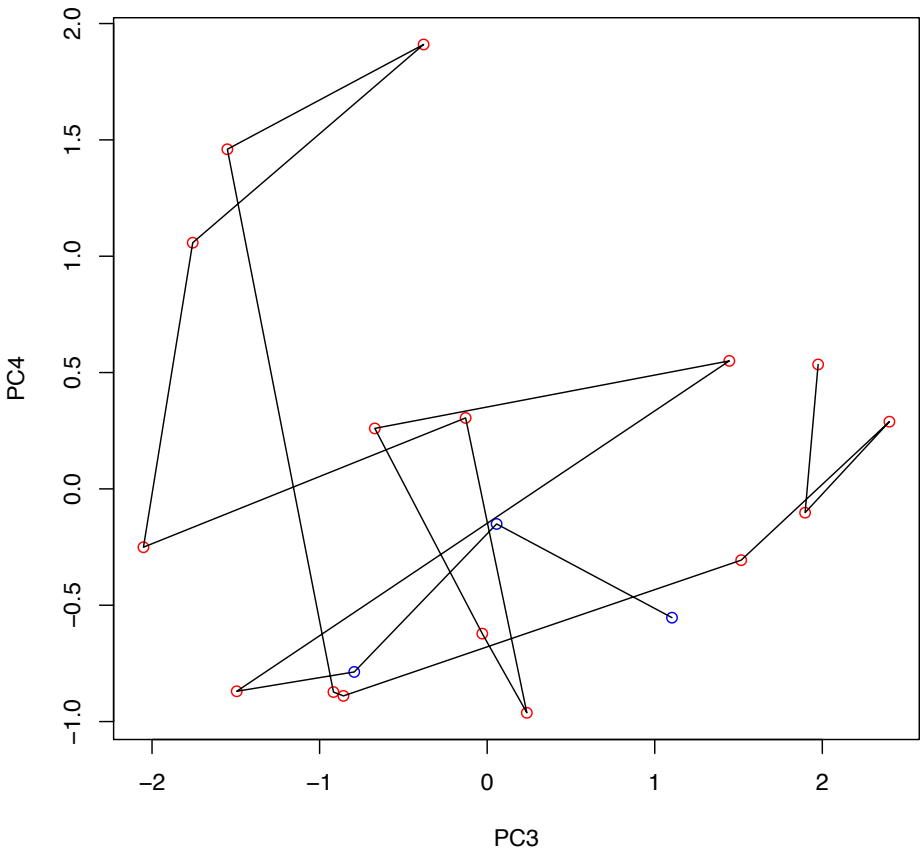


FIGURE 63 – Expérience 4 : Projection des individus sur le plan factoriel PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

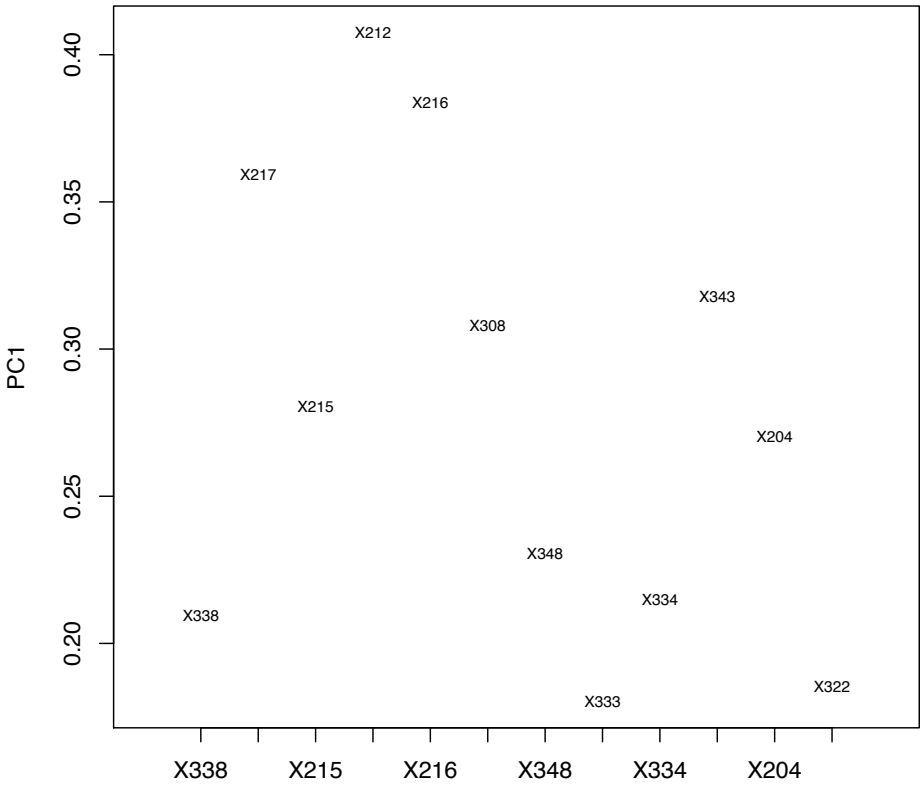


FIGURE 64 – Expérience 4 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1.

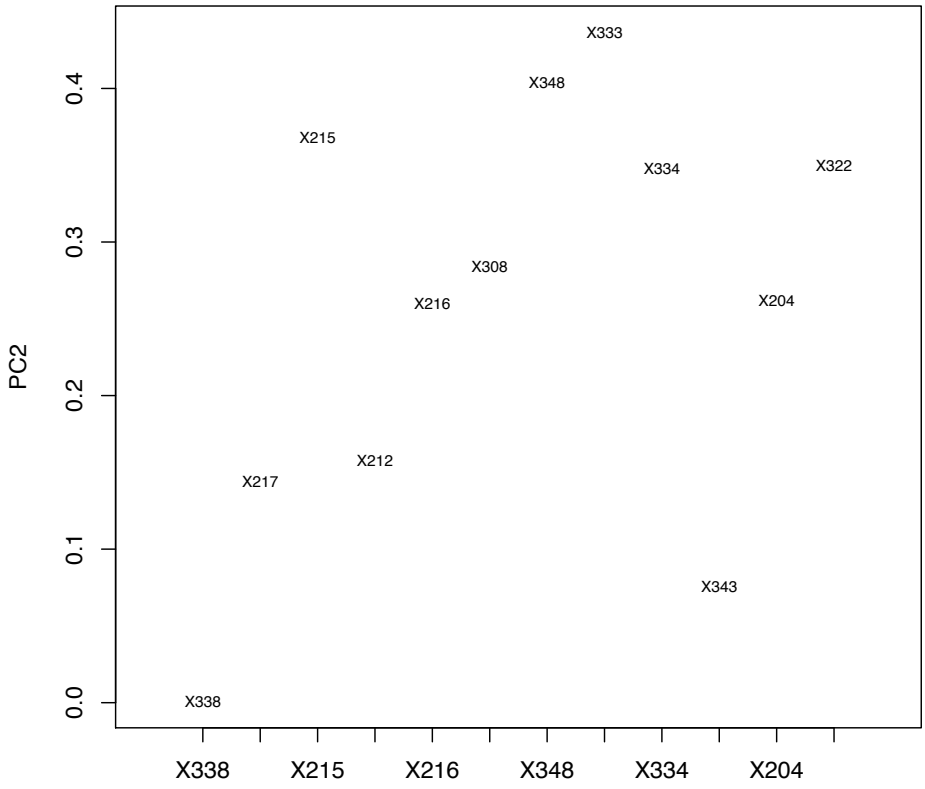


FIGURE 65 – Expérience 4 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2.

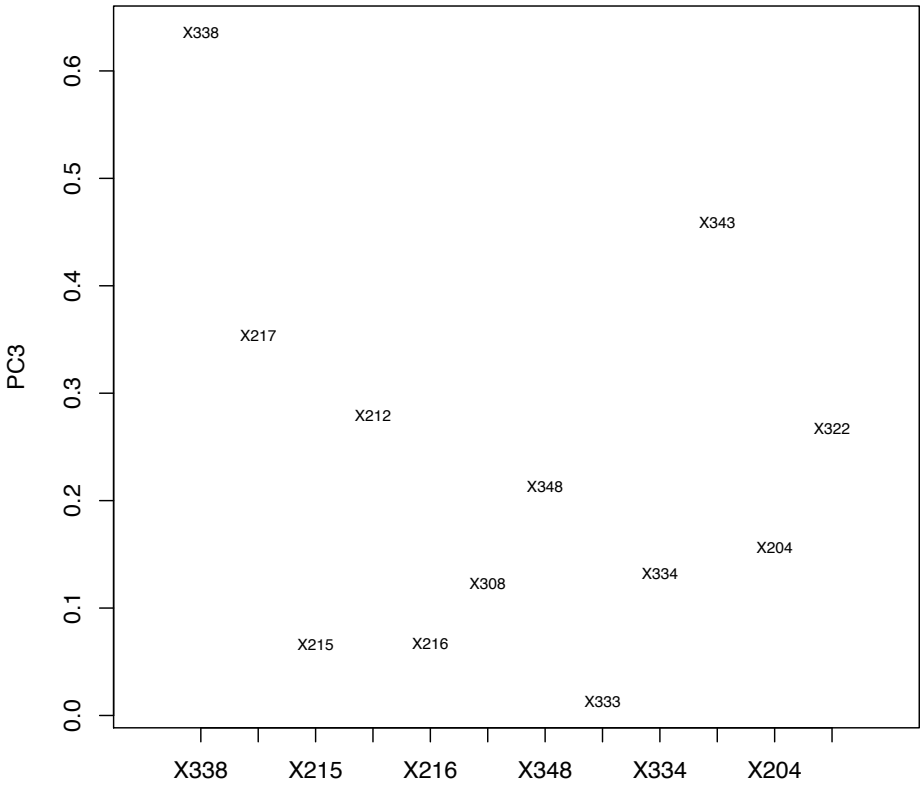


FIGURE 66 – Expérience 4 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3.

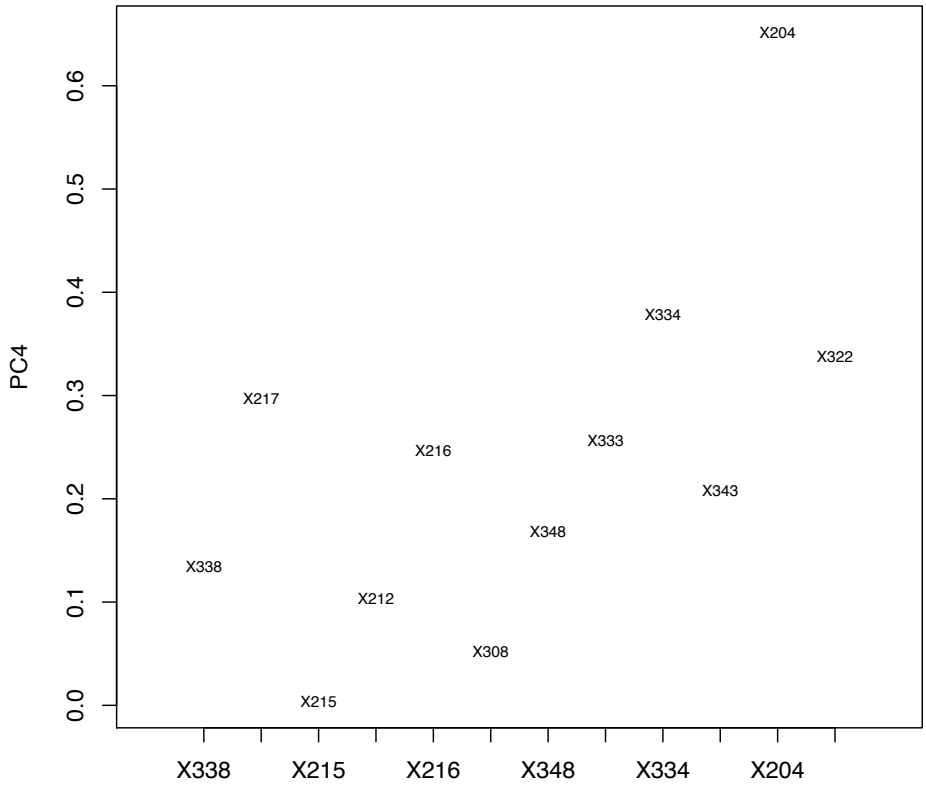


FIGURE 67 – Expérience 4 : Projection des variables sur l'axe factoriel PC4 montrant leur poids en valeur absolue dans la composante principale PC4.

B.5 Expérience 5 : ACP sur toutes les variables avec les 33 premières observations labellisées en périodes. (On ne considère pas l'après.)

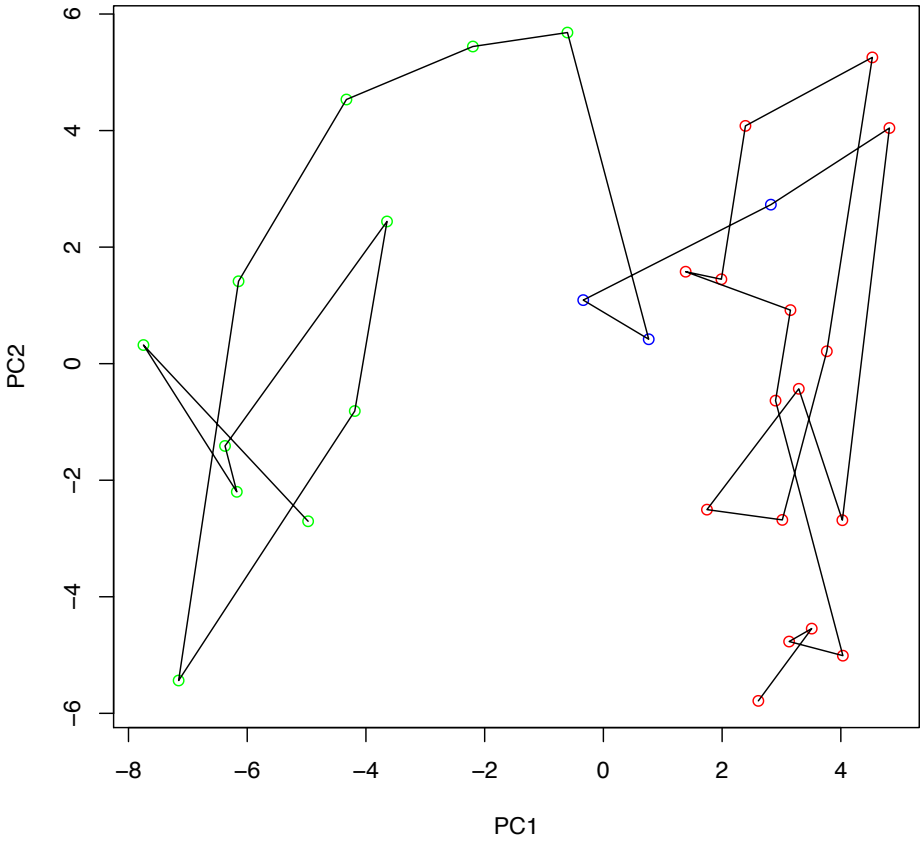


FIGURE 68 – Expérience 5 : Projection des individus sur le plan factoriel PC1× PC2. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

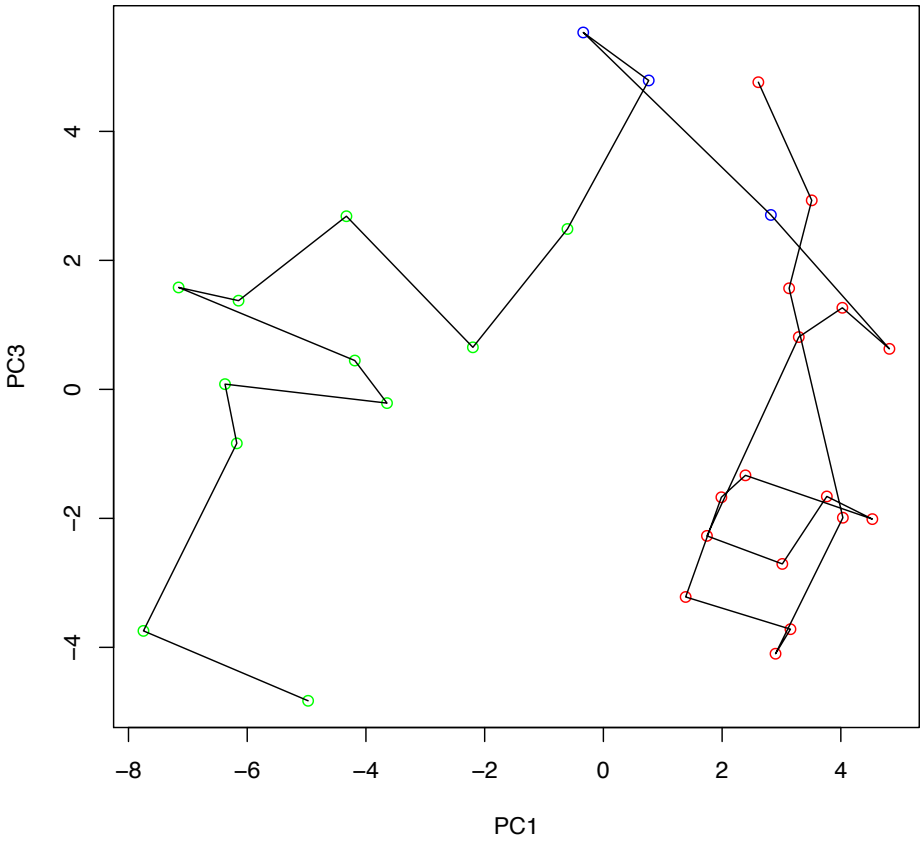


FIGURE 69 – Expérience 5 : Projection des individus sur le plan factoriel PC1× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

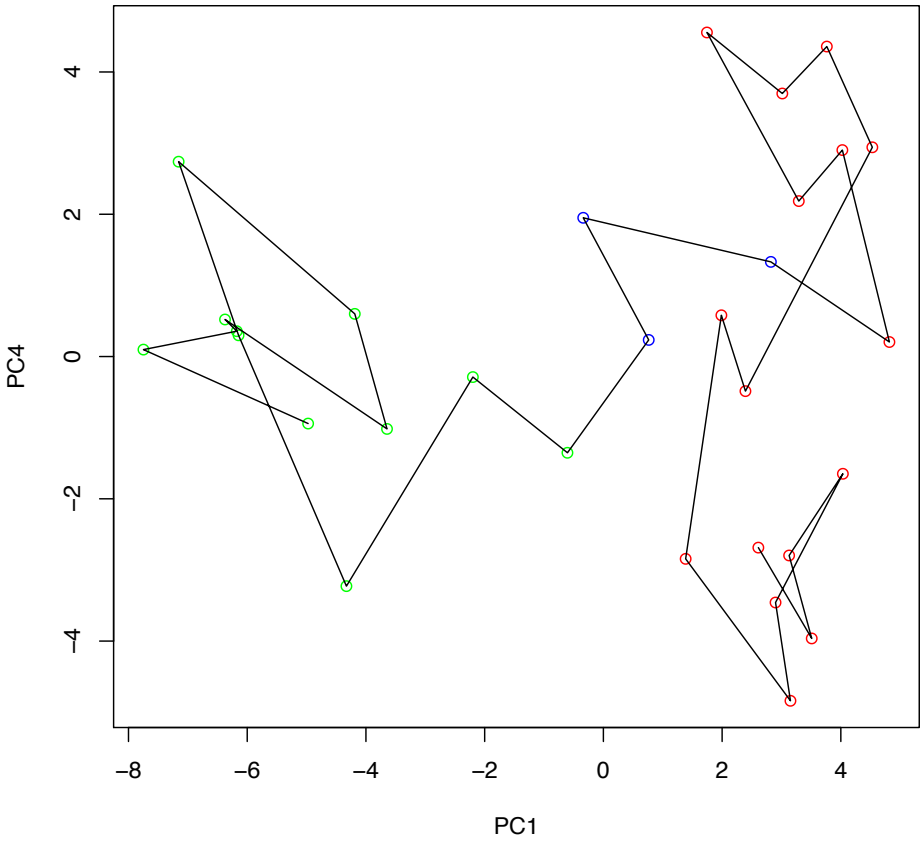


FIGURE 70 – Expérience 5 : Projection des individus sur le plan factoriel PC1× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

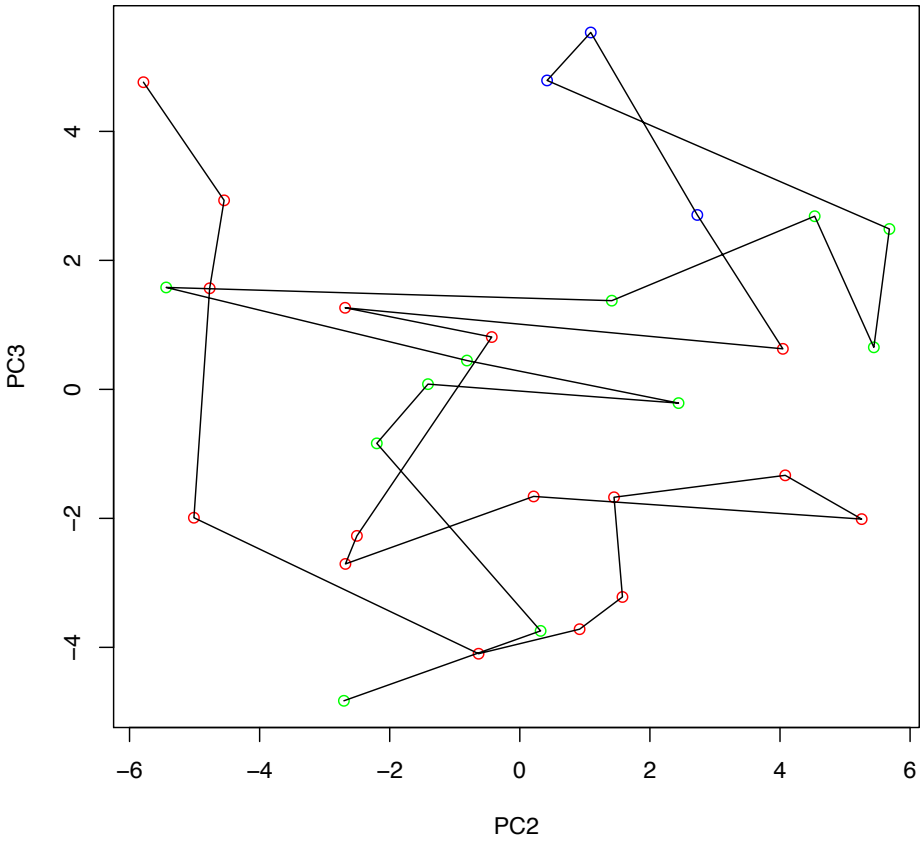


FIGURE 71 – Expérience 5 : Projection des individus sur le plan factoriel PC2× PC3. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

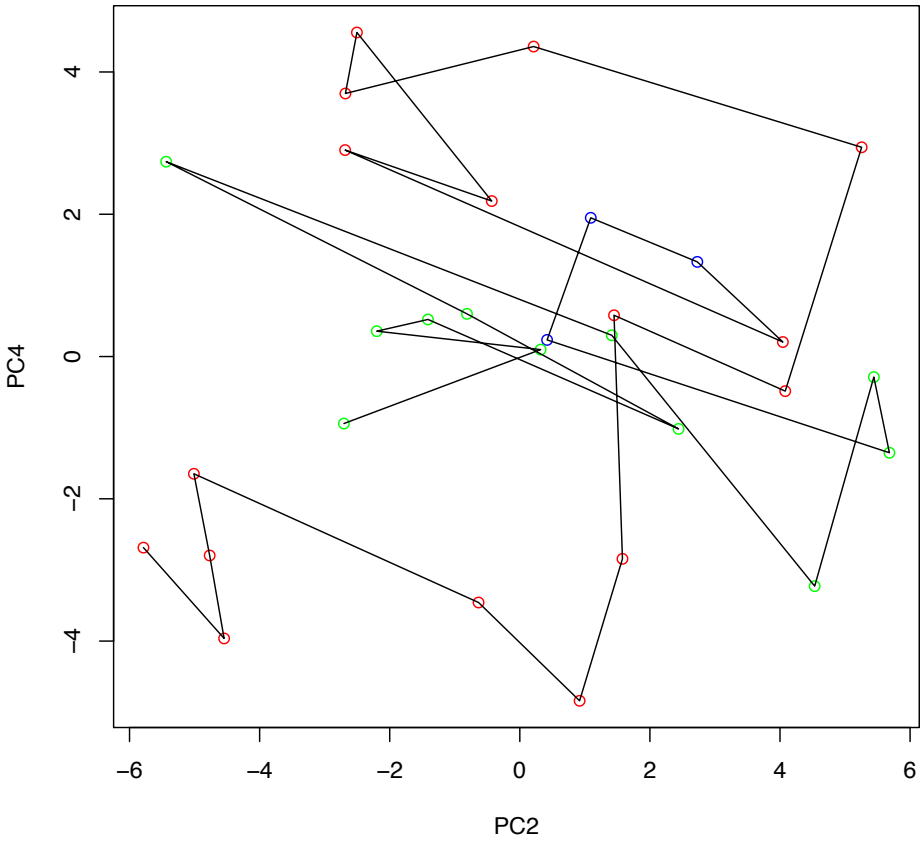


FIGURE 72 – Expérience 5 : Projection des individus sur le plan factoriel PC2× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

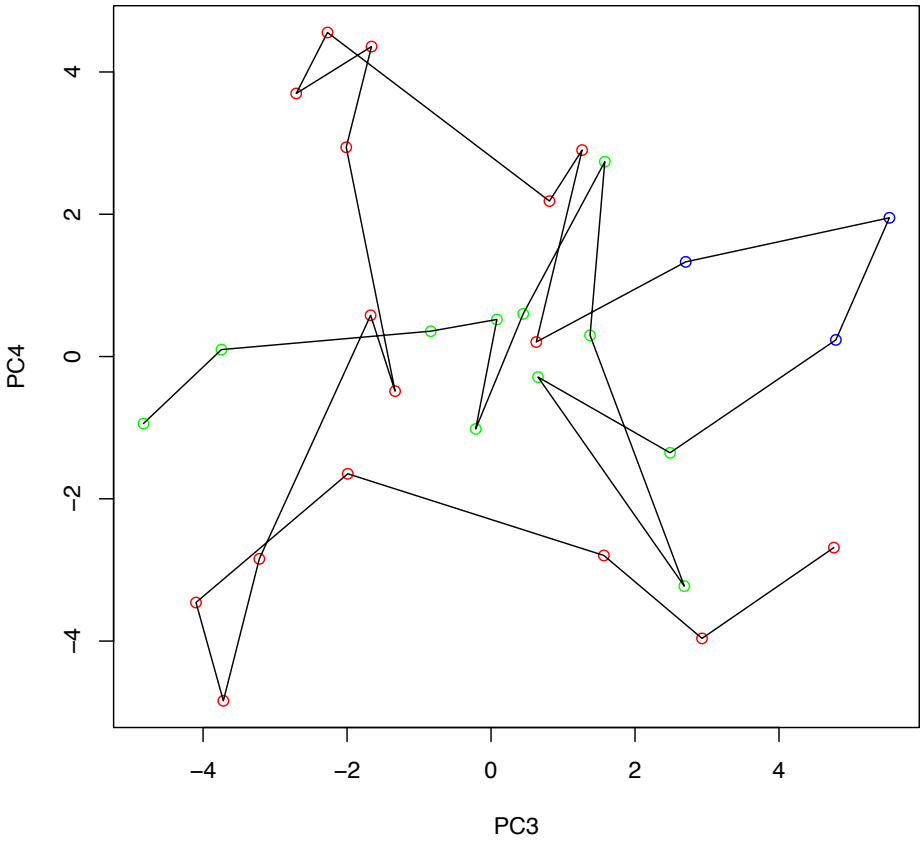


FIGURE 73 – Expérience 5 : Projection des individus sur le plan factoriel PC3× PC4. Les points correspondent à chaque semaine et portent la couleur de leur classe. La ligne noire correspond à l'évolution dans le temps.

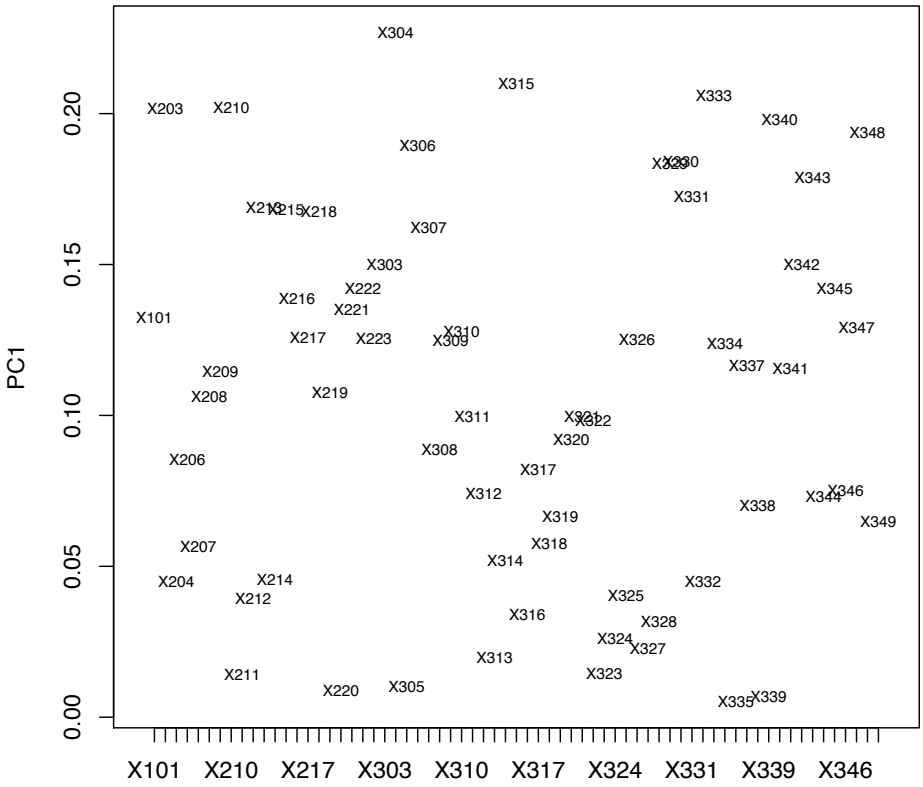


FIGURE 74 – Expérience 5 : Projection des variables sur l'axe factoriel PC1 montrant leur poids en valeur absolue dans la composante principale PC1.

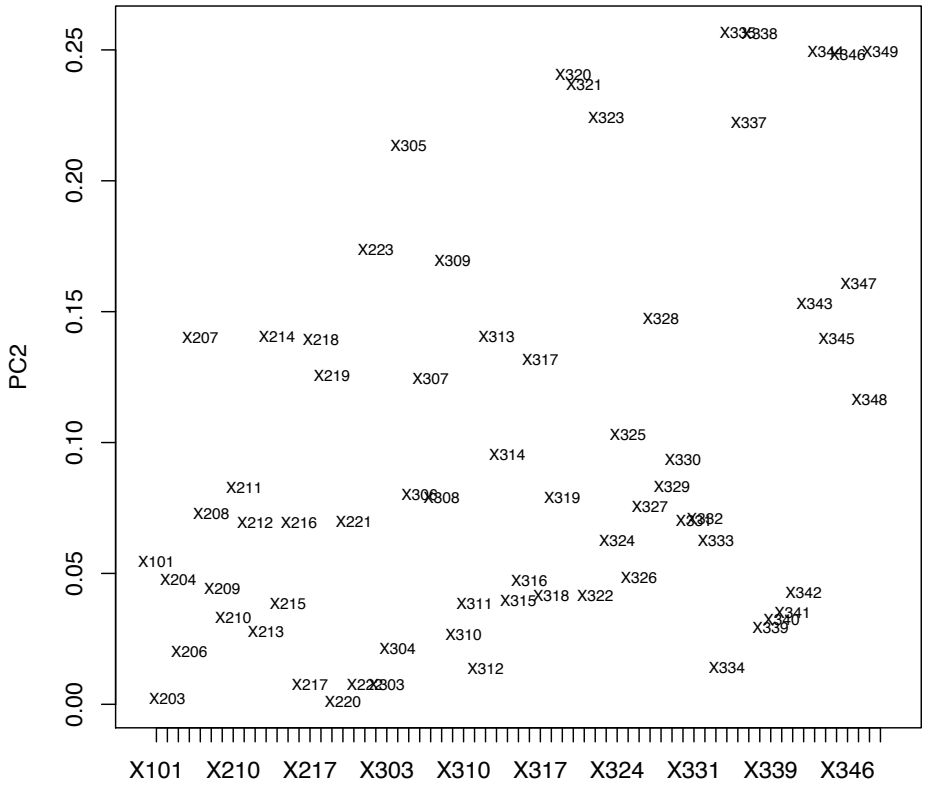


FIGURE 75 – Expérience 5 : Projection des variables sur l'axe factoriel PC2 montrant leur poids en valeur absolue dans la composante principale PC2.

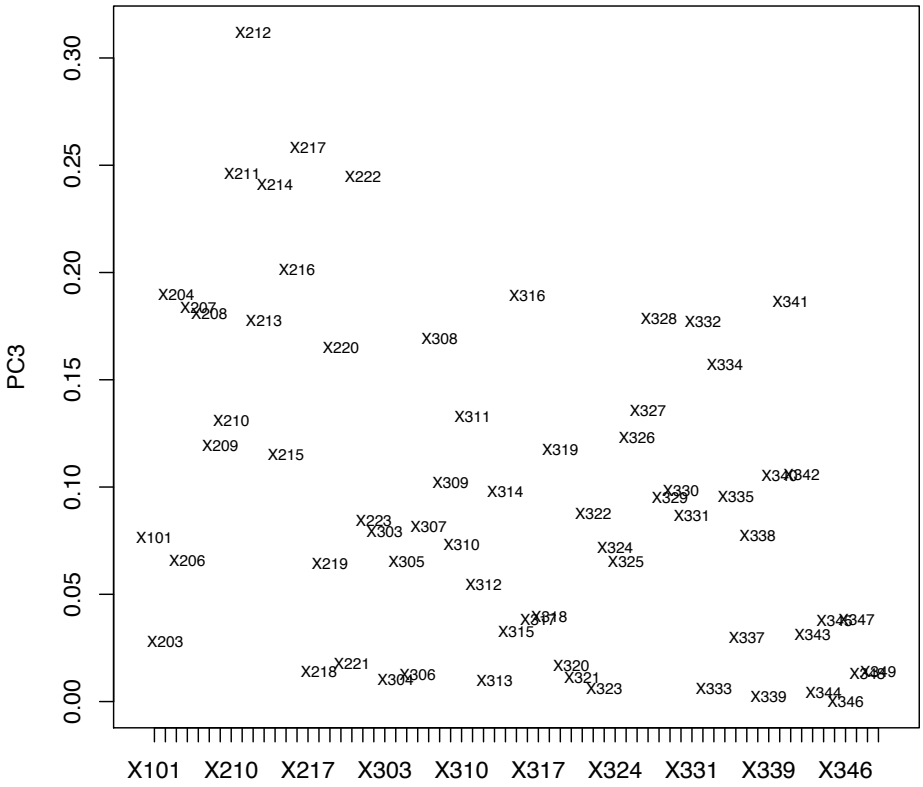


FIGURE 76 – Expérience 5 : Projection des variables sur l'axe factoriel PC3 montrant leur poids en valeur absolue dans la composante principale PC3.

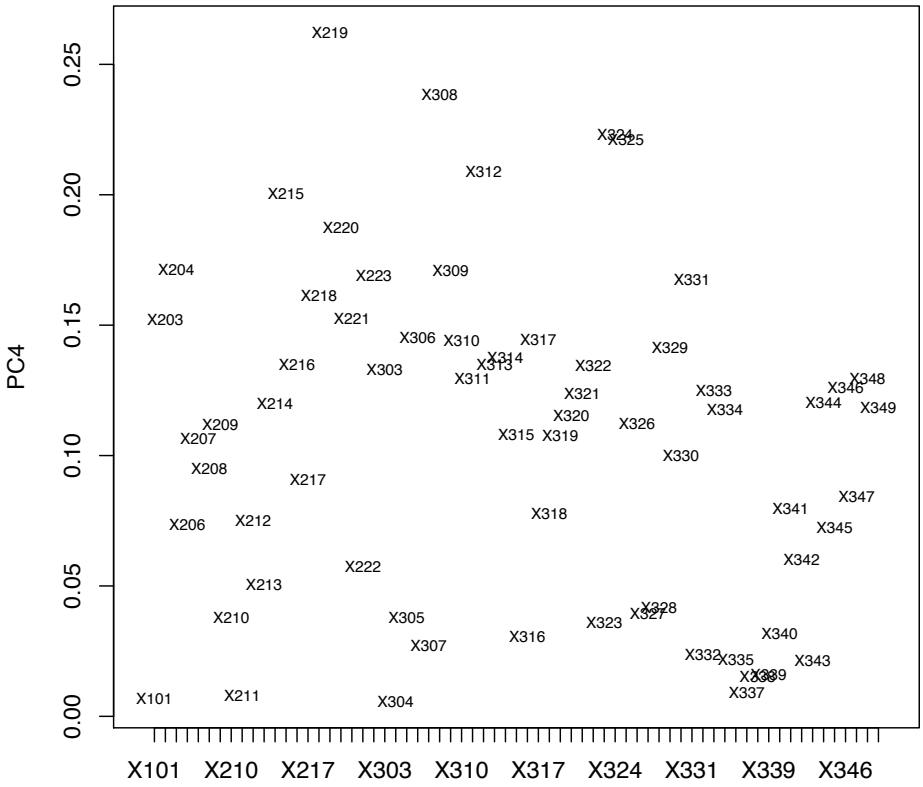


FIGURE 77 – Expérience 5 : Projection des variables sur l'axe factoriel PC4 montrant leur poids en valeur absolue dans la composante principale PC4.