Projet de Stastistiques Appliquées

Clustering sur des bases de données à valeurs manquantes

pour identifier des états sous-jacents du système économique et financier

Aaron Mamann
Aeson Feehan
Alexis Ayme
Romain Ilbert

Encadrant : Guillaume Lecué

Contents

1 Introduction

Ce projet de statistiques appliquées se déroule sous la supervision de Guillaume Lecué (CREST), en partenariat avec la Banque de France et l'Institut Louis Bachelier. Il est réalisé dans le but d'utiliser des méthodes de machine learning non supervisées, à savoir du clustering pour identifier des états sous-jacents du système économique et financier.

Le grand défi de ce projet est que ce clustering doit être réalisé sur des bases de données comportant beaucoup de valeurs manquantes. Ainsi, parmi les 4 méthodes mises en place dans ce projet,deux utiliseront des algorithmes pour remplacer les données manquantes pour ensuite clusteriser sur la base de données remplie et les deux autres consisteront à directement clusteriser sur la base de données incomplète en utilisant les algorithmes de clustering adaptés aux données manquantes.

Plus concrètement, notre base de données prend comme observations des dates (sous la forme jour/mois/année) et comme variables des indices économiques à savoir des indices boursiers (CAC 40, Eurostoxx), des indices sur le secteur bancaire (actifs bancaires consolidés ou rapport crédit/dépôts), et d'autres indices macro-économiques (comme l'inflation ou le déficit budgétaire).Ainsi, chaque date (i.e chaque observation) représente un contexte économique. Par conséquent, en regroupant en cluster les dates qui se ressemblent, chaque cluster de dates racontera un contexte économique particulier qu'on prendra le soin d'analyser.

2 Constitution de notre base de données

2.1 Séries temporelles de fréquences différentes et Apparition de valeurs manquantes

La Banque de France nous a envoyé plus d'une dizaine de bases de données différentes sous la forme de séries temporelles de fréquences différentes. Dans certains bases de données, les dates étaient relevées tous les mois, dans d'autres une fois par trimestre ou encore une fois par an, notamment au mois de janvier. Nous avons alors fusionné toutes ces bases de données. Cela a conduit à rassembler toutes ces séries temporelles de fréquences différentes. Dans cette nouvelle base, toutes ces séries temporelles ont désormais une fréquence mensuelle, ce qui génère un nombre important de valeurs manquantes : les variables relevées annuellement au mois de Janvier renvoient des valeurs manquantes de février jusqu'à décembre (et ce pour tous les ans).

2.2 Hétérogénéité des économies

Dans les bases de données apportées par la Banque de France, les variables économiques concernaient plusieurs pays dont les économies sont hétérogènes à l'instar de la France et l'Espagne. Ainsi, comme l'objectif est d'identifier des périodes économiques, nous nous

devions d'avoir des économies homogènes. Nous nous sommes donc restreints au cas de la France.

2.3 Compromis entre le nombre de variables et l'interprétabilité des clusters

La base de donnée fusionnée comportaient environ 80 variables. Pour caractériser ces clusters, qui représentent des groupes de périodes que nous devons associer à des états économiques particuliers, nous devons comparer les distributions des variables au sein de chaque cluster : par exemple nous dirons, dans le cas où le premier cluster rassemble des dates pour lesquelles l'indice CAC 40 et l'inflation sont élevés, que ce cluster représente une période de prospérité boursière et de haute inflation.

Nous avons fait le choix de retenir 17 variables parmi plus de 80 variables. **En effet, clusteriser sur 80 variables mène à créer des clusters qui ne se distinguent pas sur leurs valeurs à chacune des 80 variables mais sur leur combinaison de valeurs sur l'ensemble des 80 variables. Ainsi, en clusterisant sur ces 80 variables, si on retient 20 variables pour l'analyse des clusters, les clusters seront relativement similaires par rapport à ces 20 variables ce qui est problématique (pour avoir une vision claire des clusters on ne peut pas les comparer sur 80 variables).** Nous avons donc été obligé de réduire le nombre de variables (en aggrégeant des variables, en retirant les variables redondantes, en enlevant les variables jouant un rôle marginal dans l'économie).

Ainsi :

La base de donnée créée est donc une base où les observations sont mensuelles, à des dates comprises entre 2000 et 2019, concernent l'économie française et pour lesquelles 17 variables ont été retenues (les raisons de ces choix sont données dans la section 2.1, 2.2 et 2.3). Il est nécessaire de comprendre les variables de la base de données pour comprendre le projet. Nous avons mis en annexe les 17 variables choisies ainsi qu'une explication de leur sens économique.

3 ACP à valeurs manquantes (méthode 1)

3.1 Explication de l'algorithme à travers un cas concret

Dans cet algorithme d'ACP à valeurs manquantes [1], l'ACP est utilisée pour remplacer les données manquantes. Pour illustrer cet algorithme, considérons un exemple : une base de données avec 5 observations, deux variables, et une valeur manquante pour la variable X_2 de l'observation 4. Cet exemple est illustré par la Figure 1. Cette Figure 1 représente dans les 9

[1] Julie Josse, François Husson & Jérôme Pagès (2009) *Handling missing values in Principal Component Analysis* In Journal de la Société Française de Statistique Volume 150, numéro 2.

mini-graphiques le nuage de points des observations dans la base constituée par les variables (X_1,X_2).

La première étape de l'algorithme, visible sur les mini-graphiques 1 à 5 (Figure 1), consiste à remplir les observations manquantes par une valeur initiale. Ici, pour la valeur manquante dans la Figure 1 (c'est la variable X_2 de l'observation 4), nous remplaçons d'abord cette valeur manquante par la moyenne empirique de la variable concernée (la variable X_2). Puis, on effectue une ACP sur la base de données désormais remplie. La valeur de la projection de l'observation 4 est alors utilisée pour remplacer la valeur manquante. Plus précisément la coordonnée en X_2 de la projection de l'observation 4 est choisie pour remplacer la coordonnée en X_2 de l'observation 4. Dans cet algorithme, bien que les valeurs initialement manquantes soient remplacées, les valeurs non-manquantes à l'origine restent cependant inchangées. Sur cette nouvelle base de données, le même procédé est répété : on ré-applique l'ACP sur la nouvelle base. Les valeurs affectées à l'itération k-1 sont utilisées pour faire une ACP à l'itération k qui permettra de donner une nouvelle affectation à ces valeurs (encore une fois, seules les valeurs étant manquantes à l'état initial sont concernées, les autres valeurs de la base de données restent inchangées). On répète l'algorithme jusqu'à convergence du résultat, c'est à dire quand la valeur initialement manquante se voit toujours réaffecter la même valeur à chaque itération.

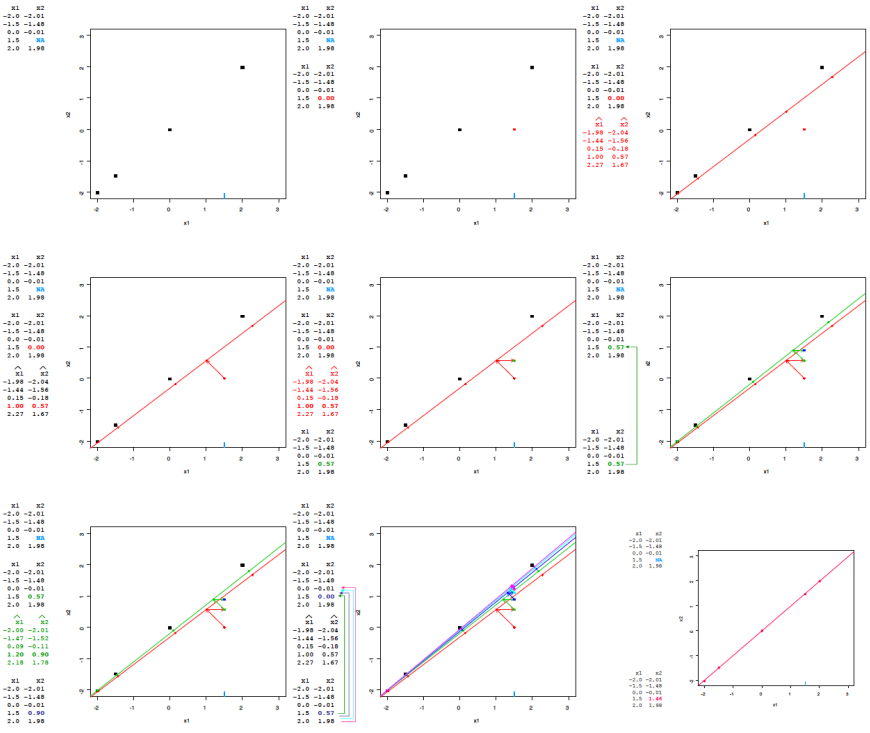

Figure 1: ACP à valeurs manquantes (François Husson, Jolie Josse)

3.2 Avantages de cette méthode de remplacement de valeurs manquantes pour notre projet :

Revenons dans le cadre de notre base de données qui représente des dates (étant les observations) caractérisées par des indices économiques (étant les variables). Chaque date représente donc un contexte économique.

Cet algorithme développé par François Husson et Jolie Josse a pour avantage d'affecter les valeurs en prenant en compte le contexte économique de chaque date.

En effet, cet ACP à valeurs manquantes a pour avantage d'affecter des valeurs aux données manquantes en prenant en compte les relations entre les lignes (i.e. les dates) et entre les colonnes (i.e. les variables) de la base de données : par exemple si à la date"01/02/2008" le déficit budgétaire n'est pas renseigné, l'algorithme va prendre en compte les autres valeurs renseignées sur cette observation. En remarquant que l'indice CAC 40 et les actifs bancaires sont faibles, l'algorithme va assigner au "01/02/2008" un déficit budgétaire similaire aux dates (i.e. aux lignes) ayant un indice CAC 40 et des actifs bancaires faibles.

Pour réaliser cet ACP à valeurs manquantes qui va compléter la base de données, on a fait appel sous le logiciel R à un package réalisé par Julie Josse et François Husson eux-mêmes (ceux qui sont à l'origine de cette méthode d'ACP à valeurs manquantes), le package est nommé "missMDA".

3.3 Clustering sur la base de donnée complétée : k-means

Maintenant que la base de donnée est complétée, on peut maintenant utiliser différents algorithmes de clustering. On a d'abord choisi un algorithme k-means .

Notons qu'après avoir complété la base de données, nous avons de nouveau fait une ACP de dimension 2 dessus et avons clusterisé sur le nuage de points projeté. Ce choix s'explique pour plusieurs raisons. Comme l'ont dit leurs créateurs (et cela peut bien se comprendre aussi), les affectations données aux valeurs manquantes par l'algorithme d'ACP à valeurs manquantes sont très pertinentes quand il s'agit de faire ensuite une ACP sur cette base de données complétée. La deuxième raison de ce choix est qu'on peut régler le k-means de manière optimale quand il s'agit de clusteriser sur un plan. En effet, le reproche souvent fait au k-means est que le choix du nombre de cluster est laissé à l'utilisateur, ainsi que le choix des centroïdes initiaux. Ces choix sont souvent non optimaux. Sur la base de données projetée sur un plan nous avons pu distinguer à vue d'oeil les clusters et donc définir de manière plus optimale les centroïdes initiaux ainsi que le nombre de clusters.

3.3.1 Caractérisation économique des Clusters

Rappelons le, notre base de donnée prend comme observations des dates. Chaque date représente un contexte économique. En effet, chaque date est caractérisée par des indices économiques (constituant les variables de la base de données). Il s'agit alors de regrouper

les dates qui se ressemblent sur ces indices économiques. Ainsi, ces clusters de dates raconte un contexte économique particulier qu'on va maintenant étudier.

On va d'abord caractériser les clusters sur le plan financier.

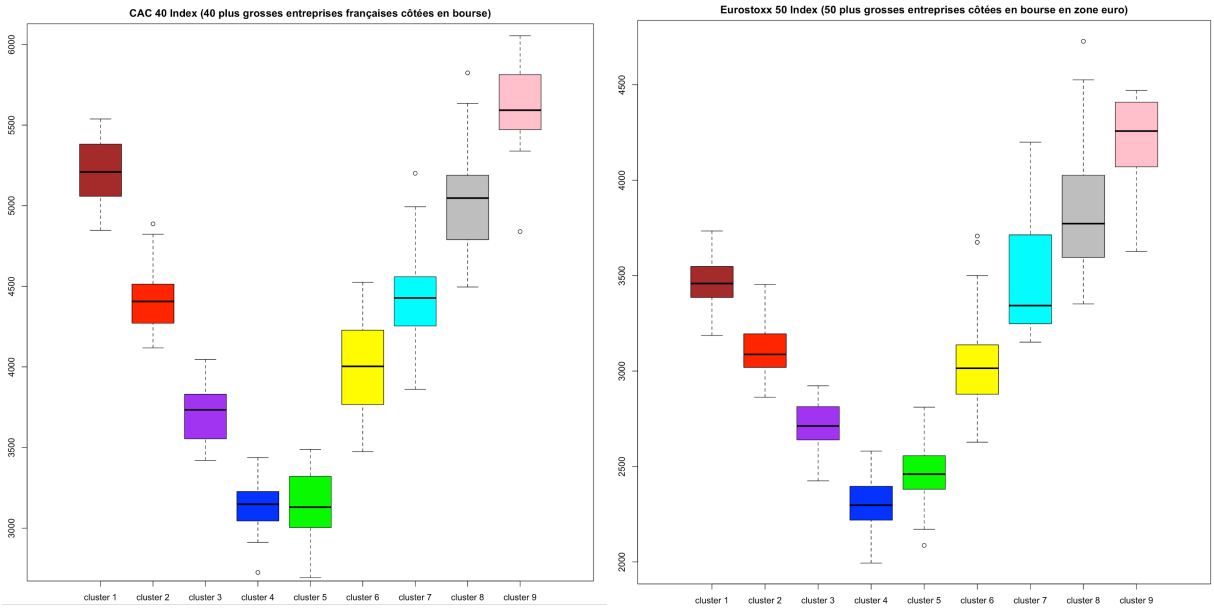

Figure 2: Comparaison (par des boxplots) des 9 clusters par rapport aux indices boursiers français et européens

On remarque d'abord que dans les deux schémas de la figure 2, les boxplots sont assez resserrés et ne se chevauchent pas d'un cluster à l'autre. Cela montre donc que chaque cluster de dates identifié représente à lui seul un état financier précis et unique.

Ce qui est également cohérent, c'est que les clusters ayant un indice CAC faible (resp élevé) ont aussi un indice Eurostoxx faible (resp. élevé).

Ainsi, alors que le cluster 4 et 5 représentent des périodes de crise des marchés boursiers, les cluster 1,8,9 représentent au contraire des périodes de prospérité de ces marchés.

On va maintenant caractériser les clusters sur la santé du secteur bancaire.

On va donc étudier la distribution dans chaque cluster des variables "actifs bancaires consolidés" et "cours/valeur comptable des banques" comme on peut le voir dans les deux graphiques (cf Figure 3). Rappelons d'abord que les actifs consolidés d'une société regroupent l'ensemble des actifs de la société mère et de ses filiales ce qui donne un bilan plus objectif de l'état financier de l'entreprise en question. Rappelons ensuite qu'un fort (resp. faible) ratio cours/valeur comptable peut signifier que l'action est sur-évaluée (resp. sous évaluée). Il peut aussi rendre compte d'un dysfonctionnement fondamental au sein de l'entreprise.

On remarque d'abord sur la figure 3 à gauche que les clusters de dates représentant des périodes de prospérité financière (resp crises financières) représentent également des périodes où les actifs des banques sont importants (resp faibles). Cependant, l'augmentation

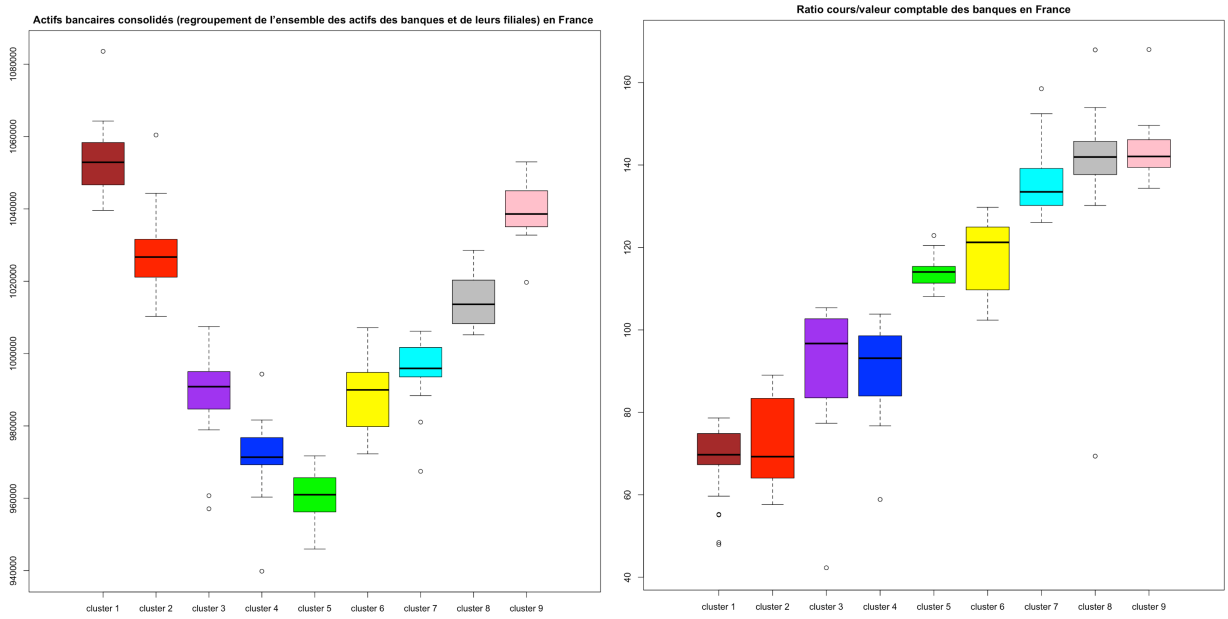

Figure 3: Comparaison (par des boxplots) des 9 clusters par rapport à la santé du secteur bancaire

de ces actifs peut cacher un décrochage entre la valeur boursière des banques et leur valeur comptable. Ainsi, le cluster 7,8,9 représentent des périodes où les actifs bancaires sont importants mais aussi des périodes où le cours des actions bancaires sont environ 140 fois supérieur à celle de leur valeur comptable ce qui témoigne que les actions bancaires sont sans doute sur-évalués ou ce qui témoigne d'un dysfonctionnement fondamental des banque. Cependant, on peut voir que le cluster 1 et 2 représentent des périodes dans lesquelles les actifs bancaires sont développés et que ce développement semble plus stable que pour les clusters 7,8,9 du fait du ratio cours/valeur comptable beaucoup plus faible.

On peut également caractériser les clusters à travers d'autres indices macro-économiques comme sur le plan l'endettement de l'état français. On s'intéresse à la balance budgétaire qui depuis des années est déficitaire et qui creuse la dette publique.

On remarque d'abord (dans la figure 13 en annexe représentant la comparaison par des boxplots des 9 clusters en fonction du déficit budgétaire), que les clusters de dates représentant des périodes de prospérité financière (resp. crises financières) où les banques détiennent beaucoup d'actifs (resp. peu d'actifs) sont aussi des périodes où le déficit budgétaire est moins important (resp. plus important).

3.3.2 Vérification de la cohérence du clustering

Concentrons nous sur la figure 4 qui représente l'évolution des clusters en fonction du temps, i.e. l'évolution des états économiques en fonction du temps. Pour qu'un clustering soit cohérent d'un point de vue économique, il est important que d'un mois à l'autre on ne

change pas souvent de clusters, i.e. que d'un mois à l'autre on ne change pas souvent d'état économique. En effet, d'un mois à l'autre il n'est pas habituel (mais il est cependant possible) de changer d'état économique. Ainsi, sur cette figure 4, on devrait avoir davantage de mois consécutifs étant dans le même cluster que de mois consécutifs dans des clusters différents i.e on devrait avoir beaucoup plus de lignes que de sauts. Ainsi, on peut voir que c'est le cas et donc que le clustering est cohérent de ce point de vue.

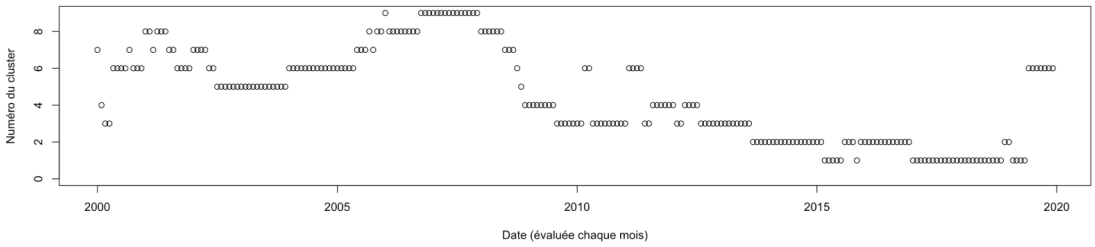

Figure 4: Évolution des clusters dans le temps représentant la succession des états économiques dans le temps (ACP à valeurs manquantes/k-means)

3.4 Autre algorithme de clustering sur la base de données complétée Density-based spatial clustering of applications with noise (DBSCAN)

Décrivons d'abord comment opère cet algorithme de clustering. Le DBSCAN commence par un point de données de départ arbitraire qui n'a pas été visité. Le voisinage de ce point est extrait en utilisant une distance epsilon ϵ. S'il y a un nombre suffisant de points (selon les minPoints) dans ce voisinage, le processus de mise en cluster démarre et le point de données actuel devient le premier point du nouveau cluster. Sinon, le point sera étiqueté comme bruit (plus tard, ce point bruyant pourrait devenir la partie du cluster). Dans les deux cas, ce point est marqué comme «visité». Pour ce premier point du nouveau cluster, les points situés dans son voisinage à distance se joignent également au même cluster. Cette procédure est ensuite répétée pour tous les nouveaux points qui viennent d'être ajoutés au groupe de cluster, et ce jusqu'à ce que tous les points du cluster soient déterminés, c'est-à-dire que tous les points à proximité du ϵ voisinage du cluster ont été visités et étiquetés. Une fois terminé avec le cluster actuel, un nouveau point non visité est récupéré et traité, ce qui permet de découvrir un nouveau cluster ou du bruit. Ce processus se répète jusqu'à ce que tous les points soient marqués comme étant visités. A la fin de tous les points visités, chaque points a été marqué comme appartenant à un cluster ou comme étant du bruit.

On n'aura pas ici la place de détailler la caractérisation économique des clusters. Par conséquent, on caractérisera les clusters (sur le plan des marchés boursiers, sur le santé du secteur bancaire, et sur la balance budgétaire) en annexe. On peut néanmoins relever des

limites de cette méthode à travers le graphique représentant les clusters à travers le temps ci-dessous.

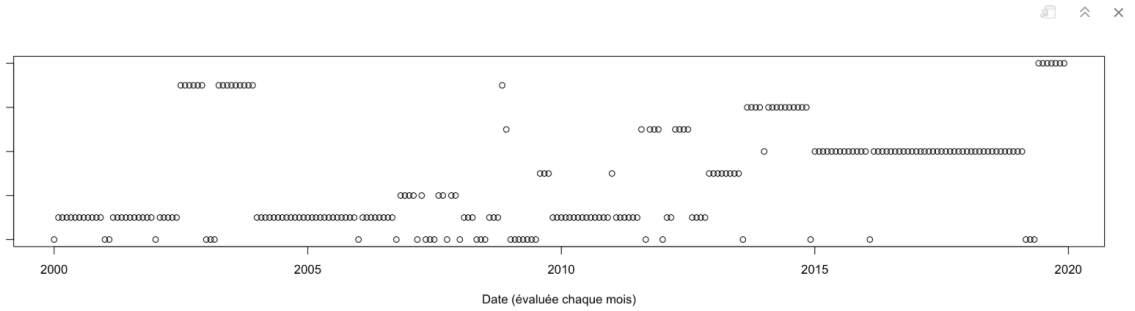

Figure 5: Evolution des clusters dans le temps représentant la succession des états économiques dans le temps (ACP à valeurs manquantes/DBSCAN)

On peut voir ici plusieurs limites. Certains clusters comportent beaucoup de dates (deux clusters rassemblent respectivement 77 dates et 45 dates sur 240 dates en tout sur la base de données) alors que d'autres pas assez (3 clusters rassemblent respectivement 4,7 et 7 dates). Les clusters rassemblant trop peu de dates racontent des périodes économiques trop précises et les clusters rassemblant beaucoup de dates racontent des périodes économiques trop vagues ce qu'on peut voir en annexe sur les boxplots qui sont relativement étalés (notamment comparés aux boxplots obtenus avec le k-means).

Cependant, ce qui est cohérent c'est qu'on voit qu'il n'y a pas beaucoup de sauts sur la figure des clusters en fonction du temps ce qui montre que d'un mois à l'autre on ne change pas souvent de cluster, i.e que d'un mois à l'autre on ne change pas souvent d'état économique, ce qui est encore une fois cohérent.

4 Fonction objectif de clustering adaptée aux données manquantes (méthode 2)

Dans cette deuxième méthode, nous allons construire une fonction objectif de clustering qui nous permette de directement clusteriser sur la base de données incomplète, ce qui diffère de la méthode précédente dans laquelle on a du d'abord compléter la base de données (i.e. affecter des valeurs aux données manquantes) pour réaliser ensuite le clustering.

L'objectif de cette partie est de définir ce que serait la fonction à minimiser de notre problème.
L'approche classique (sans données manquantes) consiste à minimiser l'énergie qui est la somme des distances au carré de chaque point à son centroïde associé.
Mais, ici définir une distance n'a de sens que si l'on prend en compte que les variables renseignées. Nous allons dans la prochaine sous-partie définir un espace et une semi-norme afin de pouvoir effectuer l'opération de la distance (qui sera une semi-distance) et celle des centroïdes.
Ensuite, nous élaborerons l'heuristique des k-means sur cet espace semi-normé et interpréterons nos résultats avec la base de données.

4.1 Espace semi normé

Soit $(X_i)_{i=1...n}$ nos observations de dimension p. Nous avons aussi la matrice Ω qui correspond à la donnée disponible (matrice de 0 et 1).
Dans ce contexte de données manquantes nous allons noter $Y_i = (X_i, \Omega_i)$ le couple de l'observation et de la donnée disponible éléments de $\mathbb{R}^p \times \mathbb{N}^p$.
Nous définissons l'addition classiquement :

$$(x, \omega) + (y, \omega') = (x + x', \omega + \omega')$$

et le produit avec un scalaire :

$$\lambda.(x, \omega) = (\lambda.x, \omega)$$

On prend alors comme semi norme :

$$N((x, \omega)) = \left(\frac{\sum_{i=1}^{p} \mathbb{1}_{\omega_i > 0} x_i^2}{\sum_{i=1}^{p} \mathbb{1}_{\omega_i > 0}} \right)^{1/2}$$

Cette norme est la somme des carrés des valeurs renseignées normalisée par le nombre de valeurs renseignées. Nous faisons cette normalisation pour ne pas que le nombre de variables renseignées ait un effet sur la norme.
La fonction de moyennisation : si $y = (x_i, \omega_i)_{i=1...p}$ alors $m(y) = \left(\frac{x_i}{\max(\omega_i, 1)}, \min(\omega_i, 1) \right)_{i=1...p}$
Cette fonction va nous permettre de définir le centroïde d'une classe.
Le but est de trouver une k-partition de $\{1, ..., n\}$ que l'on note $C_1, ..., C_k$ telle que l'on minimise la fonction :

$$\phi(C_1, ..., C_k) = \sum_{i=1}^{k} \sum_{j \in C_i} N(Y_j - M_i)$$

où

$$M_i = m\left(\sum_{j \in C_i} Y_j\right)$$

est le centroïde du cluster C_i.

Dans nos données, la valeur de ω_i sera donnée par le fait de savoir si une variable a la valeur NaN ou pas.

4.2 Heuristique des k-means

Nous avons implémenté l'heuristique des k-means en remplaçant la fonction qui calcule les centroïdes par la fonction de moyennisation explicitée dans la dernière sous partie, nous avons également remplacé la fonction de norme usuel par notre semi-norme.
Ainsi, notre algorithme de clustering ne fait jamais appel à une donnée manquante et ne cherche pas à leur attribuer une valeur. Aussi, lorsque l'algorithme calcule une distance, il le fait avec un point et un centroïde ce qui a l'avantage de rendre cette distance plus informative, car un centroïde a beaucoup de variables renseignées.
Cependant, on va retrouver les mêmes problèmes que l'algorithme classique. En effet, comme on le voit dans la figure 6, l'algorithme converge vers des pièges et l'initialisation des clusters est déterminante.

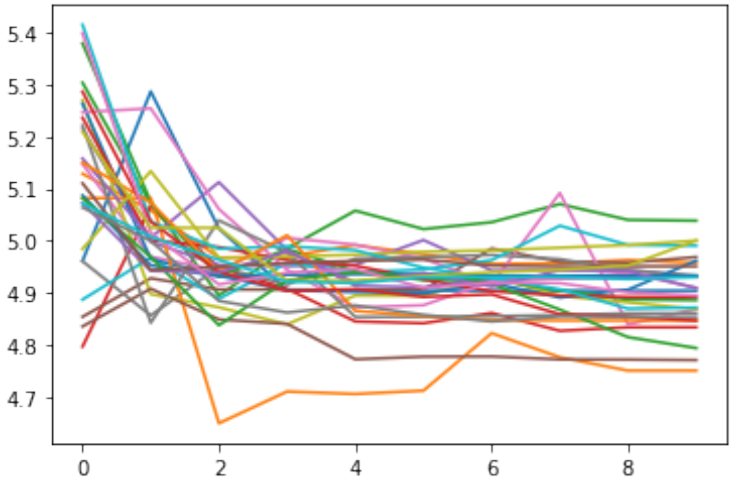

Figure 6: Logarithme de l'énergie pour différentes initialisations

4.3 Résulats

Lorsque nous avons lancé cet algorithme sur notre base de données, nous avons observé des clusters plutôt stables temporellement et des box-plots par indicateur plutôt informatifs,

deux bons critères de qualité de notre clustering.

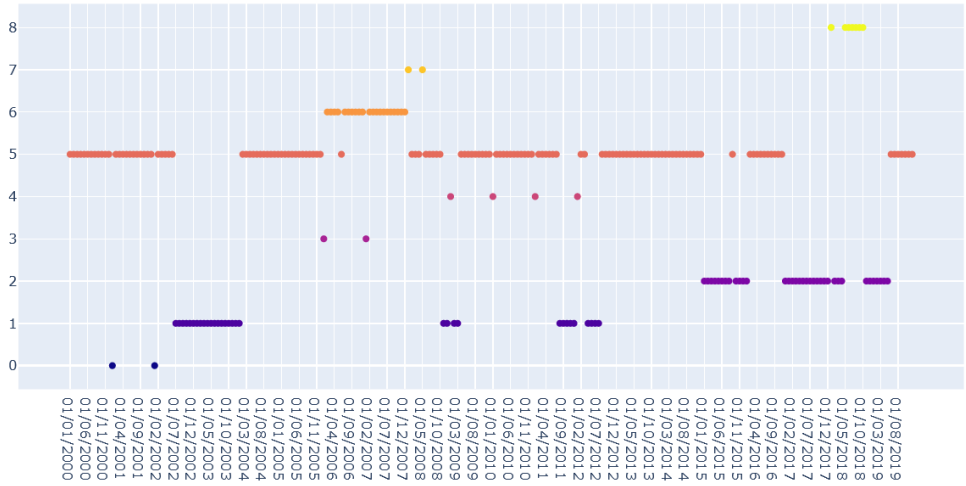

Figure 7: Cluster dans le temps

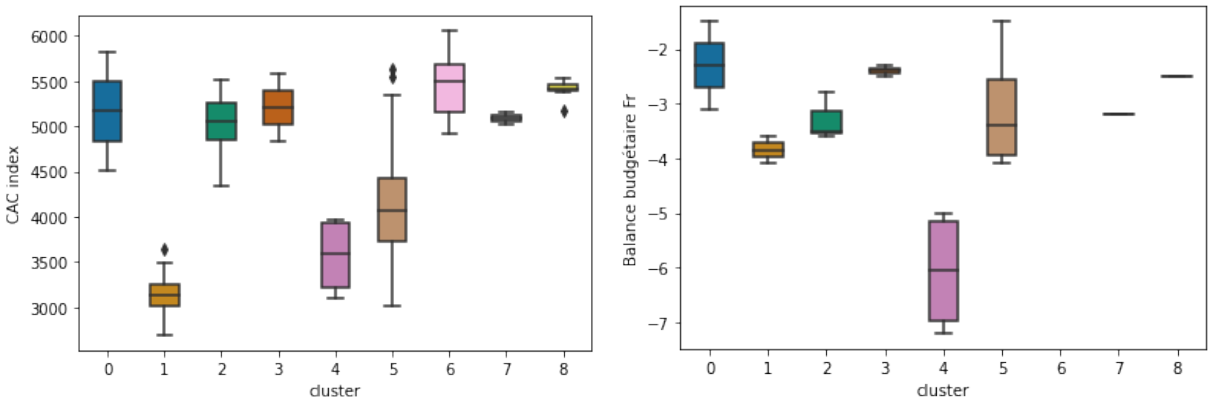

Figure 8: Comparaison (par des boxplots) des clusters respectivement par rapport au CAC 40 et à la balance budgétaire

Notre clustering a privilégié 5 clusters principaux dont :

- **Cluster 5** : situation moyenne et légèrement basse sur les indicateurs CAC et balance budgétaire, cela représente la situation la plus courante et semble être une période annonçant indifféremment une crise (cluster 1) ou une période plus favorable (cluster 2, 6 et 8).
- **Cluster 1** : marqueur d'une crise boursière.
- **Cluster 2** : période assez récente qui est caractérisée par de bons indicateurs boursiers et budgétaires stables dans le temps.
- **Clusters 6 et 8** : marqueurs d'une situation boursière exceptionnelle.

5 NIPALS (méthode 3)

5.1 Présentation de l'algorithme

L'algorithme NIPALS[2] (*nonlinear iterative partial least squares*) permet d'approcher les composantes principales d'une matrice. L'un de ses avantages est qu'il peut être adapté aux matrices à valeurs manquantes[3].

5.1.1 Méthode générale

Soient une matrice X de dimension $n \times p$ et $\mathcal{M}_k$ l'ensemble des matrices réelles de rang k et de dimension $n \times k$. On veut calculer de manière récursive les k premières composantes principales de X, notées Y_k sous forme matricielle. Si on note par ailleurs V_k la matrice des k axes principaux, les deux matrices vérifient

$$(Y_k, V_k) \in \arg \min_{A,B \in \mathcal{M}_k} \|Z - AB^t\|^2$$

où $Z = (z_i)_{1 \leq i \leq n}$ est la matrice des données centrées et réduites. Pour obtenir (Y_1, V_1), on procède comme suit:

1. Initialiser y_1 à n'importe quelle valeur (choisir une valeur approchée du vecteur propre accélère la convergence, mais n'est pas nécessaire).

2. Jusqu'à convergence, répéter les deux étapes suivantes :

   (a) Fixer y_1 et calculer v_1 comme suit:

   $$v_{1,j} = \frac{\sum_{i=1}^{n} z_{ij} y_{1,i}}{\sum_{i=1}^{n} y_{1i}^2}$$

   Si une valeur de z_j est manquante, on la passe dans le calcul de la somme (on ne somme qu'avec les données disponibles). On normalise ensuite v_1.

   (b) Fixer v_1 et calculer y_1 comme suit:

   $$y_{1,i} = \frac{\sum_{j=1}^{p} z_{ij} v_{1,j}}{\sum_{j=1}^{p} v_{1j}^2}$$

   On traite les valeurs manquantes de z_i comme ci-dessus. On normalise ensuite y_1.

Pour obtenir la deuxième composante de X, on répète les étapes ci-dessus en remplaçant Z par $Z - y_1 v_1^t$, y_1 par y_2 et v_1 par v_2. Les autres composantes s'obtiennent de la même façon, cependant on se restreint ici aux deux premières composantes.

---

[2]Wold, H. (1966) *Estimation of principal components and related models by iterative least squares.* In Multivariate Analysis (Ed., P.R. Krishnaiah), Academic Press, NY, 391-420.

[3]Martens, Harald, and Magni Martens. 2001. *Multivariate Analysis of Quality: An Introduction.* J.Wiley & Sons.

Remarque : dans l'étape 2, on fixe en général un nombre maximal d'itérations au bout duquel, si la convergence n'est pas encore assurée, on arrête quand même le calcul pour passer à la suite. Cela permet d'éviter une trop grande variabilité du temps d'exécution.

5.1.2 Modifications de l'algorithme

Premièrement, on adapte l'algorithme aux valeurs manquantes dans X. Si la valeur (i, k) est manquante, alors on passe les éléments d'indice i dans le calcul du chargement p_h, ainsi que les éléments d'indice k dans le calcul du score t_h. Plus il y a de valeurs manquantes et plus l'erreur d'approximation de T et P est grande.

Deuxièmement, la méthode numérique proposée accumule très rapidement des erreurs de calcul lors d'opérations sur les flottants. Les composantes principales calculées ne sont souvent pas orthogonales, donc on les orthogonalise à chaque itération h par la méthode de Gram-Schmidt.

Ces deux modifications de l'algorithme NIPALS sont programmées dans la fonction "nipals" du package "nipals" en R.

5.2 Lien avec le projet

Il y a plusieurs avantages à l'utilisation de l'algorithme NIPALS dans le cadre de ce projet. Il permet de calculer les premières composantes principales d'une matrice avec des données manquantes. De plus, la modification de l'algorithme basique pour s'adapter aux données manquantes est directe et cette version modifiée est déjà programmée dans plusieurs langages statistiques (R, SAS). Enfin, une spécificité de cette approche est qu'elle n'est pas une méthode d'imputation des données manquantes: on se contente ici de "passer" les données manquantes rencontrées.

Avant de présenter les résultats de cette approche, on en constate quelques limites. Notamment, lorsqu'il manque beaucoup de valeurs dans la matrice X, l'approche perd de son intérêt, car les composantes principales calculées sont très différentes des composantes de la matrice complète, voire ne convergent pas. Par ailleurs, l'observation que les composantes générées par l'algorithme ne sont pas nécessairement orthogonales montre que son implémentation génère déjà beaucoup d'erreurs de calcul sur les chargements et les scores.

5.3 Clustering et caractérisation des clusters

Afin de faciliter la comparaison avec les autres méthodes, on reprend la base de données restreinte à la France, entre 2000 et 2020 et aux variables jugées significatives sur le plan macroéconomique. On emploie de nouveau l'algorithme *k-means* sur les données projetées sur les deux premières composantes principales obtenues par l'algorithme NIPALS et on reprend les centroïdes initiaux de la méthode précédente, ce qui nous permet de garder la même numérotation des clusters.

On s'intéresse à la distribution des variables retenues dans chaque cluster. On constate que le profil économique des clusters est globalement similaire à celui de ceux obtenus par l'approche d'ACP à valeurs manquantes développée par François Husson et Jolie Josse. La grande différence est que, dans certains clusters, plusieurs variables sont très peu ou pas du tout renseignées, n'y étant pas été imputées. La caractérisation économique est donc un peu plus délicate, car on ne peut plus caractériser un cluster par une ou deux variables comme dans le cas précédent.

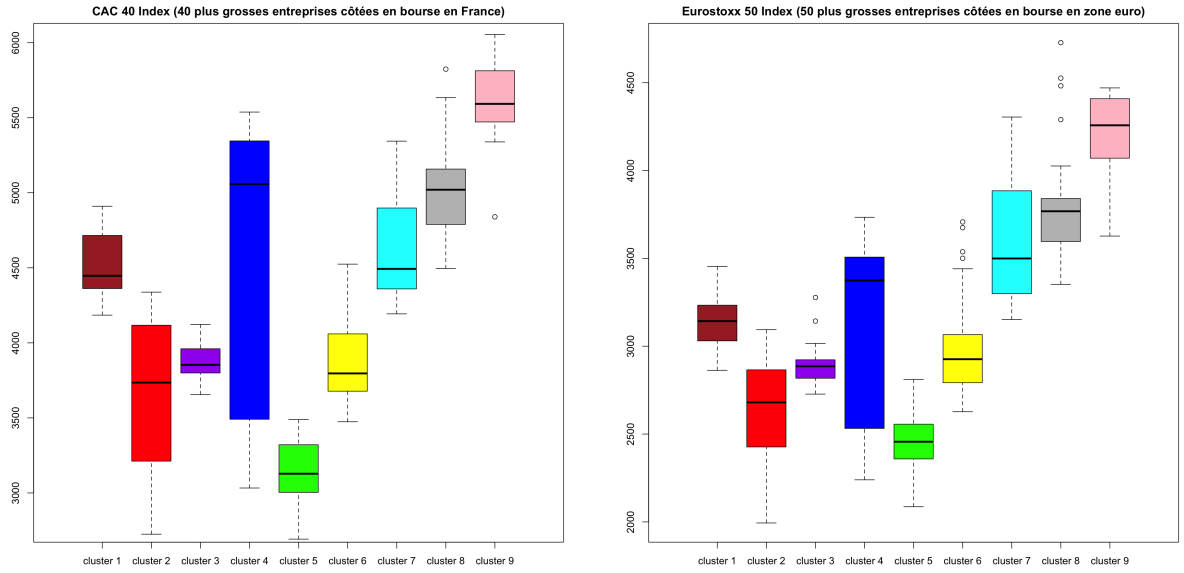

Figure 9: La santé boursière dans chaque cluster (NIPALS)

Enfin, on vérifie la cohérence du clustering avec une frise chronologique. En effet, comme les clusters correspondent à des états économiques différents, il serait surprenant d'observer un mois isolé appartenant à un cluster différent des mois autour de lui, surtout pendant une longue période de stabilité économique.

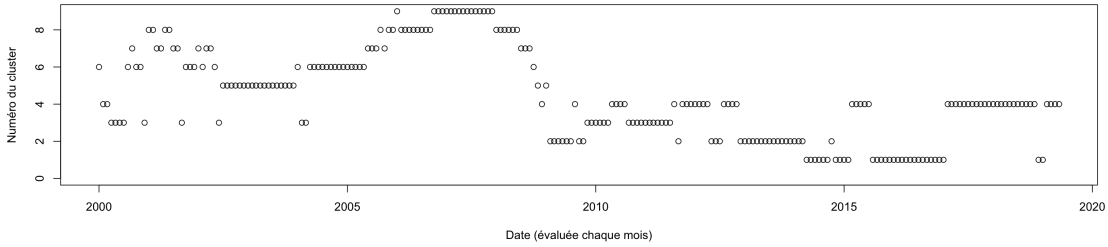

Figure 10: État économique par date (NIPALS/k-means)

5.4 Conclusion

L'algorithme NIPALS nous fournit une méthode numérique de calcul approché des composantes principales de notre base de données, qui a l'avantage d'être relativement simple et facile à adapter aux données manquantes. Elle ne fait pas d'imputation et se contente de "passer" les valeurs non-renseignées lorsqu'elles apparaissent dans les calculs.

Après clusterisation des données projetées sur les deux premières composantes obtenues, on constate que certaines variables d'intérêt sont très peu, voire pas du tout, renseignées dans plusieurs clusters. Par exemple, la part des actifs des grandes institutions financières sur le total des actifs consolidés en France n'est pas renseigné dans les clusters 6,7,8 et 9. Cela rend l'interprétation économique plus délicate, car elle nécessite la caractérisation de chaque cluster par un ensemble de variables.

Remarquons enfin que cette méthode donne les résultats les plus généraux sur les clusters, car elle ne crée aucune valeur artificielle dans la base de données. Elle peut ainsi servir pour vérifier la cohérence des imputations faites par d'autres méthodes.

6 KNN-based Missing Value Imputation (méthode 4)

6.1 Présentation de l'algorithme

Cet algorithme nous permet de compléter les données manquantes au moyen de l'étude des plus proches voisins. C'est un algorithme de clustering non supervisé. Il initialise les centroïdes de manière aléatoire, et crée des clusters en étudiant la distance (euclidienne) de chaque observation à chaque centroïde. Là encore, le nombre de centroïdes est initialisé à 9, comme la banque de France nous a suggéré, ce qui correspond au nombre d'états sous-jacents du système économique et financier. Par conséquent, chaque valeur manquante est remplacée par la moyenne (nous avons donné le même poids à chacune des variables) des valeurs des observations dans le cluster. Par exemple si X_1, X_2 et X_3 sont dans le même cluster, que le nombre de voisin est fixé à deux et qu'une coordonnée de X_1 n'est pas renseignée, alors elle sera remplacée par la moyenne de la même coordonnée entre X_2 et X_3 (si ladite coordonnée n'est pas manquante sur ces deux vecteurs) puis imputée à X_1.

6.2 Avantages de l'algorithme

Un avantage de cet algorithme est que nous n'avons pas besoin de supposer l'indépendance entre toutes les variables, mais seulement entre les clusters. En effet, il est plus aisé de supposer l'indépendance économique entre Février 2009 et Octobre 2018 qu'entre Février 2009 et Janvier 2009. Sa force est donc d'imputer des valeurs manquantes en fonction d'observations qui leur ressemblent fortement, ce qui est beaucoup plus fort qu'un algorithme qui prendrait simplement la moyenne de toutes les observations et qui n'aurait pas réellement de sens économique. Enfin, ici les hypothèses sont plus souples car la limite d'indépendance entre les observations peut être matérialisée par les frontières entre clusters.

6.3 Imputation de l'algorithme

Tout d'abord il faut savoir que la disparité des valeurs manquantes est importante selon les colonnes. Certaines n'ont que 7% de valeurs manquantes, et d'autres 96%.

Afin de combler le vide laissé par ces valeurs manquantes, nous effectuons donc une approche k-nn, avec un nombre maximal de missing value fixé à 96%. En effet, la commande KNNImputer fixe initialement un nombre maximal de missing values à 80%. Nous avons donc dû modifier ce paramètre. Nous fixons des poids uniformes pour chacune des variables, un nombre de voisins fixé à 2, et la distance euclidienne traditionnelle. La valeur manquante sera ainsi imputée par la moyenne des valeurs des deux plus proches voisins comme nous l'avons dit ci-dessus.

Maintenant que la base de données est complétée, nous allons utiliser un algorithme k-means pour le clustering.

Pour remarquer plus facilement les différences entre clusters, nous nous sommes basés sur les indices du CAC40 (variable financière) et de la balance budgétaire (variable bancaire) que nous avons comparés entre les 9 clusters. L'indice du CAC40 étant fortement corrélé à la santé économique générale et étant un indice macro-économique important, nous avons estimé qu'il pouvait être représentatif de la santé économique, de même que la balance budgétaire.

En résumé, nous avons effectué un premier clustering (KNN missing values) pour compléter notre base. Puis, nous avons effecfué un deuxième clustering (k-means) pour regrouper en clusters de dates nos vecteurs précédemment complétés.

6.4 Clustering et caractérisation des clusters

6.4.1 Caractérisation financière des clusters

Après avoir observé les boxplots de nos 9 clusters, nous remarquons des fortes disparités entre ces derniers. En effet, dans certains clusters le CAC40 dépasse les 5000 points, et dans d'autres il atteint seulement les 3000 points. Par conséquent, certains clusters témoignent d'une bonne santé financière alors que d'autres reflètent au contraire une mauvaise santé financière. Maintenant nous allons nous demander si ces différents clusters sont le reflet d'états financiers réels et si ceux-ci correspondent à des périodes bien particulières.

En observant les boxplots des clusters, et en les comparant à l'aide de l'indice du CAC40, nous pouvons les classer de la plus bonne, à la moins bonne santé financière dans l'ordre suivant (cf. figure 11) :

$$2 - 4 - 5 - 6 - 1 - 0 - 3 - 7 - 8$$

En organisant notre dataframe par ordre chronologique, nous remarquons plusieurs choses. D'abord, le cluster 8, qui reflète la plus mauvaise santé financière, est majoritairement présent en 2004 et également après la crise de 2008 : cela semble cohérent. Durant le laps de temps post-crise, entre 2009 et fin 2013 les clusters majoritairement présents sont les clusters 3, 7 et 8 qui sont les 3 clusters reflétant une mauvaise santé financière. Ensuite, en 2014, le cluster majoritairement présent est le 1, qui est un cluster "seuil" et représente une sorte

de "médiane" des clusters ainsi qu'une santé économique moyenne. Enfin, le cluster 4 est majoritairement présent entre 2016 et 2019, ce qui témoigne d'une bonne santé financière.

Nous voyons donc que nos clusters sont le reflet de l'évolution réelle des marchés financiers, ce qui prouve que notre algorithme est cohérent, et celui-ci aurait encore mieux fonctionné si certaines variables n'avaient pas 96% de données manquantes car ce clustering a été réalisé sur la base de données remplie fournie par notre algorithme KNN missing values.

6.4.2 Caractérisation des clusters du point de vue du secteur bancaire

De la même manière qu'au paragraphe précédent, nous allons caractériser les clusters du point de vue de la santé bancaire à l'aide de la variable "Balance Budgétaire". Nous remarquons tout d'abord que certaines boxplots sont assez plates. En comparant les clusters de la meilleure santé économique à la moins bonne nous trouvons ce classement : 2-4-8-5-6-0-1-3-7. Ainsi, ce classement est quasiment identique à celui du paragraphe précédent mis à part le cluster 8, qui est cette-fois ci témoin d'une bonne santé économique bancaire.

6.4.3 Résultats

Les deux graphiques ci-dessous nous montrent les boxplots avant application de l'algorithme KNN afin que les missing values ne viennent pas biaiser leur étude. Pour cette raison certaines boxplots sont assez plates, à l'instar des clusters 5 et 6 pour la balance budgétaire. Cela signifie soit que nos clusters sont très homogènes soit que le nombre d'observations à l'intérieur du cluster est relativement faible, car la variance est quasi-nulle.

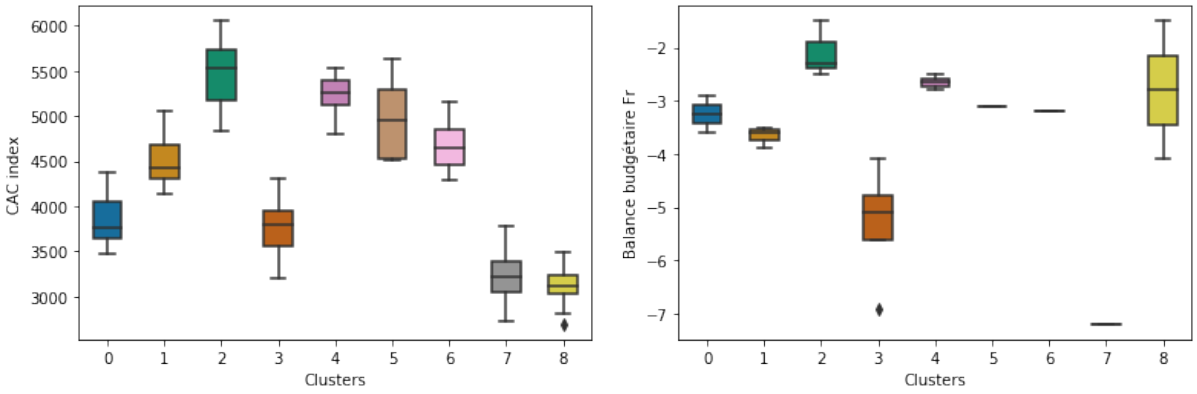

Figure 11: Comparaison des clusters par rapport au CAC 40 et à la balance budgétaire respectivement

Enfin, sur le graphique de clusters de dates ci-dessous, nous constatons une certaine régularité dans l'occurence des clusters. Cela nous montre d'une part, que la reprise économique a été relativement lente, et d'autre part, que notre algorithme est cohérent.

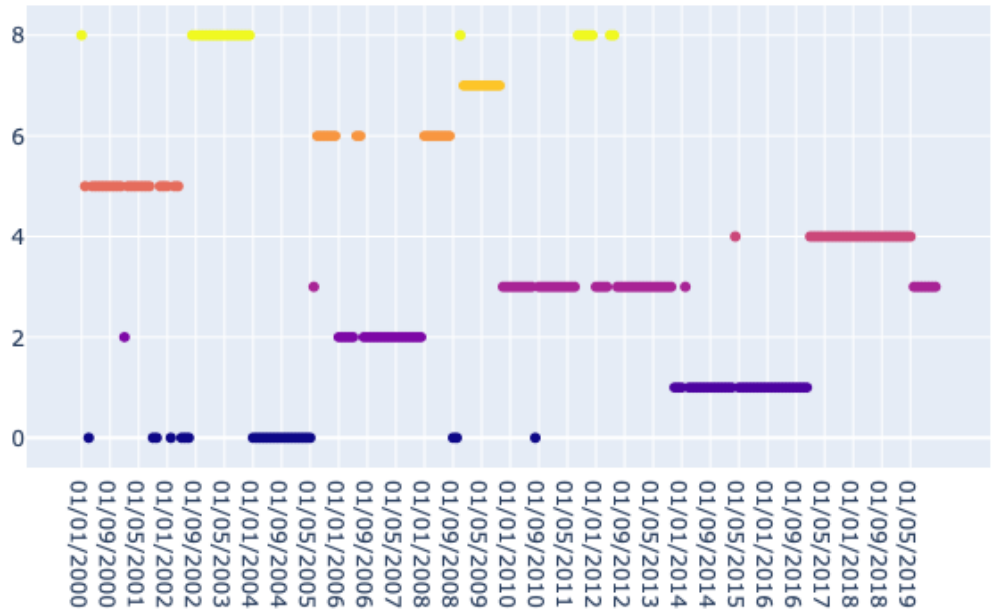

Figure 12: Clusters dans le temps

7 Conclusion

Lors de ce projet, quatres méthodes ont donc été mises en place pour faire du clustering sur une base de données à valeurs manquantes. On peut distinguer deux types de méthodes : d'un côté celles qui utilisent d'abord un algorithme pour affecter des valeurs aux données manquantes pour ensuite clusteriser sur la base de données complétée, et de l'autre celles qui font directement du clustering sur la base de données incomplète (i.e. sans faire l'étape préliminaire consistant à affecter des valeurs aux données manquantes).

Chaque méthode employée ouvre sur la formation de clusters cohérents du point de vue de la stabilité des états économiques à travers le temps (cf. figures 4,7,10,12).

Les clusters formés dans chaque méthode représentaient des états économiques uniques (pas de clusters similaires) ce qui est satisfaisant. En effet, comme on a pu le voir sur les figures 2,3,8,9,11, les boxplots de chaque cluster sont resserrées. Cela montre qu'ils représentent des états économiques particuliers. De plus, ces boxplots ne se chevauchent pas trop entre elles (sauf avec les NIPALS, en partie 5), ce qui facilite la distinction des clusters par une ou deux variables.

Dans les quatre méthodes, nous avons attribué à chaque cluster un état financier particulier (comparaison entre les boxplots de chaque cluster sur la variable CAC 40 et Eurostoxx, cf. figures 2,8,9,...), un état bancaire particulier (comparaison entre les boxplots de chaque cluster sur la variable "actifs bancaires consolidés", cf. figures 3,15,...), ainsi qu'un état d'endettement particulier (comparaison entre les boxplots de chaque cluster sur la variable "balance budgétaire", cf. figures 8,11,15).

Finalement, dans la majorité des cas, des clusters de dates assez stables apparaissent bien. On distingue d'une part des périodes de moins bonne santé économique (après 2008) et celles de reprise de l'activité économique (après 2016).

8 Annexe

8.1 Interprétation économique des variables retenues dans la base de données (nécessaire pour comprendre le projet)

Bien qu'on compte ci-dessous que treize points d'explication, on a retenu 17 variables pour la base de données (le point d'explication numéro 12 regroupe trois variables à savoir CAC 40 financier, CAC 40 biens de consommation et CAC 40 pétrole/gaz, idem pour le point d'explication numéro 11 qui regroupe deux variables à savoir CAC 40 total return index et Eurostoxx total return index)

1. **Rapport crédits/dépôts :** comme l'argent que déposent les clients (qui détiennent un compte à la banque) est donné à d'autres clients dans le cadre de l'accord de crédit, cela pose un problème économique aux banques si les crédits sont plus importants que les dépôts d'où l'intérêt de la variable du rapport crédits/dépôts.

2. **Rapport cours/valeur comptable des banques :** un fort (resp. faible) ratio cours/valeur comptable peut signifier que l'action est sur-évaluée (resp. sous évaluée). Il peut aussi rendre compte d'un dysfonctionnement fondamental au sein de l'entreprise.

3. **Rentabilité des capitaux propres :** les capitaux propres sont les ressources financières que possède l'entreprise hors dette. Ces capitaux constituent une ressource stable pour l'entreprise.

4. **Actifs bancaires consolidés :** les actifs consolidés d'une société regroupent l'ensemble des actifs de la société mère et de ses filiales ce qui donne un bilan plus objectif de l'état financier de l'entreprise en question.

5. **Actifs des GSIB :** les GSIB sont les banques dites systémiques, c'est à dire les très grandes banques qui pèsent tellement dans l'économie que leur faillite provoquerait un bouleversement économique.

6. **Inflation :** l'inflation se caractérise par une augmentation durable, générale, et auto-entretenue des prix des biens et des services.

7. **Déficit budgétaire :** le déficit budgétaire est la situation dans laquelle les recettes de l'État (hors remboursement d'emprunt) sont inférieures à ses dépenses (hors emprunt) au cours d'une année.

8. **Solde des transactions courantes :** la balance des transactions courantes regroupe les échanges de marchandises, les échanges de services, les flux de revenus et les transferts courants entre la France et le reste du monde.

9. **Indice du CAC 40 :** le CAC 40 regroupe les 40 plus grosses entreprises françaises côtées en bourse.

10. **Indice Eurostoxx :** l'Eurostoxx 50 regroupe les 50 plus grosses entreprises côtées en bourse en zone euro.

11. **Indice du rendement total (CAC 40 et Eurostoxx) :** le total return index est un index qui considère que les bénéfices, les dividendes et les surplus financiers seront réinvestis dans le temps, et qui mesure ainsi les plus-values réalisées par les actions. Un indice return est considéré par les investisseurs comme plus précis que ceux qui excluent les distributions, car il fournit une meilleure représentation de la performance financière des entreprises sous-jacentes.

12. **CAC 40 financier, biens de consommation et pétrole/gaz :** les entreprises du CAC 40 respectivement dans le secteur financier, des biens de consommation et du gaz/pétrole.

13. **Levier bancaire :** l'effet de levier désigne le recours à l'emprunt afin d'augmenter son exposition sur les marchés financiers. Le ratio de l'effet de levier nous indique la proportion d'emprunts possible à partir du montant de nos capitaux propres. L'effet de levier offre un résultat à double tranchant puisque l'investisseur augmente du même coup ses gains potentiels et ses pertes potentielles.

8.2 Caractérisation des clusters

8.2.1 Caractérisation des clusters avec la méthode de clustering k-means (sur la base de données complétée par l'ACP à valeurs manquantes)

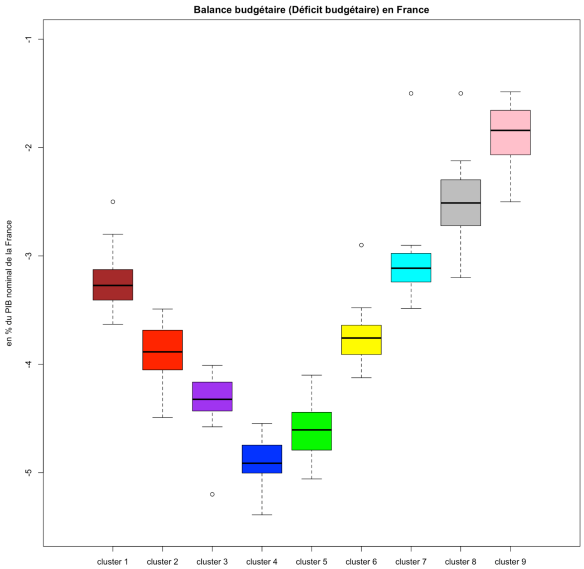

Figure 13: ACP à valeurs manquantes/k-means : Comparaison (par des boxplots) des 9 clusters sur le plan de la balance budgétaire

8.2.2 Caractérisation des clusters avec la méthode de clustering DBSCAN (sur la base de données complétée par l'ACP à valeurs manquantes)

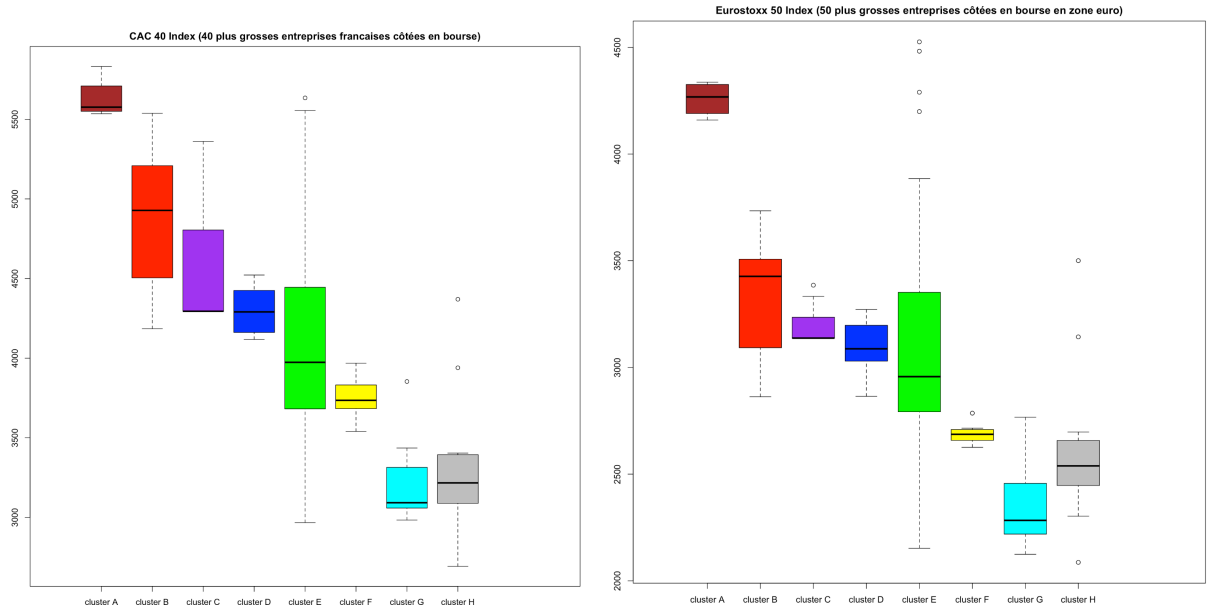

Figure 14: ACP à valeurs manquantes/DBSCAN : Comparaison (par des boxplots) des 9 clusters par rapport aux indices boursiers français et européens

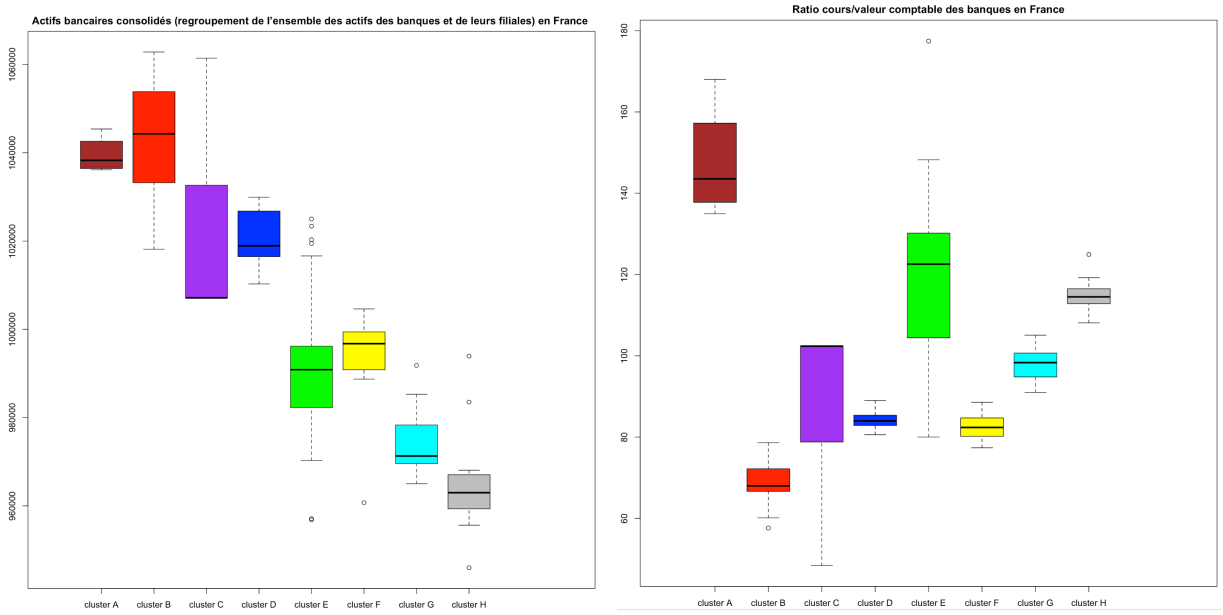

Figure 15: ACP à valeurs manquantes/DBSCAN : Comparaison (par des boxplots) des 9 clusters par rapport à la santé du secteur bancaire

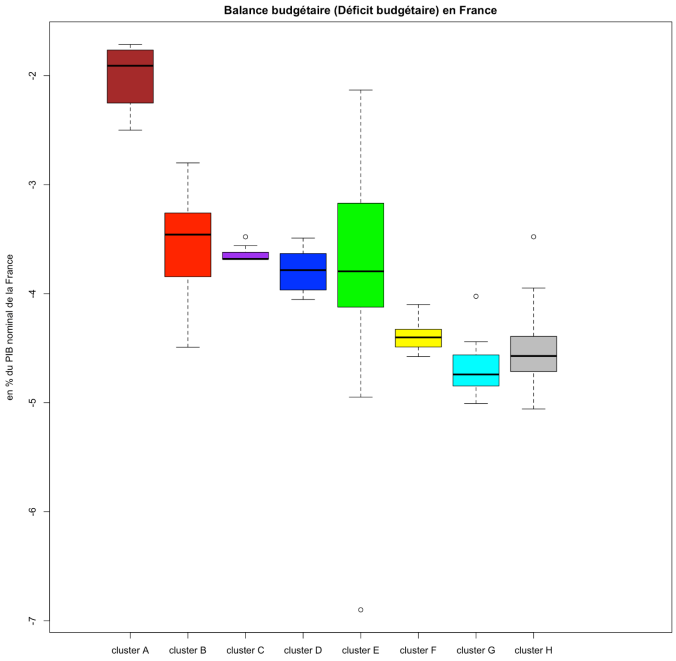

Figure 16: ACP à valeurs manquantes/DBSCAN : Comparaison (par des boxplots) des 9 clusters sur le plan de la balance budgétaire

8.2.3 Caractérisation des clusters sur la base de données initiale avant d'être complétée par l'algorithme KNN-based Missing Value Imputation

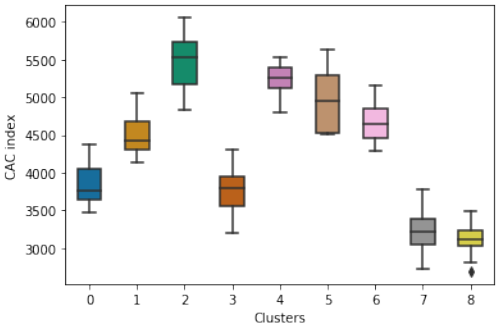

Figure 17: Clusters pour indice du CAC40

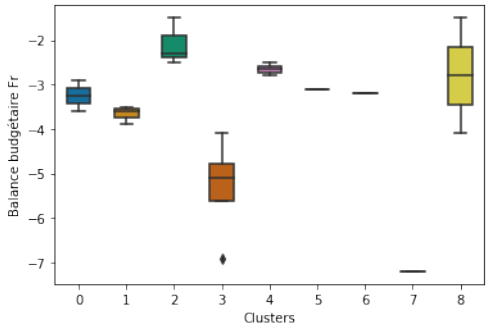

Figure 18: Clusters pour la balance budgétaire

8.3 Corrélations entre les variables

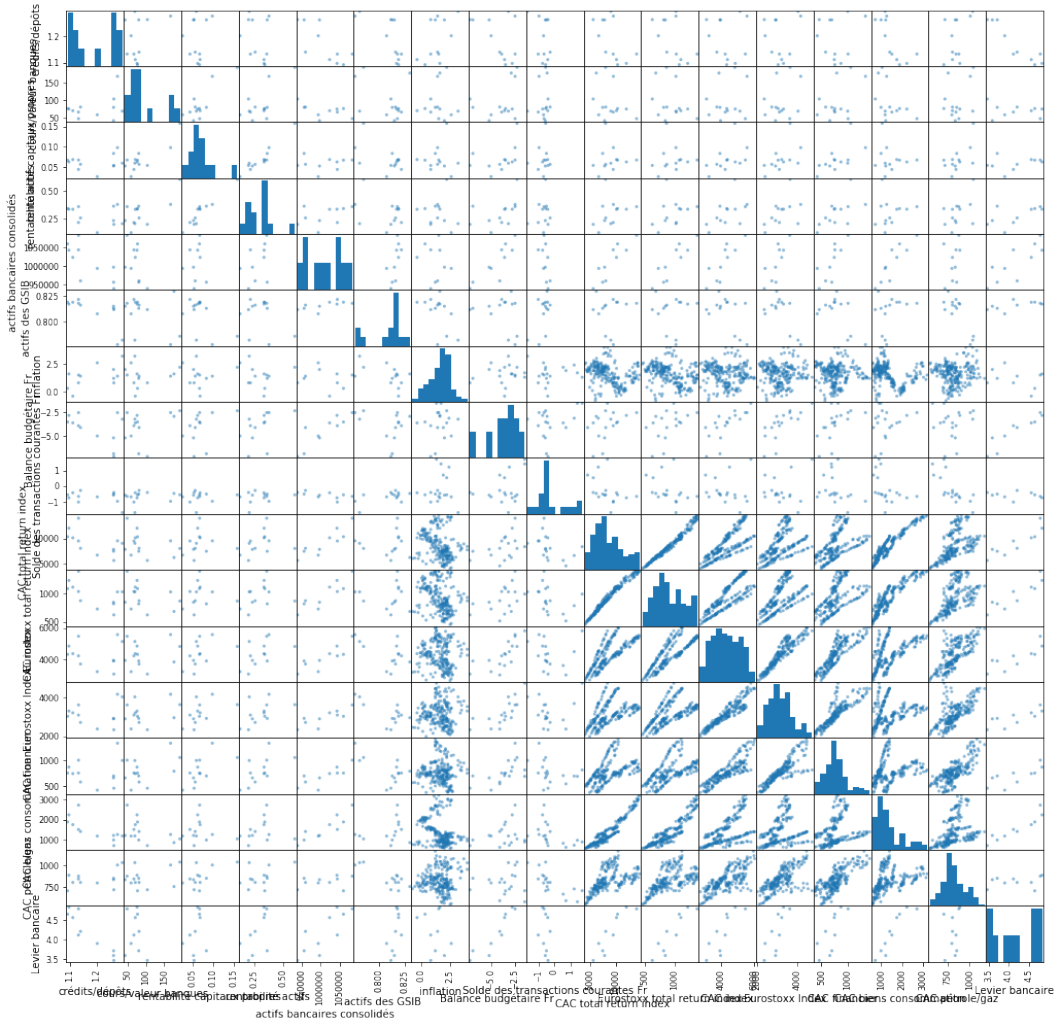

Figure 19: Corrélation entre les différentes variables de notre base de données

On remarque que les variables bancaires (situées en majorité dans les colonnes de gauche) ne sont, en majorité, ni fortement corrélées entre elles (forte dispersion en haut à gauche) ni avec les variables financières (forte dispersion en bas à gauche). En revanche, les variables financières sont fortement corrélées entre elles (forte corrélation en bas à droite). Cela nous apprend une chose importante : les données bancaires peuvent refléter une toute autre situation économique que les données financières.

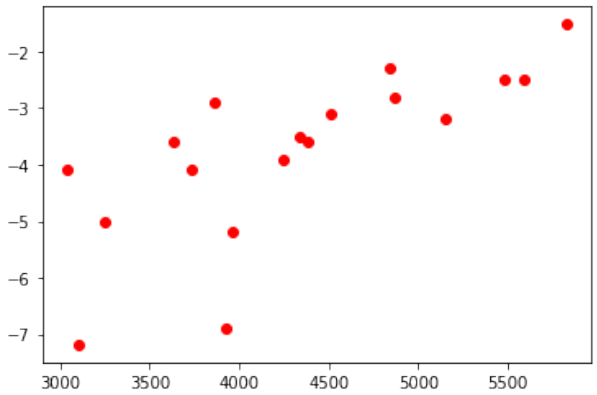

Figure 20: Corrélation entre l'indice du CAC40 et la balance budgétaire

Voici un exemple pour illustrer notre propos sur la corrélation entre les variables. Ici, les points semblent assez dispersés. Cela explique que les classements des clusters dans les boxplots pour l'algorithme KNN missing values ne soient pas exactement les mêmes pour les deux variables.