ENSAE ParisTech

Compte-rendu de Statistiques Appliquées

# Étude et prédiction des défauts sur le marché des couvertures de défaillance

HELLEBORECAPITAL

Marianne Sorba
Apolline Bouissieres
Jingmei Jiang
Xingyuan Xue

*Encadrant :* Gautier Marti
*Correspondant :* Caroline Hillairet

2 Novembre 2016 - 3 Mai 2017

# Table des matières

# Introduction

## 1.1 Hellebore Capital

Dans le cadre de notre projet de Statistiques Appliquées, nous avons travaillé pour un fond d'investissement : Hellebore Capital (HC). HC est une société de gestion spécialisée dans les arbitrages de dérivés de crédit qui travaille notamment sur le marché des Credit Default Swaps (ou couvertures de défaillance/CDS). Elle est propriétaire d'une base de données de messages (mails) concernant ces produits, qu'elle transforme ensuite en séries chronologiques exploitables. Les messages contiennent le plus souvent un bid et un ask fixés par l'intervenant pour un ticker donné ou des commentaires sur le marché. Hellebore Capital reçoit en tout plus de 20,000 messages par jour, ce qui fait un total de 40,000,000 messages pour la période étudiée ci-après.

## 1.2 Zoom sur les Credit Default Swaps

**Définition et contexte :** Un Crédit Default Swap est un dérivé de crédit. Il permet de protéger son détenteur contre un événement de crédit sur l'actif sous-jacent. Lorsqu'une entreprise ou un état veut se financer, il émet sur les marchés financiers toutes sortes de titres de dettes. L'acheteur de ces titres de dettes, lui, s'expose directement à un risque de défaut de l'entité en question. Le cœur du problème est que ces titres de dettes sont souvent émis à de longues échéances et que l'acheteur peut avoir beaucoup de difficultés à évaluer la capacité de l'entité à le rembourser. Dans ce cas, il peut également acheter un CDS à un tiers. En échange du paiement d'une prime (spread) versée à ce tiers, ce dernier s'engage à supporter le risque de crédit que supportait le détenteur de la dette en le remboursant en cas de défaut du sous-jacent. De ce fait, le prix du CDS associé à une entreprise/état est un indicateur de la confiance qui lui est accordée par les marchés : plus le prix d'un CDS est important, plus l'actif sous-jacent est considéré comme risqué par les marchés financiers. Cependant, il n'est pas nécessaire de détenir de la dette de l'actif sous-jacent pour pouvoir acheter un CDS sur cet actif (sauf pour la dette souveraine depuis la crise grecque [1]). Ils servent donc d'assurance et de moyen de spéculation. Il est important de préciser que les tiers qui assurent les acheteurs de CDS sont des institutions financières, ce sont donc elles qui supportent les risques de contreparties alors que ces crédits ne sont pas inscrits à leur bilan. Ceci a posé de graves soucis pendant la crise des dettes souveraines. De plus, le marché des CDS est un marché dit de "gré à gré" ce qui signifie qu'il n'est que très peu régulé.

**Histoire :** Le premier CDS a été créé par JP Morgan pour la compagnie pétrolière américaine Exxon. En 1994, à cause de l'échouage d'un des bateaux de Exxon, cette dernière souhaitait une ligne de crédit de $4.8$ milliards de dollars. Or, compte tenu des législations sur les fonds propres, le coup de refinancement de la banque aurait été beaucoup trop important si elle acceptait de détenir la totalité du crédit. Les CDS sont alors créés et généralisés par cette même banque. Les CDS sont considérés comme en partie responsable de la faillite de Lehman Brothers et AIG lors de la crise financière de 2008. En effet, ils permettent de passer hors du bilan des banques les crédits accordés sous forme de CDS [2].

## 1.3 Étude : détection des comportements anormaux et prédictions des défauts

Notre sujet est le suivant : Étude et prédiction des défauts sur le marché de gré à gré des couvertures de défaillance. L'intérêt pour Hellebore Capital est de leur permettre d'orienter leurs achats de CDS (les aider à parier sur les CDS qui risquent de faire défaut [3]). Pour répondre à cette problématique, nous avions à notre disposition une base contenant les séries temporelles des prix des CDS de nombreuses entreprises, regroupées par maturité (1, 3, 5, 7 ou 10 ans). Le cœur de notre travail repose sur un clustering fait au préalable : Les CDS sont regroupés en groupes distincts d'actifs qui se comportent de façon similaire, de manière à ce que l'analyse d'un cluster permette de détecter lorsqu'un CDS "s'éloigne" des autres, ce qui pourrait être un signe de comportement anormal. L'idée générale est de construire des indicateurs détectant ces signes et de s'en servir comme variables explicatives dans la régression logit pour la variable expliquée "a effectivement fait défaut". Nous en expliquerons le principe dans la suite mais nous n'avons pas eu le temps de la finir et de la tester proprement. Nous avons développé différentes méthodes permettant de quantifier l'éloignement d'un CDS au sein de son cluster, méthodes que nous allons décrire tout au long de ce rapport.

# Description de la base de données

Notre base de données est constituée des prix des CDS sur 563 entreprises (ou états). Cette base de données a été créée par Hellebore Capital à partir d'une database de messages concernant ce produit, qu'ils ont ensuite transformé en séries chronologiques exploitables. Pour chaque signature et chaque maturité, on dispose de la série chronologique du prix du CDS correspondant prise entre le 3 janvier 2006 et le 28 décembre 2016. Nous disposons des données des 5 premiers jours de la semaine. Les prix sont représentés en point de base (bps) qui correspond à un centième de point de pourcentage : Un CDS coûtant 100 bps signifie que l'acheteur s'engage à payer 1% de la valeur nominale de l'actif correspondant chaque année jusqu'à maturité en échange de la protection. Ci-dessous un exemple de l'évolution des prix de trois CDS de maturité 5 ans (Nestlé, Carlsberg et Tesco).

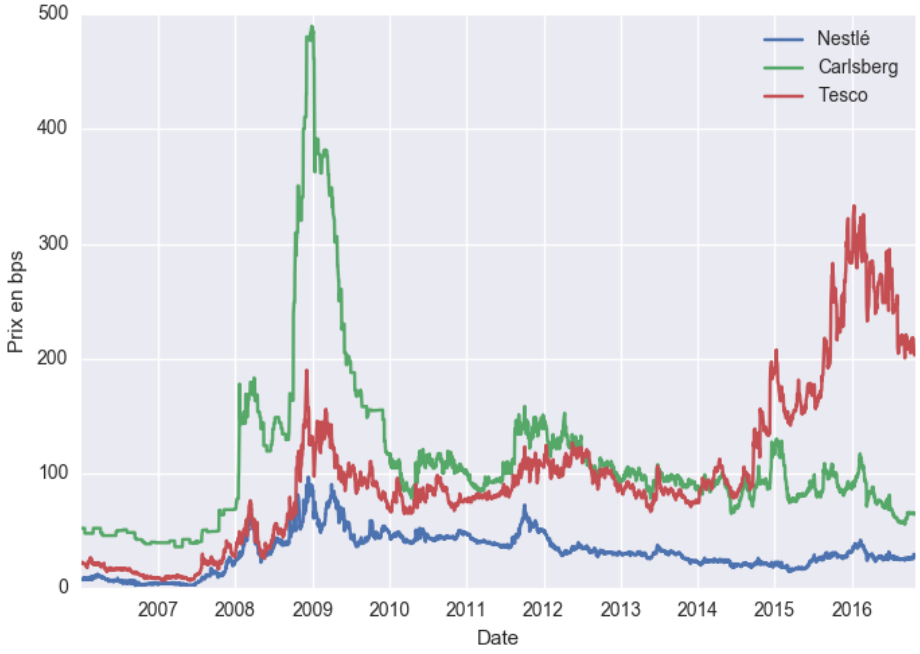



FIGURE 2.1 – Évolution des prix des CDS de Nestlé, Carlsberg et Tesco

**Répartition du prix moyen des séries temporelles en fonction de leur maturité** Intéressons nous tout d'abord à la distribution des prix de nos séries temporelles selon leur maturité. Pour chaque CDS, nous avons calculé la moyenne de son prix pour toute la période afin de mieux comprendre comment les prix sont distribués sur le marché des couvertures de défaillance. Nous nous sommes ensuite intéressées à la distribution de ces moyennes en fonction des maturités correspondantes. Vous trouverez ci-dessous le tableau récapitulatif.

| Maturité | Effectif | Minimum | Moyenne | $1^{er}$ quartile | Médiane | $3^{eme}$ quartile | Maximum |
|---|---|---|---|---|---|---|---|
| 1 ans | 563 | 8.883 | 113.9 | 30.17 | 54.72 | 110.9 | 1975 |
| 3 ans | 552 | 16.41 | 158.2 | 50.57 | 83.86 | 172.69 | 1910.24 |
| 5 ans | 563 | 25.4 | 193.8 | 72.46 | 113.23 | 211.27 | 1914.37 |
| 7 ans | 552 | 32.63 | 211.7 | 86.37 | 129.45 | 239.56 | 1863.40 |
| 10 ans | 552 | 11.11 | 146.8 | 42.96 | 75.70 | 144.89 | 1533 |

TABLE 2.1 – Répartition du prix moyen des séries temporelles en fonction de leur échéance

On observe que les prix moyens croissent avec la maturité. En règle générale, il est beaucoup plus risqué de parier à long terme sur le non défaut d'un CDS qu'à court terme. Les CDS de grandes maturités sont donc considérés comme plus risqués, et sont donc plus chers. Les prix sont très étendus, allant de 8.8 bps à 1975 bps pour les CDS de maturité 1 an. Pour chaque maturité, le prix moyen des CDS est environ deux fois plus important que son prix médian : les prix sont répartis de façon asymétrique vers la gauche de la distribution. Cette observation est confirmée par les diagrammes de Tukey de la distribution des prix moyens des CDS en fonction de leur maturité. Il y a présence importante d'outliers vers la droite de la distribution : Les outliers correspondent en majorité aux CDS des entreprises ayant fait défaut sur la période étudiée, et donc à des prix très élevés.

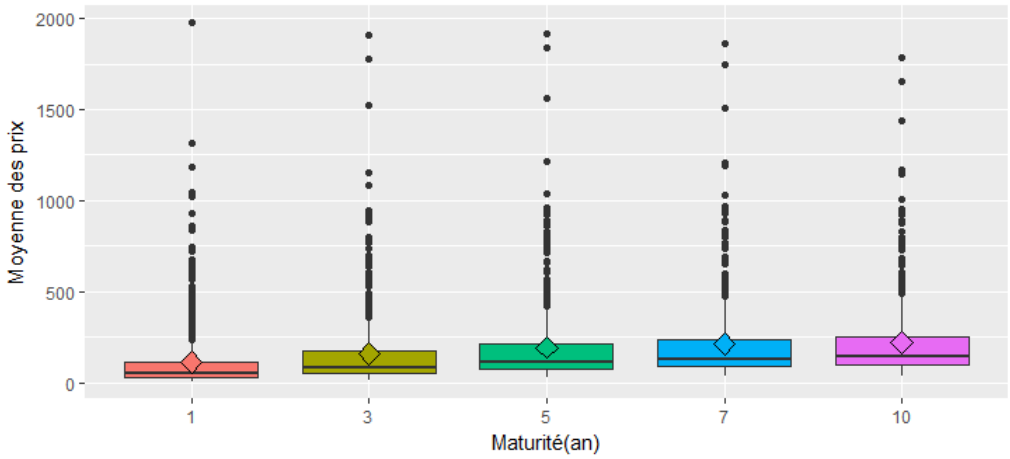



FIGURE 2.2 – Boxplots des moyennes des prix de toutes les échéances

**Statistiques descriptives des écarts-types des prix :** Intéressons nous maintenant à la distribution de l'écart-type calculé sur toute la période de nos séries temporelles selon leur maturité (correspond à la volatilité du produit financier traité). Vous trouverez ci-dessous le tableau récapitulatif.

| Maturité | Effectif | Minimum | Moyenne | $1^{er}$ quartile | Médiane | $3^{eme}$ quartile | Maximum |
|---|---|---|---|---|---|---|---|
| 1 ans | 563 | 8.883 | 180.9 | 33.1 | 68.64 | 162.31 | 2608.03 |
| 3 ans | 552 | 7.2 | 170.4 | 37.91 | 73.50 | 163.39 | 1926.8 |
| 5 ans | 563 | 9.53 | 161.7 | 41.45 | 74.77 | 156.96 | 1743.63 |
| 7 ans | 552 | 11.15 | 154.4 | 42.01 | 75.72 | 154.82 | 1631.16 |
| 10 ans | 552 | 11.11 | 146.8 | 42.96 | 75.70 | 144.89 | 1533.67 |

TABLE 2.2 – Répartition de l'écart-type des séries temporelles en fonction de leur échéance

En moyenne, les variances des CDS décroissent en fonction de leur maturité. Les CDS de petite maturité sont plus sensibles aux variations à court terme des marchés, car en cas de chocs pouvant entraîner un défaut rapide, les investisseurs parient que le CDS à 1 an va faire défaut, mais que si ça n'est pas le cas, la situation va se rétablir sur une période plus longue. Les CDS de petite maturité sont donc plus sensibles aux chocs, ce qui explique une variance plus importante.

# Le partitionnement automatique (clustering) pour constituer des benchmarks statistiques

## 3.1 Principe du partitionnement en fonction du niveau de corrélation

Monsieur Gautier Marti, notre accompagnateur pendant la durée du projet, a effectué ce clustering [4] au préalable afin de nous donner une direction pour commencer le travail de recherche sur les indicateurs. Nous allons cependant l'expliquer ici car c'est le socle sur lequel repose notre réflexion. De plus notre travail a commencé par la compréhension de ce clustering, et de ce qu'il pouvait nous apporter en regardant les corrélations au sein des clusters. Les CDS sont regroupés en clusters par la méthode de Wald. L'idée est de regrouper les actifs qui se comportent de façon similaire et qui sont sensibles aux mêmes chocs, de façon à pouvoir les comparer de manière pertinente. Par exemple, un choc sur l'offre de pétrole n'aura que très peu d'effet sur l'industrie pharmaceutique tandis qu'il en aura beaucoup plus sur le marché de l'automobile. Il serait donc judicieux de regrouper les entreprises selon leur secteur d'activité. Après analyse, c'est effectivement ce qui semble se produire avec ce clustering. Nous avons observé que les clusters regroupaient en général les entreprises selon leur région, leur taille, leur domaine d'activité et leur rating (par les agences de notation). Le clustering a été réalisé à l'aide des séries chronologiques des CDS de maturité 5 ans, qui est la maturité la plus utilisée sur le marché.

## 3.2 Construction des clusters à l'aide de la méthode de Wald

Nous allons décrire dans cette section le procédé qui a permis de construire les clusters utilisés par le suite. Soit $n$ le nombre de CDS de maturité 5 ans dont nous disposons.

1. $\forall i \in [\![1, n]\!]$, tout d'abord, il faut calculer les taux d'accroissements de la série temporelle $X^i$ en calculant la série $S^i$ des accroissements logarithmiques ( $S_t^i = \log(X_t^i) - log(X_{t-1}^i)$).
2. Il s'agit ensuite de définir une distance $d$ avec $d(S^i, S^j) = 1 - \frac{\text{cov}(S^i, S^j)}{\sigma_{S^i}\sigma_{S^j}}$
3. C'est la méthode de Ward qui a été utilisée pour effectuer une classification ascendante hiérarchique. La méthode de Ward consiste à regrouper les classes à chaque itération de façon à ce que l'augmentation de l'inertie interclasse $I_e$ soit maximum :
   Considérons $G = \left\{S^i, i \in [\![1, n]\!]\right\}$ de centre de gravité $g$ , partitionné en k classes $G_1,\ G_2,\ ..,\ G_k$ de taille $n_1,\ n_2,\ ..,\ n_k$ de centre de gravité $g_1,\ g_2,\ ..,\ g_k$. On a :
   $$I_e = \frac{1}{n}\sum_{i=1}^{k} n_i \times d^2(g_i, g)$$
   Le processus itératif s'arrête lorsque l'on a le nombre de clusters désiré. Nous disposons de 30 clusters.

## 3.3 Étude de la composition des clusters obtenus

Nous avons donc 30 clusters. Nous avons étudié leur composition. Le premier cluster est composé d'entreprises Japonaises d'industries diverses (Mitsubishi Corporation, Shimizu Corporation, Mitsui Chemicals, etc.). Les 6 clusters suivant sont également composés d'entreprises ayant leur siège social dans la même zone géographique (Australie, Inde, Chine, Corée, Europe). Le huitième cluster quant à lui regroupe toutes les multinationales dont les secteurs d'activités sont liés à l'agroalimentaire (Danone, Casino, British American Tobacco, Nestlé, Tesco, etc.). On retrouve par la suite les clusters correspondant aux secteurs de la télécommunication, du transport, de l'audiovisuel, du cinéma, de l'énergie, de la santé, de l'immobilier ainsi que le secteur bancaire. Les CDS souverains (le sous-jacent est une obligation d'Etat) sont également regroupés entre eux. Ceux de l'Europe de l'ouest composent le $11^{eme}$ cluster, ceux de l'Europe de l'Est le $12^{eme}$. On en conclu que le clustering regroupe les CDS majoritairement en fonction de leur emplacement géographique et de leur secteur d'activité. On peut donc voir clairement une segmentation des différents marchés grâce à ce clustering. Cela semble assez intuitif : les entreprises qui vendent dans un même secteur d'activité ou issue d'une même région sont soumises à des influences macroéconomiques et microéconomiques similaires, d'où

leur comportement similaire. Certains autres clusters sont moins évidents, et donc intéressants à comprendre. Ci-dessous la matrice de corrélation obtenue pour l'ensemble de nos données. Les numéros des CDS sont disposés de telle sorte que les CDS très corrélés entre eux se retrouvent côte à côte, d'où l'aspect "diagonale par bloc" de la matrice. Par exemple, les CDS appartenant au premier cluster sont numérotés de 1 à 38.

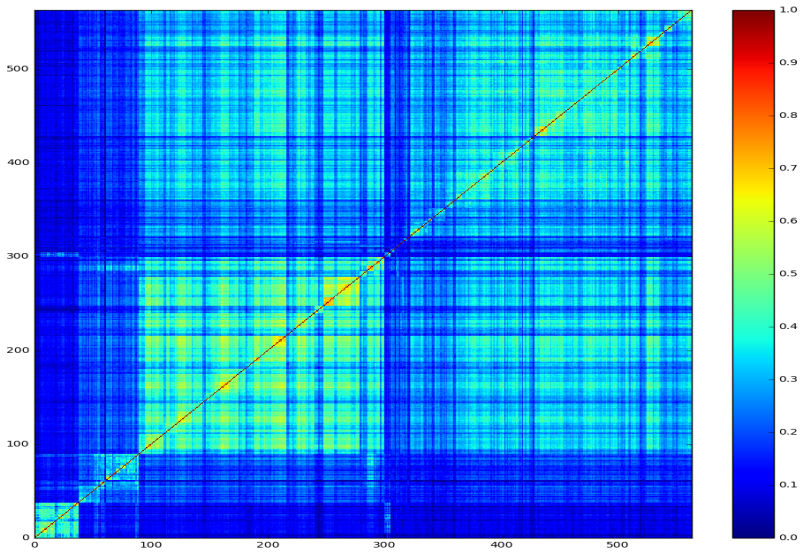



FIGURE 3.1 – Matrice de corrélation de Spearman de nos séries chronologiques

Pour y voir plus clair, regardons de plus près le premier cluster (en bas à gauche de la grande matrice). On observe la présence de blocs au niveau de la diagonale, ce qui signifie que l'on peut subdiviser davantage nos clusters en "sous-clusters" [5]. Ces subdivisions sont basées sur des critères comme le secteur d'activité ou encore le rating des entreprises (Investment grade ou High Yield). Selon les dynamiques propres à chaque pays/industrie, c'est un certain critère parmi tous ceux cités précédemment qui va avoir plus d'influence que les autres et qui va caractériser un groupe.

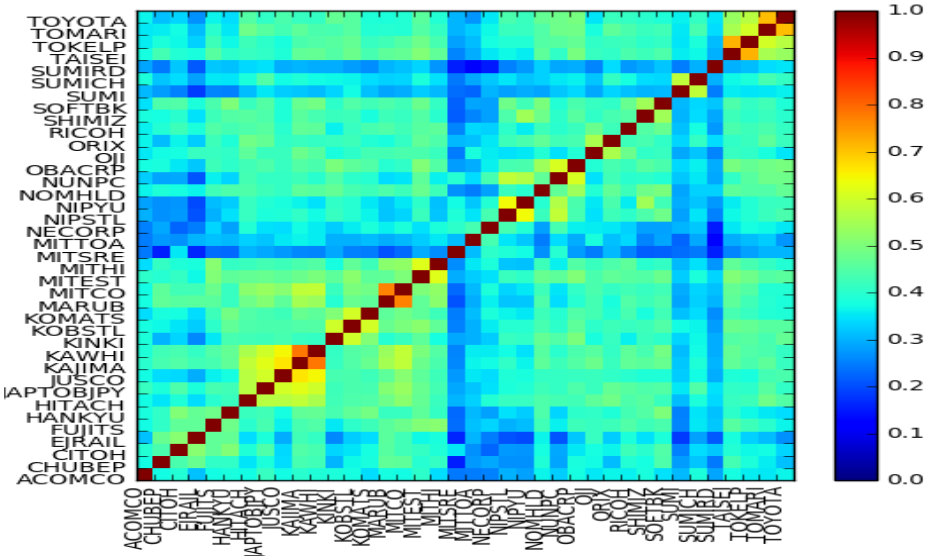



FIGURE 3.2 – Matrice de corrélation de Spearman du premier cluster

# Étude matricielle de la corrélation des CDS

## 4.1 Construction d'un indicateur basé sur les matrices de corrélation

Une fois nos clusters construits, nous avons entrepris la construction de plusieurs indicateurs de comportements anormaux pouvant aider à détecter les défauts. Le premier indicateur que nous avons construit est fondé sur l'étude matricielle de la corrélation des CDS au sein de chaque clusters. Au sein d'un cluster donné, nous avons cherché à quantifier l'évolution au cours du temps du niveau de corrélation des CDS appartenant à ce cluster. La série temporelle de l'indicateur, notée $(R_t)_{t\geq 0}$, se calcule en quatre étapes :

1. Fixer $W$ la largeur de la fenêtre glissante, et $M$ le nombre de matrices de corrélation que l'on souhaite dans notre ensemble de matrices, et $d$ la distance que l'on souhaite utiliser.
2. $\forall i \in [\![t-M, t]\!]$, calculer la matrice de corrélation des séries temporelles des CDS entre les dates $i-W$ et $i$, avec $W$ la largeur de la fenêtre glissante. On note $S_t$ l'ensemble des $M$ matrices de corrélation obtenues.
3. Calculer le rayon $R_t$ de cet ensemble de matrices $S_t$ à l'aide de la distance $d$ (le rayon est définit comme la plus grande distance entre l'une des matrices et le barycentre de l'ensemble).
4. Reprendre 2 et 3 au temps $t+1$ pour calculer $R_{t+1}$

La série temporelle $(R_t)_{t\geq 0}$ obtenue est notre indicateur final. Naïvement, une grande valeur de $R_t$ signifie qu'il y a une matrice de $S_t$ qui est très éloignée du centre de $S_t$, signe d'un changement plus ou moins brusque du niveau de corrélation des CDS au sein d'un cluster. Un grand rayon s'explique soit par une modification dans la composition du cluster (rajout d'un CDS) soit par une modification dans le comportement de ces CDS. Il s'explique le plus souvent par le fait qu'un CDS s'éloigne des autres et devient un "outlier" au sein de son cluster. Dans tous les cas, ce rayon peut être utilisé comme un indicateur pour prédire les défauts.

## 4.2 Recherche d'une distance matricielle adaptée

L'indicateur dépendant énormément de la distance matricielle utilisée dans le calcul, la prochaine étape de notre travail consiste à rechercher la distance la plus adaptée. Nous avons commencé naïvement par la distance euclidienne (norme de Frobenius, $Tr({}^tMM)$). Nous avons obtenu quelque chose de très bruité et donc difficilement exploitable (cf. Annexe 1). Nous nous sommes donc intéressées à des métriques plus spécifiques convenant à des modèles probabilistes. La première que nous avons retenue est la divergence de Kullback-Leibler [6], [7] qui est une mesure de dissimilarité entre deux distributions de probabilités. Pour des distributions $P$ et $Q$ continues de densités respectives $p$ et $q$, la divergence de Kullback-Leibler est définie par :

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

Ainsi, l'estimation de la divergence de Kullback-Leibler entre deux matrices de corrélation $P = (p_{i,j})_{i,j\leq n}$ et $Q = (q_{i,j})_{i,j\leq n}$ est :

$$\sum_{i\leq n, j\leq n} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j}}\right)$$

Cela définit la divergence de Kullback-Leibler non-paramétrique. En utilisant cette divergence, nous avons obtenu des résultats plus concluants, quoique toujours très bruités (cf. Annexe 2). Nous avons donc entrepris de définir notre modèle de façon paramétrique afin de pouvoir utiliser la divergence de Kullback-Leibler paramétrique. Nous avons fait l'hypothèse que l'accroissement $S_t^i$ de $X^i$ à la date $t$ suit une loi normale $\mathcal{N}(0, \sigma_{j,t})$ avec $j$ le numéro du cluster.

La divergence de Kullback-Leibler entre deux vecteurs gaussiens $S_t = (S_t^1, ..., S_t^{n_j})$ et $S_{t+1} = (S_{t+1}^1, ..., S_{t+1}^{n_j})$ de variance $\Sigma_t$ et $\Sigma_{t+1}$, avec $n_j$ le nombre de CDS au sein du cluster $j$, est [8] :

$$D_{\mathrm{KLP}}(X_t \| X_{t+1}) = \frac{1}{2}\left(\log\left(\frac{det(\Sigma_t)}{det(\Sigma_{t+1})}\right) + tr(\Sigma_t^{-1}\Sigma_{t+1}) - n\right)$$

L'estimé de la variance du vecteur gaussien $S_t$ se calcule en multipliant la matrice de corrélation à la date t par $\widehat{\sigma_{j,t}}^2$, que l'on estime avec les données de la fenêtre glissante. En utilisant cette divergence, nous avons obtenu un indicateur bien moins bruité et très prometteur. Cette divergence n'étant pas symétrique, nous avons étudié son comportement lorsque nous inversions les rôles de $X_t$ et $X_{t+1}$. Nous avons ensuite définit une distance basée sur cette divergence :

$$D(X_t \| X_{t+1}) = \frac{1}{2}(D_{\mathrm{KLP}}(X_t \| X_{t+1}) + D_{\mathrm{KLP}}(X_{t+1} \| X_t))$$

Vous trouverez en Annexe 3 l'évolution simultanée de ces trois distances.

## 4.3 Choix de la largeur des fenêtres glissantes

Nous avons donc retenu la distance de Kullback-Leibler paramétrique symétrique pour notre indicateur. Après avoir choisi la distance à utiliser, nous avons cherché les paramètres idéaux $M$ et $W$. Nous ne voulions pas d'une largeur de fenêtre $W$ trop grande, d'une part, car la conjoncture économique implique que la corrélation au sein d'un cluster évolue dans tous les cas, ce qui ne doit pas être interprété comme un signe de défaut. D'autre part, une trop grande valeur de $W$ ne permettrait pas de détecter rapidement un changement brutal dans le comportement d'un CDS. Inversement, un $W$ trop petit ne permettrait pas de capter un changement dans la tendance à moyen terme d'un CDS au sein de son cluster. Le même raisonnement s'applique pour $M$. Pour trouver les paramètres idéaux, nous avons effectué une recherche par quadrillage. Pour ce faire, nous avons fait varier $M$ entre 0 et 50, et W entre 0 et 100, et nous avons calculé le rapport signal sur bruit $(\frac{moyenne}{écart-type})$ de la série temporelle de l'indicateur correspondant. Les paramètres maximisant ce rapport sont $K = 40$ et $W = 70$. Vous trouverez ci-dessous une illustration de notre indicateur pour le premier cluster regroupant des entreprises Japonaises

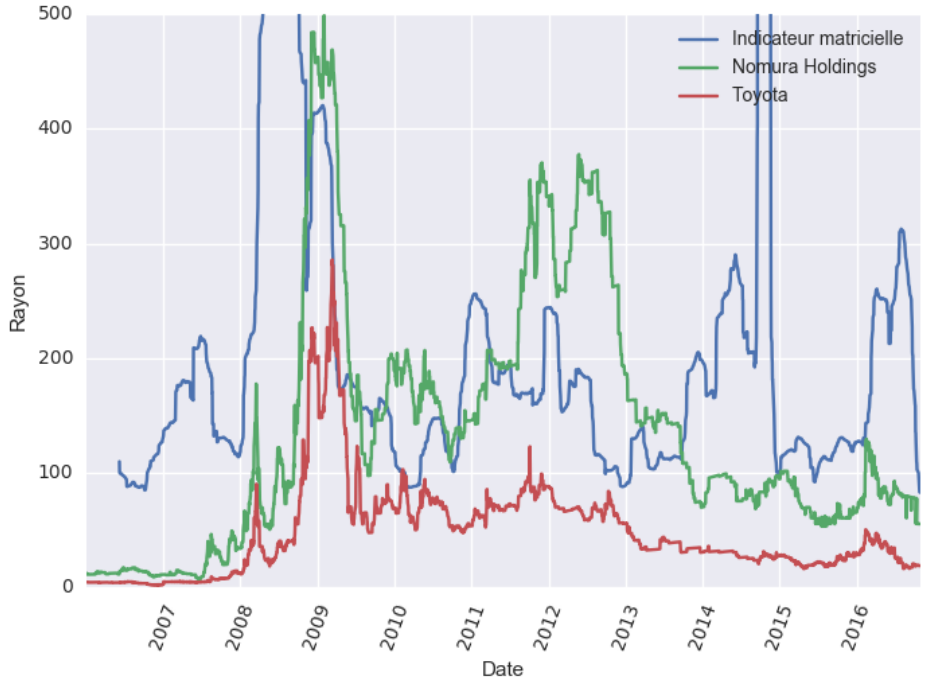



FIGURE 4.1 – Indicateur KL final du premier cluster représenté avec deux entreprises japonaises

Notre indicateur final a réussi à détecter l'effet qu'a eu la crise des subprimes sur le marché des couvertures de défaillance au Japon en 2008 ainsi que la récession subie par le Japon au troisième semestre de 2014[9].

# Utilisation des méthodes d'analyse des séries temporelles

## 5.1 Quantifier un mouvement anormal d'un CDS dans son cluster : première approche

Cet indicateur repose sur le postulat que les CDS ont un comportement similaire au sein d'un même cluster. En effet, la classification a permis de rendre compte de corrélations entre certains CDS, et ainsi de les classifier en clusters. Nous avons donc commencé par construire un indicateur "naïf" qui permet de rendre compte d'un mouvement anormal d'un CDS.

**Démarche :** Nous avons tenté de définir ce que pouvait être le comportement général des CDS au sein d'un même cluster. L'idée qui s'imposait naturellement était, à chaque temps $t$, de calculer la moyenne ou la médiane des valeurs de chaque CDS en ce point. Nous avons choisi la médiane afin que ce premier indicateur soit robuste et peu sensible aux valeurs aberrantes : en effet, lorsqu'un CDS a un comportement anormal (fort risque de faire défaut), celui-ci prend de très fortes valeurs, que nous voulons exclure de notre indicateur du comportement "normal" des CDS d'un cluster (voir en Annexe 4 un exemple sur le cluster numéro 2).

Ensuite, comme chaque CDS a son propre comportement habituel au sein de son cluster, nous avons quantifié celui-ci en calculant son éloignement médian à la série médiane du cluster :

$$eloignementMedian = med(X_t - X_t^{med})$$

Enfin, nous avons calculé les variations du CDS par rapport à ce comportement habituel : Si le CDS varie trop à la hausse par rapport à son éloignement habituel du centre du cluster, on peut alors dire qu'il effectue un mouvement anormal. Une des premières difficultés a été de quantifier ce "trop". Est ce qu'une hausse de $10\%$ d'éloignement par rapport à son éloignement habituel peut être considéré comme une forte hausse. En fait, l'indicateur est trop bruité pour que nous arrivions à définir un palier assez sensible sans détecter des milliers d'anomalies. Nous nous sommes donc intéressées à des modèles de séries temporelles afin de modéliser nos CDS.

## 5.2 Modèles de séries temporelles et distance au centre dans l'espace des paramètres estimés

Pour ce nouvel indicateur, l'idée est de modéliser les CDS par un modèle de séries temporelles, d'en extraire les paramètres après avoir entraîné le modèle, puis de calculer la distance dans l'espace de ces paramètres d'un CDS au centre de son cluster. On trouve cette approche dans la littérature notamment : Un modèle autorégressif, où même ARIMA[10], peut être ajusté aux séries. On compare ces modèles avec une métrique particulière (Piccolo 1990 ; Maharaj 2000 ; Xiong et Yeung 2002 ; Piccolo 2007, et Corduas et Piccolo 2008).

Cependant, nous nous heurtons à deux difficultés majeures : nos séries temporelles sont hétéroscédastiques ce qui pose problème pour l'application des modèles vus en cours (tels que ARMA ou ARIMA). De plus, il est très difficile de correctement spécifier les modèles (paramètres p,d,q de l'ARIMA par exemple) car chaque série est différente et il est très compliqué de trouver un moyen algorithmique pour spécifier ces paramètres automatiquement. Ceci étant dit, nous avons tout de même construit notre premier indicateur à l'aide d'une modélisation ARIMA (en essayant d'automatiser la spécification des paramètres), puis pour aller plus loin, nous avons considéré un modèle prenant en compte l'hétéroscédasticité conditionnelle des séries : le modèle ARCH.

**Comment spécifier les modèles ?** Pour spécifier les modèles, nous avons choisi de passer par la série médiane de chaque cluster, de l'étudier automatiquement afin d'approximer les paramètres, et de considérer ces mêmes paramètres pour toutes les séries du cluster. L'exemple développé par la suite correspond à celui du premier cluster.

### 5.2.1 Modèle ARIMA

**Rappels théoriques sur le modèle ARIMA :** Le modèle ARIMA permet de modéliser des séries temporelles non stationnaires mais qui pourront le devenir par différenciation, c'est-à-dire en considérant la série $\Delta^d X_t = (1-L)^d X_t$ où $L$ est l'opérateur de retard.

**Définition :** Un processus ARIMA(p,d,q) est un processus intégré d'ordre d, autorégressif sur p termes et considérant également comme variables explicatives q moyennes mobiles. Il vérifie l'équation suivante :

$$\forall t \Phi(L)(1-L)^d X_t = \Theta(L)\epsilon_t$$

où :

$$\Phi(L) = \mathbf{1} - \phi_1 L - \phi_2 L^2 ... - \phi_p L^p$$
$$\Theta(L) = \mathbf{1} + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q$$

et $(\epsilon_t)$ est un bruit blanc de variance $\sigma^2$.

**Détermination de l'ordre de différenciation et stationnarité :** Pour utiliser un modèle ARIMA, nous devons avoir une série stationnaire au moins au second ordre. C'est-à-dire de moyenne et de variance constante dans le temps. Or, nos séries présentent des tendances ainsi que de l'hétéoscédasticité. Pour la correction de la tendance de la série, cela peut se faire facilement par différenciation. Pour ce qui est de la variance, nous pouvons la corriger par des transformations logarithmique (si la variance croît) ou exponentielle (si la variance décroît). Or ici, la variance n'est certes pas constante, mais elle semble suivre un comportement plus compliqué qu'une simple croissance ou décroissance. C'est précisément ce point là qui nous a poussé plus tard à étudier les modèles ARCH. Cependant, pour le moment, aux risques de faire de grosses approximations, nous corrigerons la variance de manière naïve à l'aide des méthodes citées précédemment.

**La transformation Box-Cox :** Nous avons choisi d'utiliser la transformation Box-Cox pour corriger la variance. Elle permet de rendre les données normalement distribuées et s'écrit sous cette forme :

$$B(x) = \frac{x^\lambda - 1}{\lambda} \quad \text{si} \quad \lambda \neq 0$$
$$B(x) = log(x) \quad \text{sinon}$$

L'exemple suivant montre la détermination du paramètre $\lambda$ pour la série médiane du premier cluster à l'aide de la méthode du maximum de vraisemblance. Le paramètre $\lambda$ correspond au pic de la courbe des maximum de vraisemblance.

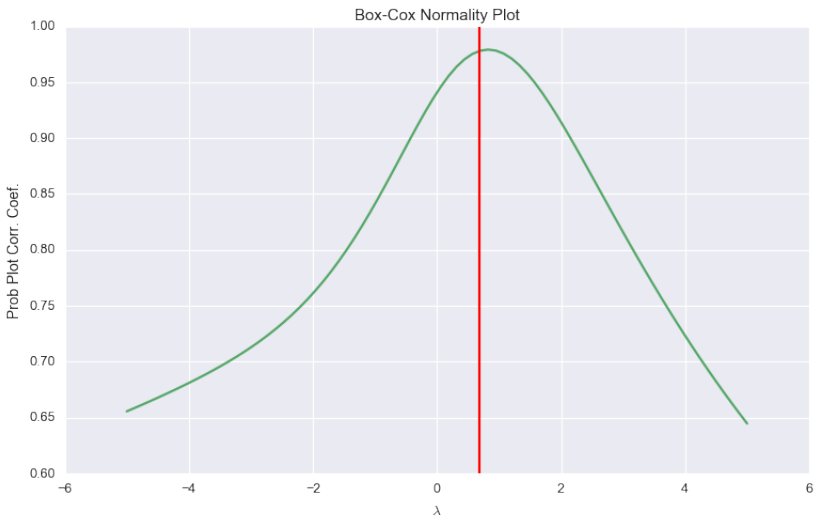



FIGURE 5.1 – Détermination du paramètre $\lambda$ de la procédure Box-Cox de la série médiane du cluster 1

**Ordre de différenciation :** Dans la littérature, l'ordre de différenciation généralement retenu pour les séries financières est de 1. Pour déterminer correctement l'ordre de différenciation, on s'intéresse à la fonction d'autocorrélation. Si celle-ci décroît lentement (courbe bleue sur le graphique suivant), alors la série doit être différenciée. Si après différenciation on obtient une courbe qui tend vite vers 0 avec des autocorrélations faibles (courbe verte sur le graphique suivant), alors la série est devenue stationnaire. Lorsqu'on différencie une deuxième fois la série médiane, l'écart-type des résidus augmente, confirmant une sur-différenciation. On retient donc un ordre de différenciation égal à 1, le paramètre $d$ de la procédure ARIMA de la série médiane sera donc égal à 1.


FIGURE 5.2 – Autocorrélations de la série médiane non différenciée et différenciée du cluster 1

**Détermination des termes autorégressifs AR :** En regardant la fonction d'autocorrélations partielles, il est possible de déterminer le nombre de termes autorégressifs à inclure dans le modèle. Le décalage auquel l'autoccorélation partielle disparaît (n'est plus significative) nous donne la valeur du paramètre $p$ donnant le nombre de termes autorégressifs à inclure dans le modèle. Dans le cadre de notre exemple, nous allons inclure 4 termes autorégressifs.

**Détermination des termes moyennes mobiles MA :** En regardant la fonction d'autocorrélations (et non d'autocorrélations partielles), il est possible de la même manière que précédemment de déterminer le nombre de termes MA à inclure et donc la valeur du paramètre $q$. Dans le cadre de notre exemple, nous avons inclus 3 termes moyennes mobiles pour la série médiane du premier cluster.

**Fit du modèle $ARIMA(4,1,3)$ et résultats quant à sa pertinence :** Après avoir entraîné le modèle, les tests de student de chaque coefficient permettent de donner une idée de leur significativité. Nous avons observé que le quatrième coefficient AR de la série médiane du premier cluster est peu significatif. Cependant, lorsque l'on veut utiliser le modèle $ARIMA(3,1,3)$ pour palier ce problème, l'algorithme python bloque car il détecte une série non stationnaire. Ceci est dû à la variance de la série qui tantôt croît, tantôt décroît, et qui semble dépendre de la variance passée de la série. Ce problème nous amène à considérer un autre type de modèle : les modèles $ARCH$ qui permettent de considérer une série avec une variance conditionnellement au passé. Cependant, nous allons dans un premier temps utiliser le modèle $ARIMA$ classique pour calculer la distance des CDS au centre de leur cluster.

### 5.2.2 Distance au centre après modélisation par le modèle $ARIMA$

Au niveau algorithmique, nous choisissons les mêmes paramètres p, d et q pour modéliser la série médiane du cluster (qu'on définit comme le centre du cluster) et la série étudiée appartenant au cluster. De cette manière, nous sommes certaines d'avoir la même dimension pour les vecteurs des paramètres estimés. De plus, comme elles appartiennent au même cluster, on peut supposer qu'elles se ressemblent assez pour pouvoir utiliser la même spécification du modèle $ARIMA$.

**Distance utilisée :** Pour mesurer la distance d'un CDS au centre de son cluster, nous avons donc choisi d'utiliser une f-divergence (la total variation distance) sur les vecteurs des paramètres estimés. Elle permet de mesurer la distance entre deux distributions de probabilité par la formule qui suit : pour $P = (p_1, ...p_k)$ et $Q = (q_1, ..., q_k)$ :

$$\delta(P, Q) = \frac{1}{2} \sum_{j=1}^{k} |p_j - q_j|$$

Plus la valeur donnée par l'indicateur pour un CDS est grande, plus ce CDS est considéré comme risqué car de plus en plus éloigné du comportement moyen de son cluster.

Cependant, compte tenu des erreurs qui surviennent à cause de la non stationnarité de certaines séries, nous n'obtenons un résultat que sur la moitié des CDS étudiés. Cela nous empêche de faire une analyse approfondie des clusters. En effet, il suffit que l'algorithme n'arrive pas à calculer le modèle pour un outlier au sein d'un cluster et l'analyse en devient entièrement faussée. Nous nous intéressons donc à un autre modèle de séries temporelles.

### 5.2.3 Prise en compte de l'hétéroscédasticité : modèle $ARCH$

La modélisation précédente ne permet pas de rendre compte de la volatilité des CDS, mesure pourtant centrale en finance. Pour essayer d'inclure une meilleure modélisation de la volatilité des CDS, nous nous sommes intéressées au modèle $ARCH$ [11] (autoregressive conditional heteroscedasticity) qui permet de faire dépendre la variance de la série de son passé, et donc d'alterner des périodes de forte volatilité avec des périodes de volatilité plus faible. Grâce au modèle ARCH, nous pouvons observer que la distribution des rendements est une gaussienne à queue lourde, c'est-à-dire que ce modèle arrive à capturer des événements rares, ce qui dans le cas de la détection des défauts peut nous être d'une aide précieuse.

**Présentation du modèle $ARCH$ :** Le modèle $ARCH(q)$ est un modèle non-linéaire dans lequel la variance conditionnelle tient compte du passé de la série. Le modèle s'écrit comme suit :

$$X_t = z_t \sqrt{h_t}$$

où $z_t$ est un bruit blanc gaussien et où $h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2$ La variance conditionnelle s'écrit donc :

$$V[X_t | X_{t-i}] = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2$$

Les moyennes conditionnelle et non conditionnelle sont nulles.

**Choix du paramètre :** Nous allons donc appliquer le modèle $ARCH$ à nos séries différenciées (car il est nécessaire que le processus soit de moyenne nulle pour lui appliquer le modèle) et nous allons arbitrairement choisir $q = 1$ (pour des raisons de facilité, de rapidité des algorithmes et du fait que c'est souvent une modélisation prise dans les ouvrages pour modéliser facilement les séries financières).

### 5.2.4 Analyse des résultats de l'indicateur issu du modèle $ARCH$

Nous commençons par nous intéresser à l'indicateur calculé sur toute la période donnée (et non sur une fenêtre glissante). Regardons les diagrammes de Tukey de l'indicateur pour chaque cluster.

On observe que certains cluster ont subis de forts chocs sur la période (cf figure 5.3), notamment les clusters :

1. Cluster numéro 9 : Il contient des entreprises du secteur automobile, des compagnies aériennes ainsi que du secteur du bâtiment, notamment Volkswagen et Lufthensa qui ont été toutes deux en grandes difficultés financières en 2015 (affaire des moteurs polluants de Volkswagen en 2015 et crash de l'A320 sur la compagnie Lufthensa en 2015).
2. Cluster numéro 15 : Il contient des compagnies du secteur du digital et du médical. Nous n'avons pas trouvé d'explication évidente à cette détection.

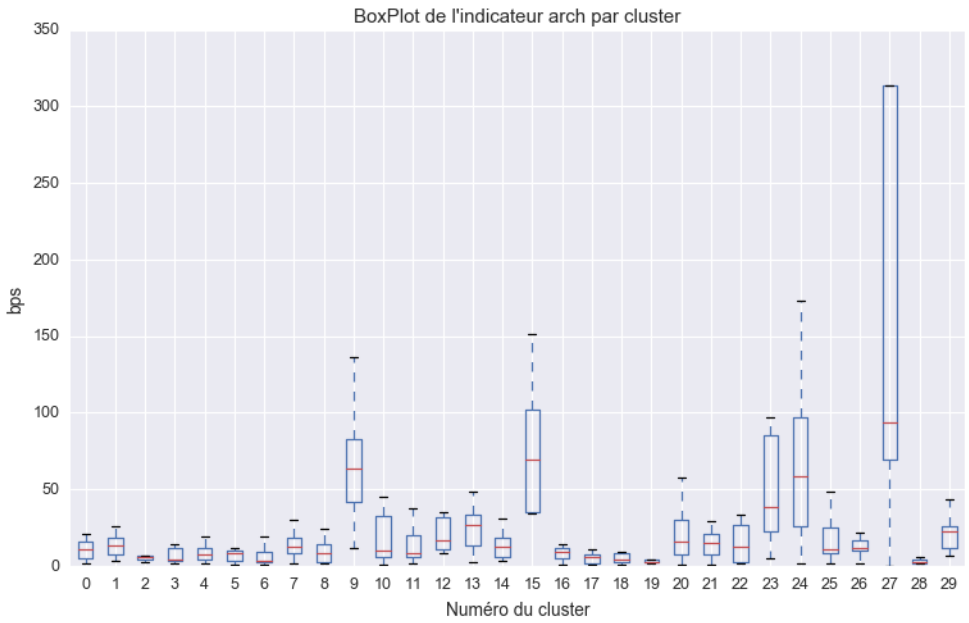



FIGURE 5.3 – Diagramme de Tukey de l'indicateur ARCH par cluster

3. Cluster numéro 23 : Il contient des compagnies qui sont dans le design, la construction et la vente d'immobilier aux Etats-Unis telles que Hovnanian Entreprises ou M.D.C. Holdings. Il n'est pas surprenant que nous détections ce cluster compte tenu de la crise immobilière qui a touché le marché américain en 2007.
4. Cluster numéro 24 : Il contient énormément de CDS de compagnies assez différentes (acier, électricité, immobilier, financier). On peut supposer que ce cluster "divers" a été touché différemment selon le CDS pendant la crise des subprimes, ce qui explique que nous le détection.
5. Cluster numéro 27 : C'est le cluster qui a clairement subi les plus gros chocs. Il contient des compagnies d'assurances américaines spécialisées dans l'assurance de prêt immobilier (Mortgage insurance) qui permet à des ménages à faible capacité financière de pouvoir obtenir des prêts qu'ils n'auraient pas obtenus sinon. Ce sont donc des entreprises qui ont été en très grandes difficultés pendant la crise des subprimes.

# Étude de la structure des échéances des CDS

## 6.1 La structure des échéances : une courbe habituellement croissante et concave

On appelle structure des échéances d'un CDS, la courbe qui représente son prix pour différentes maturités.

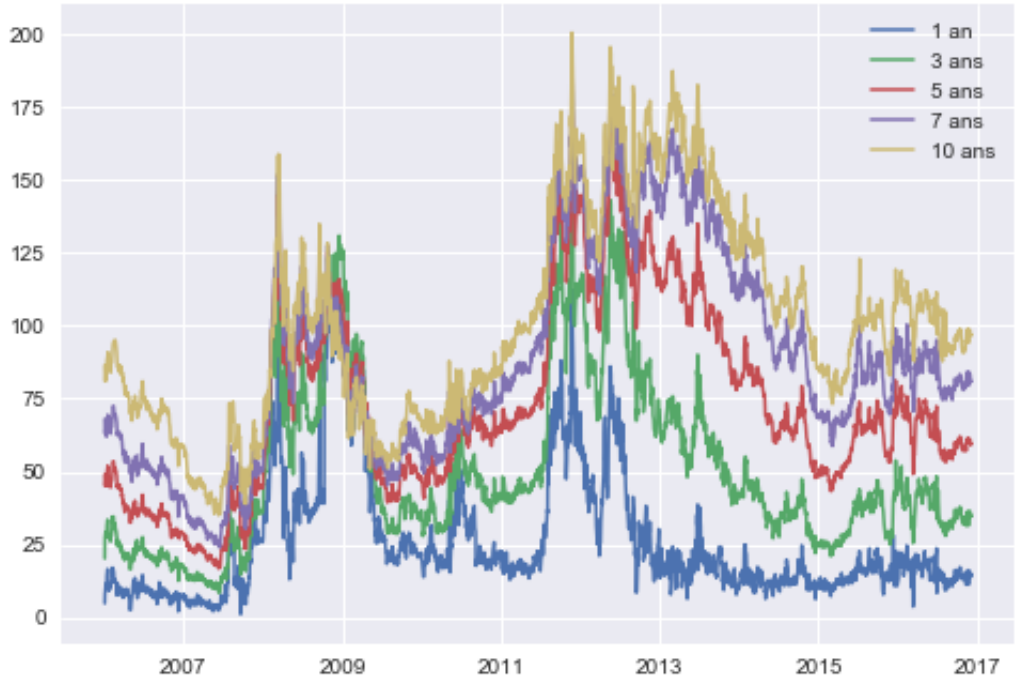


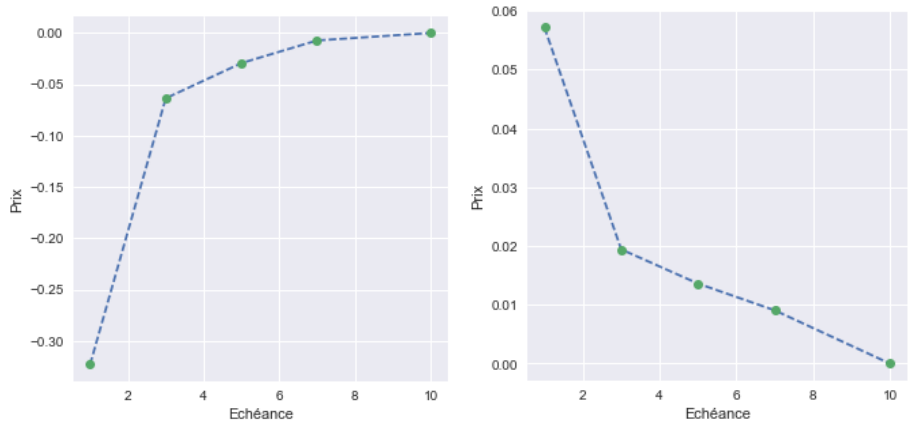

FIGURE 6.1 – Séries temporelles de la structure des échéances d'Orange

En intégrant les structures des échéances dans la construction de notre indicateur, nous pouvons capter de nouvelles informations utiles à la détection d'un potentiel défaut d'un CDS étudié.
Lorsqu'une entreprise est vue de manière positive par les investisseurs, ceux-ci estiment que la probabilité de défaut du CDS à court terme est faible, hors période agitée sur les marchés. Par exemple, on constate que la courbe du prix à 10 ans est en dessous de la courbe du prix à 3 ans. Cependant, même si le produit est jugé sûr, il est beaucoup plus risqué de parier à long terme sur le non défaut du CDS. Le prix d'un CDS de maturité 1 an est donc toujours plus faible (car moins risqué donc on cherche moins à se protéger contre un événement de crédit) qu'un CDS de maturité 10 ans sur le même sous-jacent. Une courbe de structure des échéances normale est donc croissante en la maturité [12] [13]. En outre, dans 70% de cas, elle est concave. En période agitée sur les marchés ou bien lorsqu'une entreprise est en difficulté, la courbe a tendance à s'aplatir, voir à changer de concavité (donc à devenir convexe) si les investisseurs prévoient un défaut rapide compte tenu de la situation de l'entreprise. Le graphique 6.2 présente deux courbes de structure des échéances d'un même CDS en différentes situations économiques.

## 6.2 Construction d'un indicateur basé sur la structure des échéances

L'idée est ici de construire un indicateur qui quantifie le niveau de croissance et de concavité de la structure des échéances d'un CDS. Supposons que $p_{t,j}(x_i)$ est le prix en $t$ d'un CDS $j$, avec $x_i$ l'échéance ($x_1$=1, $x_2$=3, $x_3$=5, $x_4$=7 et $x_5$= 10 ans). En notant $\Delta p_{t,j}(x_i) = p_{t,j}(x_i) - p_{t,j}(x_{i-1})$ la différenciation à l'ordre 1 et $\Delta^2 p_{t,j}(x_i) = \Delta p_{t,j}(x_i) - \Delta p_{t,j}(x_{i-1})$ la différenciation à l'ordre 2, nous définissons la série temporelle suivante :

$$C_{t,j} = \sum_{i=2}^{5} \Delta p_{t,j}(x_i) - \sum_{i=3}^{5} \Delta^2 p_{t,j}(x_i)$$




*Note :* En mai et juin 2006, France Télécom et Wanadoo deviennent Orange. Cette fusion explique l'instabilité subie par le CDS d'Orange sur cette période et l'inversion de la courbe de structure des échéances.

FIGURE 6.2 – Structures des échéances normale (à gauche, d'Orange le 18 janvier, 2006) et anormale (à droite, d'Orange le 30 mai, 2006)

En période non agitée, on s'attend à ce que les valeurs de cette série soient relativement grandes du fait de la concavité et de la corrélation positive entre le prix et l'échéance de la structure des échéances. En effet, on aurait $\forall i \in [\![1, 5]\!]$ et $\Delta p(x_i) \geq 0, \Delta^2 p(x_i) \leq 0$. Inversement, en période agitée, on s'attend à des valeurs plus faibles synonymes d'une diminution du niveau de la corrélation positive et de la concavité de la structure des échéances. Pour supprimer les fluctuations transitoires de façon à souligner la structure à plus long terme, nous avons retenu comme indicateur final la moyenne mobile de $C_{t,j}$ sur 20 jours (un mois) :

$$\overline{C_{t,j}} = \frac{1}{20} \sum_{i=0}^{19} C_{t-i,j}$$

.

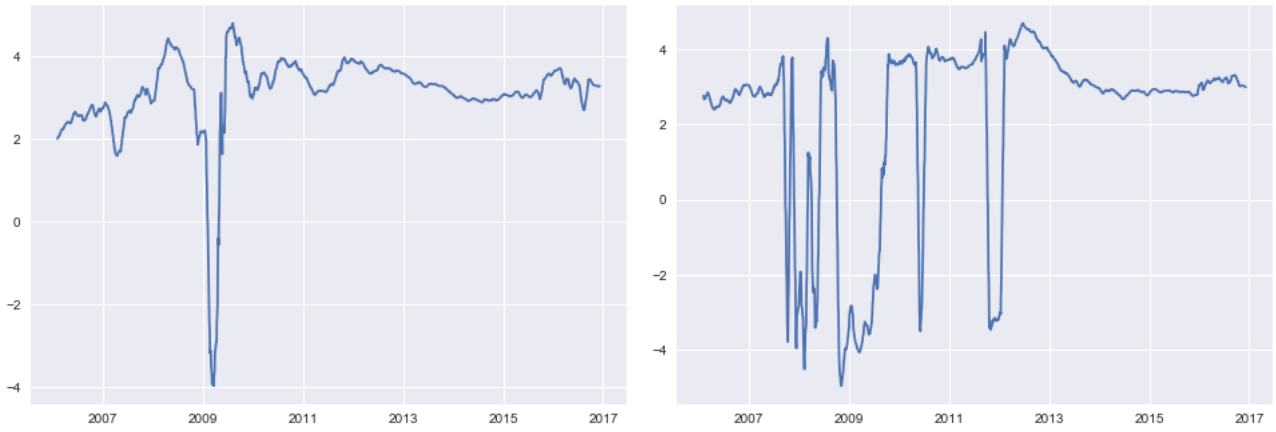



FIGURE 6.3 – Indicateur de normalité d'Arconic Inc. et de The Goldman Sachs Group, Inc.

Sur le graphique 6.3, on constate qu'en 2009, l'indicateur d'Arconic Inc. a fortement baissé en raison de l'effet qu'a eu la crise des subprimes sur la structure des échéances. Mais l'entreprise s'est rapidement rétablie. On constate un deuxième pic moins visible vers fin 2016. En effet, à cette période, l'entreprise Alcoa s'est scindée en deux et une partie de son activité a rejoint Arconic [14]. Le pic traduit donc une petite baisse de confiance des marchés au moment de la fusion, qui se sont cependant très vite rassurés. Pour Goldman Sachs Group, l'indicateur a baissé plus fortement et est resté plus longtemps en bas en raison de son implication directe dans la crise de 2008 et de l'insécurité des banques provoquée notamment par les faillites de Lehman Brothers et AIG.

Le graphique ci-dessous montre les séries temporelles de l'indicateur de 7 clusters. On voit que l'indicateur diminue brusquement pour ces clusters en 2009. Le cluster 1 contenant exclusivement des entreprises japonaises a un indicateur qui ne dépasse jamais 0, ce qui reflète l'économie ralentie du Japon. On note que l'indicateur du cluster 3 (rassemblant des institutions financières indiennes) présente de fortes variations en 2016. Cela est dû à la réforme financière indienne de 2016 qui a entamée la confiance des investisseurs dans le roupie [15].


FIGURE 6.4 – Indicateurs de normalité de 7 clusters

Pour le cluster qui rassemble les pays d'Europe de l'Est (Graphique 6.5), l'indicateur distingue bien la Pologne qui est le seul pays dont l'indicateur ne diminue pas en 2008 (La courbe en violet). En effet, la Pologne a été applaudie pour sa performance pendant la crise financière [16].


FIGURE 6.5 – Indicateurs des pays d'Europe de l'Est

# La Théorie des Graphes appliquée à notre étude

Dans cette partie, nous avons utilisé la théorie des graphes pour modéliser nos clusters, et utilisé certaines de leurs propriétés pour en extraire des indicateurs. Tout d'abord, quelques définitions :

**Graphe non orienté :** Un graphe non orienté est un ensemble de sommets $V$ reliés par un ensemble d'arêtes $E$. Il est dit connexe si quelques soient $u$ et $v$ dans $V$, il existe une succession d'arêtes permettant d'accéder à $v$ à partir de $u$. Il est dit pondéré si chaque arête est affectée d'un nombre réel positif.

**Arbre :** Un arbre est un graphe non orienté, connexe et acyclique : quelques soient $u$ et $v$, il existe un unique chemin permettant d'accéder à $v$ à partir de $u$

**Arbre couvrant :** Étant donné un graphe non orienté connexe $G(V, E)$, un arbre couvrant $T(V, E')$ de $G$ est un arbre inclus dans ce graphe qui connecte tous les sommets du graphe.

**Arbre couvrant minimal (ACM) :** Soit $G(V, E)$ un graphe non orienté connexe dont les arêtes sont pondérées d'un poids $W(u, v)$. Un arbre couvrant minimal $T_{min}(V, E')$ est un arbre couvrant de $G$ dont la somme des poids des arêtes est minimale. Il est unique si les poids sont différents (cf. Annexe 5)

**Chaîne :** Soit $T$ un arbre couvrant minimal d'un graphe pondéré $G(V, E)$. La chaîne reliant $u$ à $v$, est définit comme l'unique suite finie d'arêtes consécutives reliant $u$ à $v$. Sa taille est définie comme la somme des poids des arêtes la constituant.

## 7.1 Construction des graphes et de leur arbre couvrant minimal

Nous avons modélisé chaque cluster par un graphe non orienté connexe pondéré. Chaque CDS constitue un sommet du graphe et chaque arête est pondérée de la distance entre les deux CDS la constituant. La distance considérée ici est $\frac{1-\rho_{i,j,t}}{2}$ avec $\rho_{i,j,t}$ la corrélation entre le CDS $i$ et $j$ calculée entre $t-60$ et $t$. Pour chaque cluster et pour chaque date $t$, nous avons construit l'arbre couvrant minimal de son graphe à l'aide de l'algorithme de Prim[17]. L'algorithme de Prim est un algorithme qui calcule un arbre couvrant minimal $T$ d'un graphe connexe pondéré non orienté $G$. Les étapes de l'algorithme sont les suivantes :

1. Initialiser $T$ composé de : $\begin{cases} \text{Un sommet } s_i \text{ choisi de façon aléatoire} \\ \text{Aucune arêtes} \end{cases}$
2. Considérer toutes les arêtes de G qui relient un sommet appartenant à T et un sommet qui n'appartient pas à T. Parmi celles-ci, choisir l'arête dont le poids est le plus petit. Ajouter à T cette arête et le sommet correspondant.
3. Répéter 2 jusqu'à ce que tous les sommets de $G$ soient dans $T$.

Vous trouverez en Annexe 7 une représentation de l'arbre couvrant minimal du premier cluster constitué des 38 CDS numérotés de 1 à 38, pour $t = 11/08/2016$

## 7.2 Construction de deux indicateurs ACM

Après avoir obtenu l'arbre couvrant minimal de chaque cluster pour chaque date $t$, nous avons construit deux indicateurs : un indicateur ACM global, qui est le même pour chaque CDS appartenant au sein d'un même cluster, et un indicateur ACM local, qui diffère pour chaque CDS.

**Indicateur ACM global :** Pour une date $t$ donnée, nous avons défini l'indicateur ACM global d'un cluster comme le diamètre de l'arbre couvrant minimal de son graphe. Le diamètre d'un arbre couvrant $T$ est définit comme la taille de la plus grande chaîne reliant deux sommets de $T$. Les arêtes étant pondérées par un poids inversement proportionnel au niveau de corrélation des CDS les constituant, une valeur importante de l'indicateur signifie un niveau de corrélation bas au sein du cluster. Ceci pourrait être interprété comme un signe avant-coureur de défaut.

**Indicateur ACM local** Soit $t$ une date donnée, $C$ un cluster et $T$ l'arbre couvrant minimal de son graphe. Nous avons définit l'indicateur ACM local d'un CDS appartenant à $C$ comme la taille de la plus grande chaîne possible d'un CDS à un autre sommet de T. Vous trouverez en Annexe 7 et 8 un exemple de l'évolution de ces deux indicateurs.

## 7.3 Étude de la corrélation des indicateurs

Après avoir construit tous nos indicateurs, il est intéressant d'étudier à quel point ils sont corrélés entre eux. L'idéal aurait était d'obtenir des indicateurs qui arrivent tous plus ou moins à détecter un choc ou une anomalie sur le marché. On aurait donc obtenu des indicateurs corrélés, mais pas trop de sorte à ce que chaque indicateur apporte une information supplémentaire. Pour illustrer le niveau de corrélation de nos indicateurs, nous avons choisi de représenter les indicateurs du CDS de Volkswagen de maturité 5 ans.

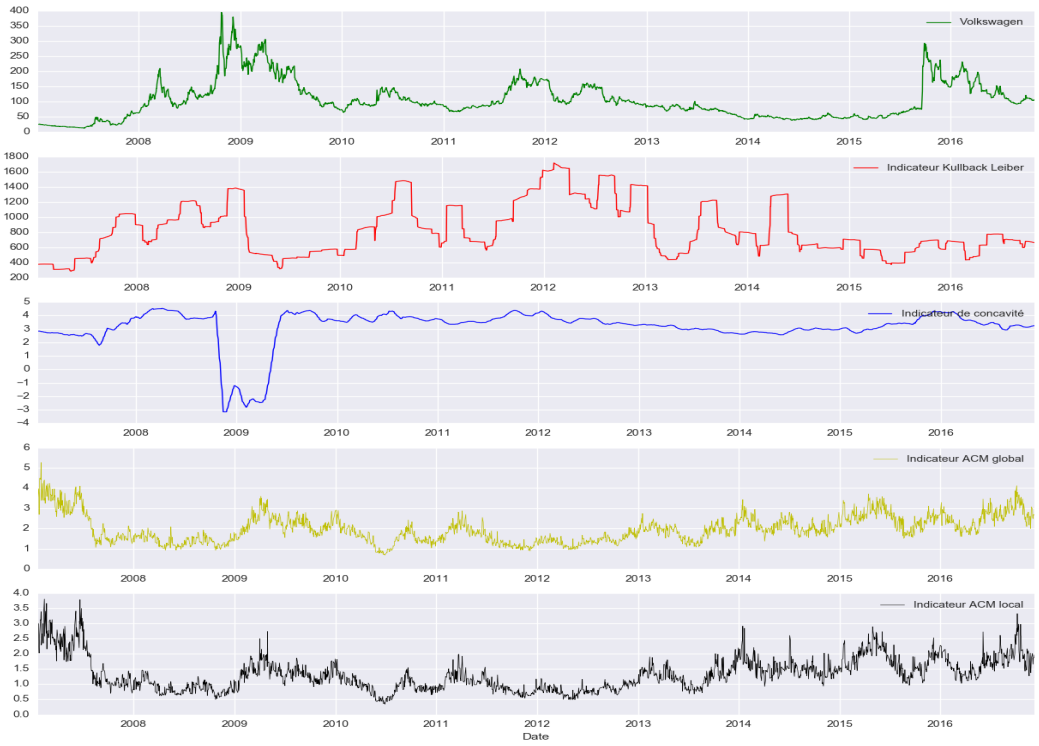



FIGURE 7.1 – Évolution des différents indicateurs du CDS de Volkswagen de maturité 5 ans

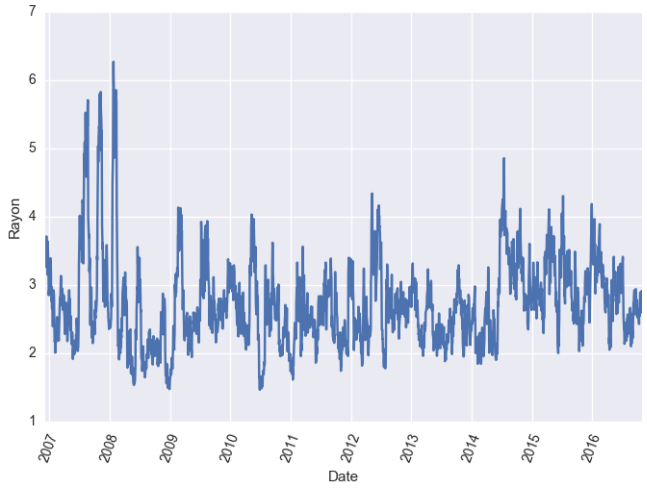
Plusieurs observations peuvent être faites. Premièrement, on remarque que certains indicateurs sont sensibles seulement à des variations très importantes du prix de Volkswagen, notamment l'indicateur de concavité, tandis que d'autres sont sensibles à des variations moins importantes (Kullback-Leibler), qui pourraient être liées à un changement de structure du cluster contenant Volkswagen. Hormis l'indicateur ACM global et local qui évoluent de façon quasi-identique, les indicateurs n'évoluent pas de la même manière, ce qui est une bonne chose compte tenu du fait que l'on cherche à maximiser l'apport d'information.

# Conclusion

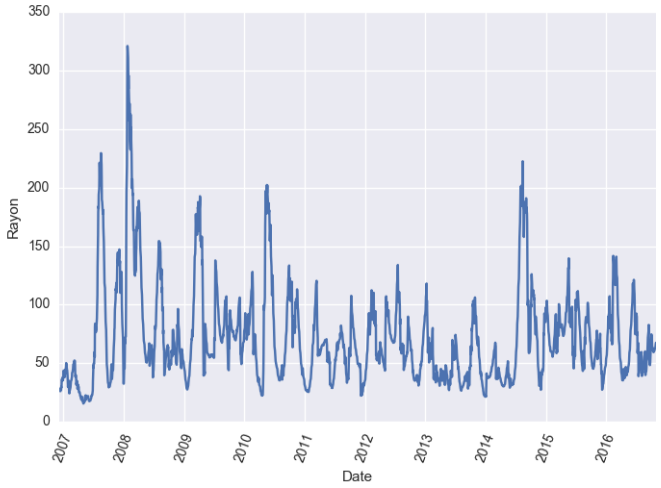
**Ce que nous aurions pû faire dans la continuité de notre travail :** Le modèle de régression logistique bayésienne est un modèle de régression généralisé permettant d'expliquer une variable binaire, en supposant une densité $apriori$. Dans notre exemple, nous souhaitons faire une prédiction à court terme des défauts sur le marché des CDS. Autrement dit, nous pouvons faire une régression logistique bayésienne de nos indicateurs sur notre variable à expliquer : "A fait défaut à la date t", "N'a pas fait défaut à la date t". Une fois le modèle estimé, nous pourrons effectuer une prédiction sur un échantillon test, puis évaluer cette prédiction. Cependant, beaucoup de problèmes entre en compte : beaucoup de nos indicateurs sont corrélés entre eux et nous avons sans doute omis certaines variables (variables exogènes d'un choc économique ou autre...), et nous devons définir sur quelle période de temps nous allons traiter nos indicateurs. En effet, le but de ce projet de statistiques appliquées est de prévoir les défauts à court terme. Faire une régression sur toute la période (10 ans) n'aurait pas beaucoup de sens, surtout compte tenu des forts changements de contexte économique durant la période considérée. De plus, effectuer une régression chaque jour paraît impossible de manière pratique : il n'y a pas eu assez de défauts dans l'historique pour qu'il y ai suffisamment de données sur chaque jour. Une solution intermédiaire pourrait être d'évaluer sur une période de 1 à 3 mois.

**Conclusion générale :** Pendant nos 6 mois de travail sur ce projet, nous avons dû faire face à beaucoup de difficultés contingentes à l'utilisation de modèles théoriques sur des données réelles. De plus, tirer des conclusions de nos indicateurs a parfois été difficile et nous avons dû essayer nombre de modèles avant de pouvoir nous lancer dans des analyses cohérentes. Malgré tout cela, nous avons pu construire des méthodes pertinentes pour analyser le marché des couvertures de défaillance bien qu'il nous a été impossible de prévoir un défaut.

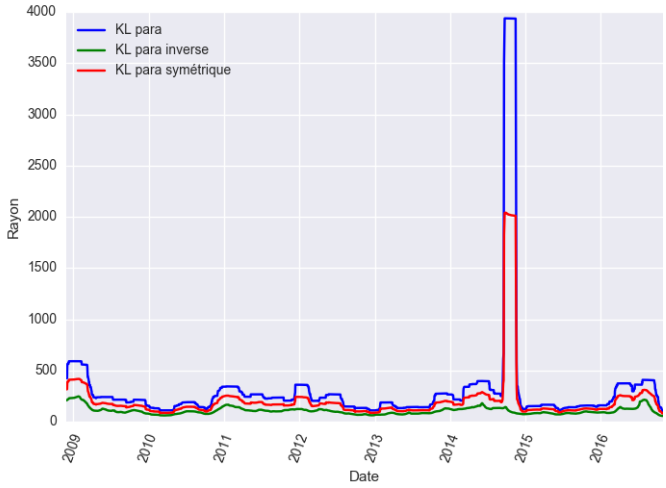
# Annexes



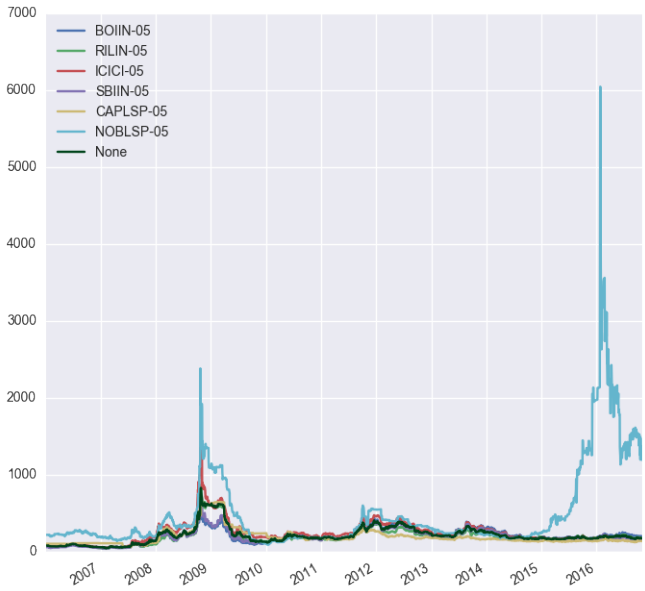

Annexe 1 – Indicateur matricielle du premier cluster, distance euclidienne



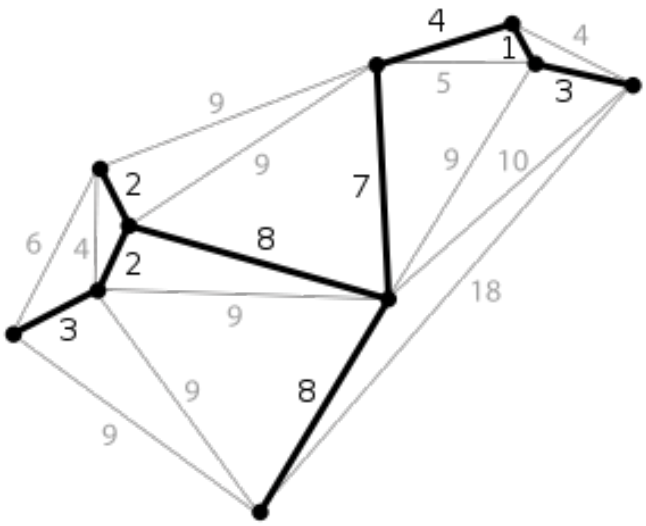

Annexe 2 – Indicateur matricielle du premier cluster, distance de Kullback-Leibler non-paramétrique



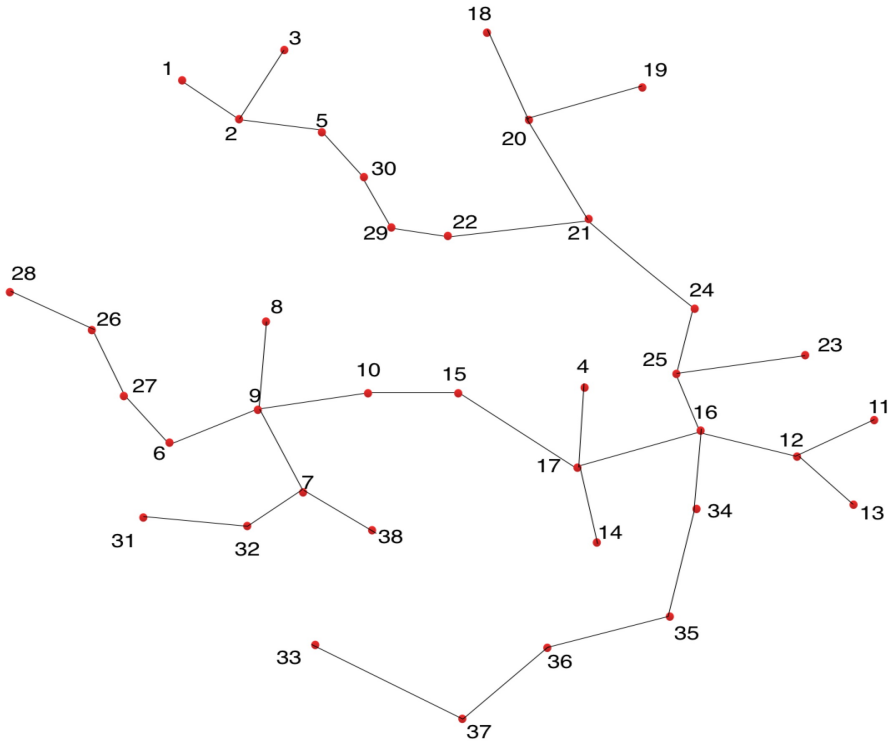

Annexe 3 – Indicateur matricielle du premier cluster, Kullback-Leibler



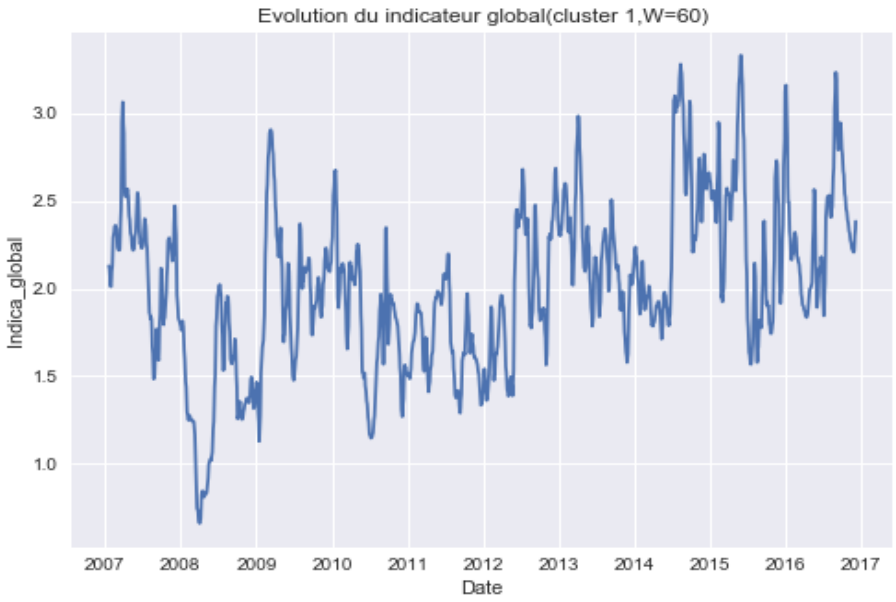

Annexe 4 – Cluster numéro 2 avec sa médiane (en vert foncé)



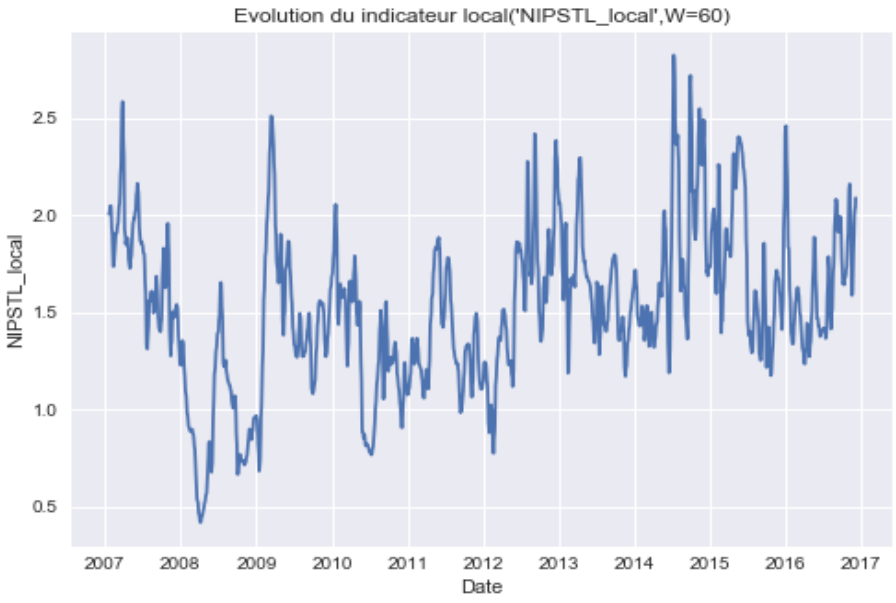

Annexe 5 – Exemple d'arbre couvrant minimal d'un graphe




Annexe 6 – ACM du premier cluster, 11/08/2016




Annexe 7 – Évolution de l'indicateur ACM global du premier cluster




Annexe 8 – Évolution de l'indicateur ACM local du CDS de Nippon Steel Sumitomo Metal Corporation à maturité 5 ans

# Bibliographie


[1] Phillipe Herlin. Cds : Pire que la dette, le produit financier qui pourrait provoquer la faillite des banques françaises. *Atlantico*, (21), sept 2011.

[2] "Wikipedia". "credit default swap, wikipedia, the free encyclopedia", "2004".

[3] Virginie Coudert and Mathieu Gex. The credit default swap market and the settlement of large defaults. Working Papers 2010-17, CEPII, 2010.

[4] Gautier Marti, Frank Nielsen, Philippe Donnat, and Sébastien Andler. *On Clustering Financial Time Series : A Need for Distances Between Dependent Random Variables*, pages 149–174. Springer International Publishing, Cham, 2017.

[5] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Clustering financial time series : How long is enough ? In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2583–2589, 2016.

[6] S. Kullback. The Kullback-Leibler distance. *The American Statistician*, 41(4) :340–341, 1987.

[7] Frank Nielsen and Ke Sun. Guaranteed bounds on the kullback-leibler divergence of univariate mixtures. *IEEE Signal Process. Lett.*, 23(11) :1543–1546, 2016.

[8] Y. Lu, A.M. Stuart, and H. Weber. Gaussian approximations for probability measures on rd. submitted.

[9] Adam Reichardt. Le japon replonge en récession, crise politique en vue. *Challenges*, 11 2014.

[10] Didier Delignières. Séries temporelles – modèles arima. *Séminaire EA "Sport – Performance – Santé"*, mars 2000.

[11] Elizabeth Howard. La volatilité selon les modèles garch. *La revue d'Opus Finance*, dec 2012.

[12] Matthias Laub Stefan Truck and Svetlozar T. Rachev. The term structure of credit spreads and credit default swaps - an empirical investigation. sept 2014.

[13] The slope of the term structure of credit spreads : An empirical investigation. *CFA Digest*, 37(4) :22–24, 2007.

[14] Gil Bousquet. L'usine alcoa devient arconic. *La Dépêche*, 11 2016.

[15] Institute of Economic Growth Pravakar Sahoo. Can indian financial reform build better banks ? *East Asia Forum*, 06 2016.

[16] Adam Reichardt. Poland and the global economic crisis : Observations and reflections in the public sector. *Journal of Finance and Management in Public Services*, 10(1), 1987.

[17] A. Kershenbaum and R. Van Slyke. Computing minimum spanning trees efficiently. In *Proceedings of the ACM Annual Conference - Volume 1*, ACM '72, pages 518–527, New York, NY, USA, 1972. ACM.