UNIVERSIDAD AUTONOMA DE MADRID

Rapport de stage

effectué à Madrid, au sein de la société

BME BOLSAS Y MERCADOS ESPAÑOLES

Sébastien Ta

Étudiant en 3ème année à Sup Galilée/Université Paris 13 en spécialité mathématiques appliquées et calcul scientifique

Analyse statistique sur les ratios d'inefficience de liquidation-règlement

Mai-Septembre 2017

Ingénieurs sup Galilée

UNIVERSITÉ PARIS 13

Table des matières

1 Remerciements

Je tiens à remercier l'ensemble des personnes qui m'ont permis de réaliser ce stage, m'ont permis d'écrire ce rapport et d'avoir cette expérience unique. Tout d'abord je remercie M.Santiago Carrillo Menéndez de l'Université Autonome de Madrid qui a été mon premier contact et m'a beaucoup aidé dans la recherche du stage en Espagne. Ayant fais part de mes intentions de trouver un stage dans le domaine de la finance, il m'a permis de trouver ce stage dans l'entreprise BME qui convenait tout à fait.

Je tiens également à profondément remercier mon maître de stage M.Jose Manuel Ortiz, chef du département PTI au sein de l'entreprise BME. Son accueil, sa disponibilité, les méthodes de travail et les échanges professionnels que nous entretenions ont été grandement enrichissant et me font désormais sentir beaucoup plus à l'aise dans les domaines et contextes auxquels j'ai été confronté.

Je souhaite aussi remercier tout les membres de l'équipe du PTI de l'entreprise BME, qui m'ont accueilli et qui ont été très gentils et disponibles lorsque j'avais des doutes et questions. J'ai grandement apprécié les échanges que j'ai eu avec eux qui m'ont également permis d'améliorer mon espagnol et de me sentir plus confiant avec cette langue.

Je remercie mon professeur de statistiques et probabilités de l'Université Autonome de Madrid, M.Antonio Cuevas, pour ses précieux conseils et aides lorsque j'eus des doutes.

Je remercie aussi Mme.Rosa María Torrado Martín Pallomino du département des ressources humaines de BME pour son aide précieuse lors de toutes les démarches administratives.

Je souhaite également remercier tout ceux qui m'ont soutenu, ma famille, mes amis, pendant ce stage et les moments difficiles rencontrés.

2 Introduction et présentation de l'entreprise

2.1 Notes d'introduction

Le rapport ci-présent a été rédigé en français mais une version a aussi été rédigé en espagnol. Afin de rendre compte de la meilleure compréhension possible et d'appréhender un vocabulaire qui nécessairement est partagé entre plusieurs langues, j'ai essayé de mettre autant que possible et dès que je le savais, les traductions en anglais et en espagnol des termes couramment rencontrés dans le domaine étudié.

2.2 Introduction

Le stage effectué s'est déroulé au sein de la société BME : bolsas y mercados españoles sur une durée d'approximativement 4 mois allant du mois de mai à celui de septembre 2017. L'entreprise est situé à Madrid et la possibilité d'y réaliser un tel stage fait également suite à un séjour d'étude Erasmus de deux semestres à l'Université Autonome de Madrid (UAM : Universidad Autónoma de Madrid), dans le cadre de la dernière année d'étude à l'école d'ingénieur Sup Galilée/Université Paris 13 en spécialité MACS : mathématiques appliquées et calcul scientifique. Au cours de ce stage et au sein du département PTI : post trade interface, j'ai pu apprendre le fonctionnement général de la société et ses divers rôles ainsi qu'être l'auteur de recherches et d'études statistiques afin de répondre à un besoin exprimé par le PTI.

2.3 BME : bolsas y mercados españoles

BME est un groupe d'entreprises financières cotées à l'Ibex 35, qui correspond à l'indice boursier crée par BME elle-même. Cette indice est le principal indice boursier de la bourse de Madrid et est composé de 35 entreprises possédant un poids différent qui varie en fonction de leur capitalisation boursière. BME regroupe plusieurs acteurs et sociétés en son sein dont les principaux sont les suivants :

- Les 4 marchés boursiers les plus importantes du pays à savoir : bolsa de Madrid, bolsa de Barcelona, bolsa de Valencia, bolsa de Bilbao.
- Iberclear : le dépositaire central espagnol.
- MEFF : marché de futures et options espagnols.
- AIAF : marché d'actifs à revenus fixes (Fixed income/Renta fija).
- Latibex : le marché de valeurs latino-américaines, cotées en euro.

2.4 Entités

Je décrirai dans cette partie les différentes composantes de BME, et utiliserai la comparaison avec des équivalents européens afin de mieux comprendre et d'apprivoiser la terminologie qui diffère d'un pays à l'autre.

BME est composée de plusieurs départements ou entités distinctes qui intéragissent entre elles et remplissent leur rôles respectifs pour le bon fonctionnement du système financier actuel. Chez BME, les départements sont les suivants : le marché, la chambre de compensation, le dépositaire central, le PTI. Ces mêmes entités en espagnol pour respecter les termes employés : el mercado, la entidad de contrapartida central (ECC), el depositario central de valores (DCV), el PTI (Post Trade Interface).

Le marché

Le marché correspond au département ou sont effectuées chaques jours d'ouverture, de très nombreuses transactions en temps réel entre les différents intervenants : clients, particuliers, banques, entreprises et tout ceci dans un lieu communément appelé les salles de marché. Au sein du marché même, on peut trouver trois sous-départements distincts en fonction du type de titres (*securities* en anglais ou *valores* en espagnol) ou d'instruments financiers concernés, du mode d'investissement. Le premier est le suivant : en français, un département où sont gérés des instruments financiers à revenu fixe, *fixed income* en anglais. En espagnol, on utilisera le terme *renta fija*. Le second département est l'alternative au premier département ou sont gérés des instruments financiers cette fois-ci à revenu variable, *variable income* en anglais, *renta variable* en espagnol. Celui-ci fait souvent référence aux échanges d'actions ou d'espèces directement (*cash* en anglais ou *efectivo* en espagnol). Le troisième département concerne les produits dérivés financiers (*derivatives* en anglais ou *derivados* en espagnol) comme les futures, les forwards ou les options. Des définitions plus précises sont données en fin de rapport dans l'annexe.

Par exemple, au sein du marché, en *Renta variable*, se déroule une transaction entre deux partis ou clients $C1$ et $C2$. Ceux-ci n'entre pas en contact directement et passe par l'intermédiaire du marché. Chacun d'entre eux passent un ordre (achat/vente ou *compra/venta*) à un membre du marché respectif (*miembro del mercado*) qui à leur tour transfèrent l'ordre au marché. Les membres du marché peuvent être différentes entités à savoir par exemple des courtiers (*brokers*), des banques (*bancos*), des sociétés d'investissements (*sociedad de valores*)...

À l'issue de ces étapes, le marché transfert alors un ordre, appelé *negociación* (littéralement négociation) à l'entité suivante, la chambre de compensation (CCP en anglais ou ECC en espagnol) dans le processus de transaction global.

Chambre de compensation

La chambre de compensation (*clearing house* en anglais ou *cámara de compensación*) est le principal intermédiaire qui agit entre deux partis voulant effectuer une transaction entre eux. C'est un organisme financier assurant le bon déroulement d'une transaction et garantissant à chaque parti ce qui leur est dû. Ainsi la chambre de compensation devient le porteur du risque de contrepartie, c'est-à-dire le risque encouru en cas d'impossibilité d'un des parties à assurer sa fonction dans la

transaction en question. On dit que la chambre de compensation assure le principe de novation (*la novación* en espagnol). Voir l'annexe pour la définition. Pendant le stage, j'ai rencontré les termes ECC ou également CCP faisant tout deux référence à la chambre de compensation selon la langue : *entidad de contrapartida central* ou *central counterparty*. La notion de contrepartie et donc d'intermédiaire aux deux partis (acheteur/vendeur ou débiteur/créditeur) est clairement présente. Les chambres de compensation opèrent généralement à travers plusieurs pays et sont souvent respectives de certains marchés. On cite comme chambres de compensation européennes : LCH.Clearnet pour Paris et Londres, Eurex Clearing, EuroCCP. . . Enfin la chambre de compensation au sein de BME est BME Clearing et agit sur le marché de gré à gré (marché OTC) des produits dérivés financiers.

Au travers de nombreuses procédures complexes, spécifiques selon les cas et veillant à assurer le bon déroulement d'une transaction, la chambre de compensation assure deux rôles principaux :

1. La novation qui comme nous le disions plus haut assure l'intermédiarité de la chambre de compensation entre les deux partis. Ainsi, la chambre de compensation devient le vendeur pour l'acheteur et l'acheteur pour le vendeur.
2. La compensation (*netting* en anglais ou *compensación* en espagnol) qui consiste à déterminer le solde à livrer ou à reçevoir pour chaque parti, solde exprimé en titres, espèces, produits financiers.

Dans le cas de BME Clearing, il existe plusieurs types de membres lui appartenant : les membres compensateurs (simple et pour compte propres) et les membres non compensateurs (individuels et généraux). Les différences entre ces quatre types de membres s'expriment principalement par le droit d'accès ou non aux registres des opérations et aux détails des contrats en cours.

La chambre de compensation et ses divers membres reçoivent donc les ordres émit par le marché. Ces ordres dont le principe de novation s'applique désormais sont alors traités sous formes d'opérations (*operaciones* en espagnol). La chambre de compensation applique également la compensation, couvrant le risque de toute transaction qu'elle gère, en le portant. Cela signifie que si l'un des partis (acheteur ou vendeur) dans la transaction est défaillant (ne peut fournir les titres ou l'espèce), elle se substitue à lui pour régler à la place ce qui est dû. Un tel service n'est néanmoins pas sans contraintes ou coût. Afin de pouvoir remédier aux éventuelles défaillances, la chambre de compensation demande soit dès le début de la transaction une marge que l'on appelle marge initiale (*deposit* en anglais ou *depósito* en espagnol) ou régulièrement durant la transaction des marges appelés marges additionnelles. Ces marges sont fixées et calculées selon des standards et moyens très préçis et dépendent de chaque cas. Des procédures variées peuvent être appliquées dans le cas d'une défaillance de l'un des deux partis. Il y a énormément de cas distincts et cela ferait l'objet d'un document entier, mais on peut au moins retenir que la chambre de compensation peut effectuer une compensation directement en titres lorsque cela est possible, par exemple avec l'aide de contribuables tiers préablement sélectionnés. Si à la fin, il y a toujours défaillance d'un des partis, la chambre peut toujours compenser les opérations en cours par un montant en espèces déterminé également en fonction de la transaction.

Lorsqu'une transaction est validée par la chambre de compensation et que chaque parti peut reçevoir ce qui lui est dû, la chambre de compensation envoit alors des instructions de règlement-livraison

(*settlement instructions* en anglais ou *instrucciones de liquidación* en espagnol) au dépositaire central (réellement, plusieurs instructions sont envoyés, une par membre compensateur de chaque côté de la transaction).

Dépositaire central

Le dépositaire central est un organisme financier ou sont regroupés et comptabilisés les titres et valeurs détenus par le dépositaire central lui-même ou bien par les différents acteurs à savoir les clients, les banques, les courtiers ou encore les intermédiaires financiers. On parlera de dépositaire central en français comme terme générique, de CSD pour *central securities depository* en anglais et enfin de DCV pour *depositario central de valores* en espagnol. On peut comparer cette organisme avec l'image du coffre-fort, lieu ou l'on regroupe une quantité importante d'un bien, ici les titres. Il en existe plusieurs en europe, propre à chaque pays ou bien des dépositaires centraux internationaux existent également : Euroclear et Clearstream. En France on trouvait la CCDVT : caisse centrale de dépôts et de virements de titres, mais celle ci n'existe plus et actuellement l'ensemble des titres sont détenus par Euroclear et plus spécifiquement Euroclear France. Dans le cas de BME, le dépositaire central est Iberclear. D'autres exemples sont CBF : Clearstream Banking Frankfurt ou bien Monte Titoli en Italie.

Iberclear reçoit donc les instructions de règlement-livraison émises par la chambre de compensation, vérifie la concordance des instructions reçues (*matching* en anglais ou *caso/casada* en espagnol) et réalise le transfert des titres entre les comptes des partis en question. Formellement, les instructions de règlement-livraison génèrent à leur tour de nouvelles instructions au sein d'Iberclear même, envoyés à d'autres sous-entités ou participants de la chaîne globale de la transaction. Notamment, Iberclear ou le dépositaire central doivent également envoyer un retour de la bonne livraison des titres et de la transaction entre comptes (*resultado de la liquidación* en espagnol). En l'occurence, le dépositaire central interagit avec les comptes des entités participantes (*entidad participantes* en espagnol) qui sont les sociétés, entreprises, banques dont les titres et produits financiers sous-jacent sont concernés.

PTI : Post Trade Interface

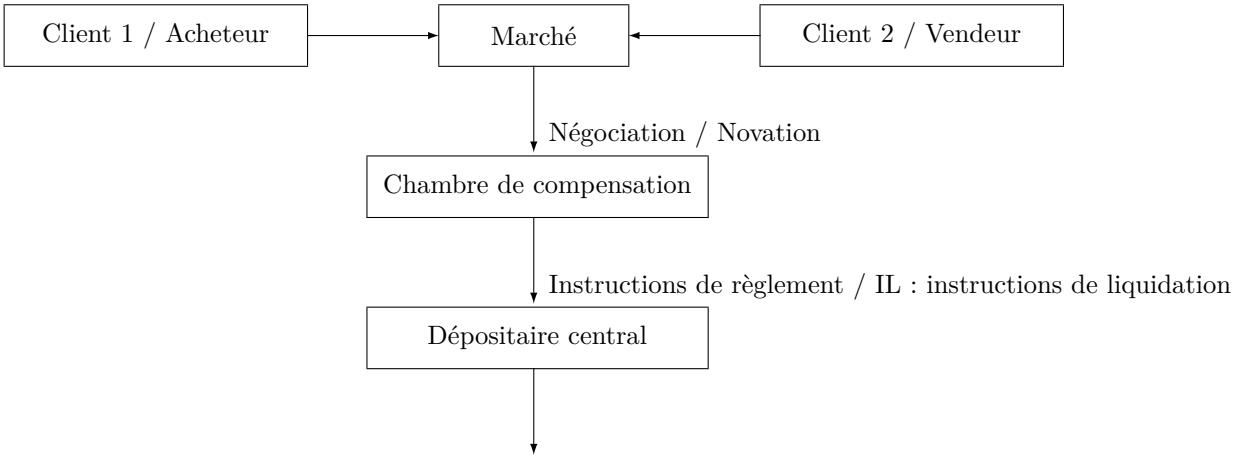
Le PTI (back office) pour *post trade interface* en anglais est un département de BME qui assure un rôle de gestion, de vérification également et fournit les informations entre toutes les entités concernées, celles citées auparavant (marché, chambre de compensation, dépositaire central) mais également d'autres (entités participantes, clients, détenteurs des comptes, membres du marché, membres compensateurs ...). Celui-ci reçoit également les informations au préalable de ces mêmes entités. En particulier : le marché informe des ordres et négociations effectués ou en cours, la chambre de compensation informe des opérations et des instructions de règlement-livraison, le dépositaire central informe de ses propres instructions de règlement-livraison également ainsi que du bon déroulement des règlements en espèces ou en titres (*resultado de liquidación*). Enfin les membres du marché et les entités participantes doivent également communiquer une information spécifiquement pour le PTI : l'appropriation (*la titularidad* en espagnol). L'idée est que l'on puisse avoir les renseignements adéquats et inhérents aux titres ou instruments en question lors d'une transaction (noms, adresses, contacts et informations complémentaires). Ainsi les membres du marché ou entités participantes font ce qu'on appelle en espagnol *comunicar la titularidad* que l'on traduirait approximativement en français par *renseigner l'appartenance*, sous-entendu des titres.

En particulier, les différentes entités ont l'obligation d'envoyer divers informations : les données statiques (titres actifs, membres du marché actifs, comptes concernés, les relations et positions concernant les opérations), les données liées aux instructions de règlement, les données liées aux opérations sur titres (*corporate actions* en anglais ou *eventos corporativos* en espagnol).

Les droits d'accès aux informations et à quels types d'informations sont bien évidemment dûment réglementés, ceci par le PTI. Tout ceci est spécifique du type de transaction en jeu et des partis en question, nous n'allons donc pas rentrer dans les détails.

C'est au sein de l'équipe du PTI que j'ai effectué mon stage, commençant ma formation en apprenant le fonctionnement global des départements de BME et de ses interactions avec le marché espagnol, européen et international.

Résumé du processus global d'une transaction



Cas particuliers

Tout ce qui a été cité précédemment est une manière standard de procéder à une transaction. Cependant, ce n'est pas l'unique manière de faire au sein du marché et de nombreuses autres possibilités existent, faisant intervenir ou non l'une des entités dont nous avons parlé jusqu'à maintenant. Ces cas sont néanmoins plus rares et représentent une part plutôt faible de l'ensemble des transactions effectué (du moins dans le cas de BME : environ 5% des opérations). Par exemple, on peut très bien décider de se priver de l'intermédiaire qu'est la chambre de compensation et donc de ne pas subir les marges imposées par celle-ci au risque de d'exposer au risque de contrepartie dans le cas d'une faillite d'un des deux partis. Ceci est notamment le cas dans les marchés de gré à gré (marché OTC : *over the counter* en anglais). Ainsi dans ce cas, le marché transmet directement les ordres (négociations) au dépositaire central. Un autre cas est celui ou ni le marché, ni la chambre de compensation n'interviennent dans le processus. Seul le dépositaire central joue un rôle. Un dernier cas est celui où les transactions s'effectuent spécifiquement entre la chambre de compensation et le dépositaire central, car en effet le dépositaire central possède également ses propres comptes et titres et peut agir en tant qu'un des partis de la transaction.

3 Sujet de stage

3.1 Analyse du problème

Introduction du problème

Le problème proposé est un travail concernant le suivi et le contrôle d'opérations effectuées au sein de BME. Le travail en question concerne l'étude de données quotidiennes que gère l'entreprise afin d'en déduire des informations et d'entreprendre des démarches appropriées.

Données étudiées

Les données étudiées sont des ratios d'efficience, ou plutôt d'inefficience (*ratios de incumplimientos* en espagnol) compris entre 0 et 1 que possèdent chaque entité. L'inefficience concerne ici la liquidation ou le règlement des titres et montants en espèces. Ce qu'on appelle entité ici sont les entités participantes, donc des sociétés, entreprises ou encore des banques ou bien même le marché comme défini dans la partie 2 de ce rapport. Ces ratios sont quotidiens et peuvent également être en plusieurs exemplaires selon le type de paiement (APMT/FREE ou *contra pago/libre de pago* en espagnol ou encore DvP/FoP en anglais, voir l'annexe) ou bien selon le sens de la transaction (achat/vente). Nous avons donc accès à une liste de ratios compris entre 0 et 1 par entité remontant jusqu'à une date donnée (un an environ dans l'exemple que nous donnerons). Cette liste s'enrichie quotidiennement de la valeur du jour. Dans l'étude menée, plusieurs fichiers semblables seront amenés à être examinés. Les ratios d'efficience sont calculés à partir de valeurs quotidiennes concernant un volume de transactions menées : titres réglés (*settled securities* en anglais ou *valores liquidados* en espagnol), titres non réglés (en attente de règlement, *securities pending settlement* en anglais ou *valores pendientes de liquidación* en espagnol), espèces réglées (*settled cash* en anglais ou *efectivo liquidado* en espagnol), espèces non réglées (en attentes de règlement, *pending cash* en anglais ou *efectivo pendiente de liquidación*). Pour faire simple et synthétiser l'ensemble, un volume liquidé/réglé (*settled* ou *liquidado*) et un volume en attente de liquidation/non réglé (*pending* ou *pendiente*).

Exemple :

Date	ID Entité	Paiement	Tit.liq	Tit.att	Cash.liq	Cash.att	Ratio
01/06/2016	1	APMT	896985	16165	5418508	178419	0.0319

Table 1 – Exemple d'une donnée d'un jour

Ainsi pour cet exemple, le ratio est définit de manière à rendre compte de l'efficience de la liquidation un jour donné et se calcul comme suit, pour le type de transaction APMT :

$$Ratio = \frac{Cash.att}{Cash.liq + Cash.att}$$

Pour un type de transaction FREE, la formule utilise à la place les titres en question :

$$Ratio = \frac{Tit.att}{Tit.liq + Tit.att}$$

avec comme notations : tit pour titre, liq pour liquidé, cash le montant en espèce, att pour en attente de liquidation.

Ceci n'est qu'un exemple de type de données étudiées. Lors de mon stage, j'ai commencé avec ce type de données et passé donc beaucoup de temps avec. Cependant, une fois que le travail eut été plus accompli, les démarches à entreprendre eurent été choisis et les codes eurent été écrits, on me confia alors d'autres données similaires, mais dont les ratios d'efficience se calculaient un peu différemment en fonction des paramètres d'entrées qui différaient (je parle ici des valeurs Tit.liq, Tit.att, Cash.liq, Cash.att). Ceci n'a pas été grave car l'important au final était d'avoir une liste de ratios et ce fut précisément ces ratios qui nous intéressaient pour l'étude menée.

Travail demandé

Le travail proposé concerne l'étude de ces données. L'entreprise acquiert chaque jour ces données afin de rendre compte de l'activité des entités et de leur capacité à liquider les titres ou le cash impliqué. Le but est de définir une démarche complète et de trouver des seuils d'alertes propres aux ratios à partir desquels l'entreprise BME et plus précisément le PTI considèrera qu'il y a un problème dans la liquidation. Auparavant de l'étude menée, BME et donc le PTI utilisait déjà un système afin de définir ces seuils d'alertes. Cependant, les démarches entreprises dans ce but n'étaient pas justifiées et il manquait une base pour expliquer pourquoi agissait-on ainsi.

Démarches utilisées auparavant :

Chaque jour, la valeur du ratio est comparée à la moyenne des valeurs des ratios des 60 jours précedents sommée à deux fois l'écart type. Si l'entité en question est le marché, on sommera cette fois-ci à non pas deux fois mais trois fois l'écart type.

Cette méthode était basée sur la "*règle des trois sigmas*" (ou encore "*règle 68, 95, 99.7*") que l'on rencontre dans de nombreux domaines et qui fait intervenir pour justification la loi normale.

Énoncé : "*Pour une distribution normale, presque toutes les valeurs se situent dans un intervalle centré autour de la moyenne et dont les bornes se situent à 3 écarts types de part et d'autre.*"

Bien évidemment "*presque toutes les valeurs*" est une notion vague et cela signifie ici que 68% de la population se situe à un écart type de part et d'autre de la moyenne, 95% à deux écarts types et 99.7% à trois écarts types. Pour une variable aléatoire X suivant une loi normale $N(\mu, \sigma^2)$:

$$P(\mu - \ \sigma \leq x \leq \mu + \ \sigma) \approx 0.6827$$
$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$
$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$$

Plusieurs problèmes se posent alors à nous concernant l'utilisation d'un tel procédé. Le premier étant le fait que son utilisation suppose que les données en question suivent une loi normale, ce qui ne semble pas être le cas ici. En effet nous savons que le support des données est le domaine $[0, 1]$. Nous pouvons également le vérifier visuellement en traçant les histogrammes des données et voir que la courbe de la distribution ne semble pas suivre une distribution normale. Enfin nous pouvons utiliser différents tests d'hypothèses statistiques afin de conclure de manière plus précise quant aux lois en question. Nous verrons ceci dans la partie suivante *3.3 Démarches statistiques et inférence de distributions*.

Ci-dessous des exemples de tracés d'histogrammes obtenus à partir des 90 derniers jours (un ratio par jour) pour deux entités distinctes.

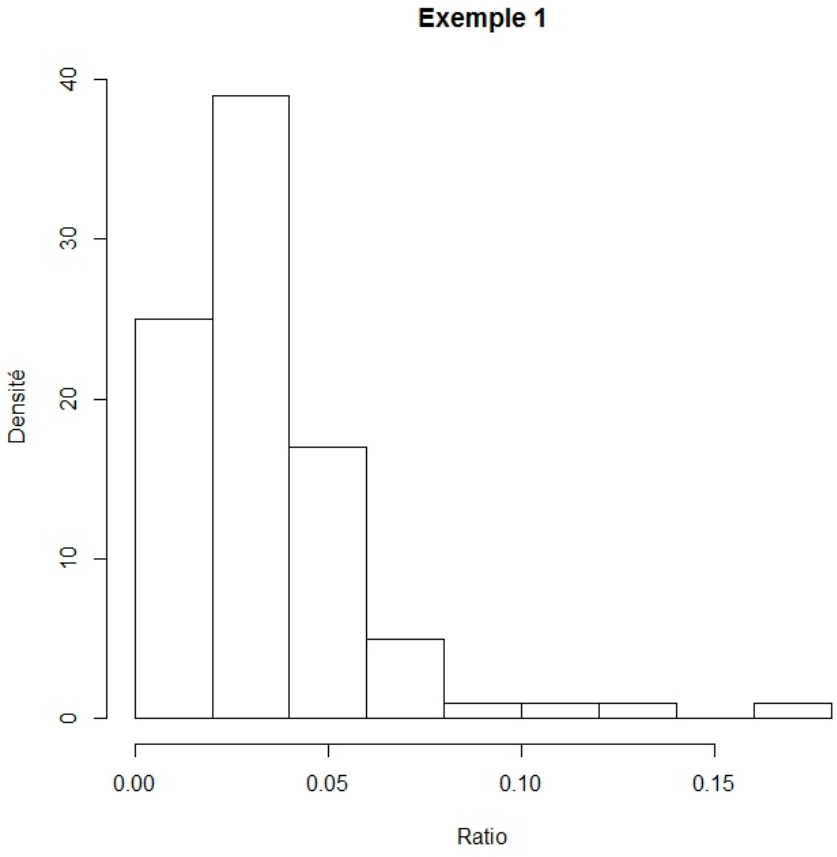


FIGURE 1 – Histogramme des données étudiées

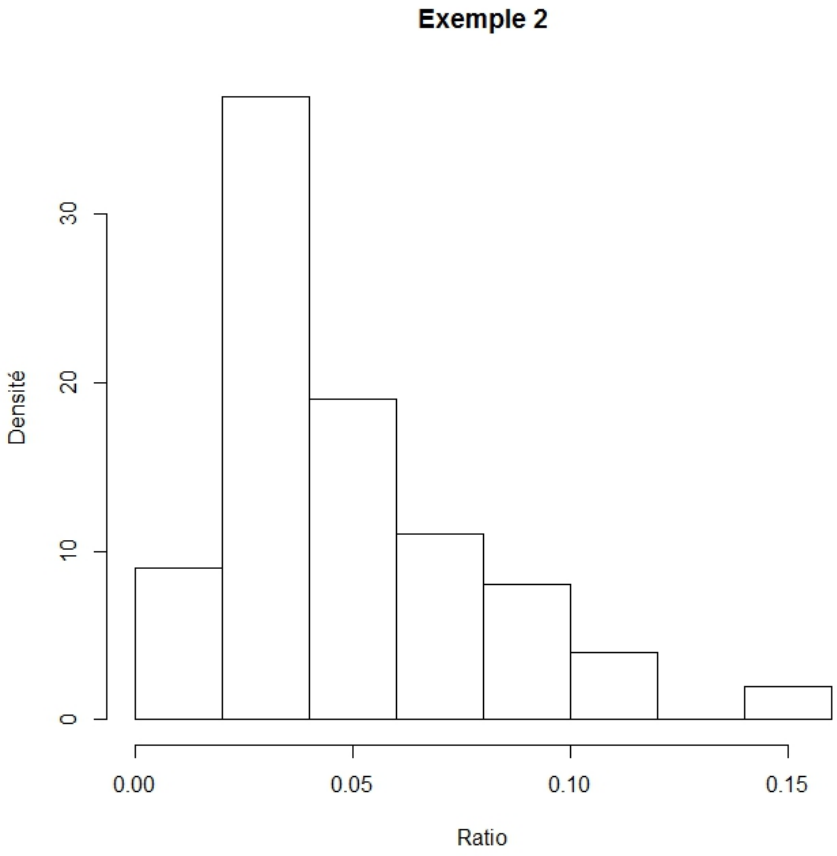


FIGURE 2 – Histogramme des données étudiées

3.2 Support de travail : logiciel R

FIGURE 3 – Logiciel R, version 3.1.4 utilisée

Afin de répondre à la requête fournie, j'étais donc contraint d'effectuer une analyse statistique. Le logiciel et langage utilisé eut été R, qui répondait parfaitement à la situation pour plusieurs raisons. Le logiciel est tout d'abord libre de droit et gratuit, étant donné qu'il n'était pas encore utilisé par le PTI, cela facilitait les démarches d'installation et d'acquisition. Je précise que l'intention du PTI était de pouvoir utiliser une procédure fiable et rapide et ce de manière quotidienne : les seuils d'alertes seront detérminés et calculés chaque jour à partir de données antérieures. Ensuite R est un logiciel et langage relativement utile et proposant de nombreuses extensions en fonction des données étudiées, en fonction de l'étude statistique à laquelle on s'intéresse. Ainsi, étant donné que la recherche d'une réponse au problème posé s'est faite de manière progressive, j'ai pu utiliser et me renseigner grâce à la très large documentation propre à R et à ses extensions, à leur utilisation et aussi grâce aux nombreuses sources d'informations désormais accessibles.

Les données fournies étaient initialement contenues dans un fichier Excel. J'ai effectué mes premières recherches sur Excel même mais j'ai rapidement compris que les possibilités étaient limitées et peu pratiques. Tracé de graphiques, de distributions, statistique descriptive et statistique inférentielle sont beaucoup plus réalisables via R. Il m'a suffit donc d'importer les données nécessaires des fichiers Excel afin de les exploiter dans R.

3.3 Démarches statistiques et inférence de distributions

Démarche initiale

Afin de répondre au problème posé, l'idée initiale était la suivante :

1. Considérer les ratios d'efficience de chaque entité en question, un échantillon composé des 90 dernières valeurs de ratio d'efficience antérieures à aujourd'hui (pour trouver le seuil d'alerte du jour). Le nombre 90 n'est pas choisi au hasard et le détail de ce choix est expliqué dans la partie 3.4 qui suit, *Problèmes rencontrés*.
2. Inférer la meilleure distribution possible à l'échantillon de données considéré, celle qui correspondrait le mieux et représenterait au mieux la tendance indiqué par l'échantillon. Bien entendu les termes "meilleure, mieux" sont vagues et il faut définir en quel sens une distribution est-elle plus adaptée qu'une autre pour modéliser un comportement, une tendance représentée par l'échantillon de données.
3. Une fois cette distribution trouvée, définir à l'aide de celle-ci le seuil d'alerte que l'on souhaite avoir, indicateur de l'anormalité dans la liquidation ou règlement des titres et espèces.
4. Comparer dès lors la valeur du ratio du jour de chaque entité avec leur seuil d'alerte respectif défini et conclure quant à l'anormalité et à l'efficience du processus de liquidation ou règlement de chacune des entités.

Ceci est l'idée initiale à partir de laquelle se sont développées les recherches que j'ai effectué. Habituellement, lorsque l'on souhaite inférer une distribution sur un échantillon de valeurs, on commence par essayer de représenter les données afin d'avoir un premier aperçu et que la visualisation nous donne des pistes à suivre : histogrammes, graphiques, courbe de la densité (estimation par noyau ou *kernel density estimation* en anglais). Ensuite grâce à ces représentations et à l'expérience, on peut deviner et déterminer les possibles distributions qui s'adapteront le mieux aux données. Enfin une fois les distributions inférées et donc les paramètres trouvés, l'important restera de vérifier la qualité de l'inférence faite et de valider ou bien de choisir entre plusieurs distributions une manière de les départager.

Cependant cette démarche ne s'applique pas aussi facilement dans mon cas, car la procédure se devait d'être au maximum la plus automatisé possible afin de pouvoir l'utiliser chaque jour et pour de nombreux échantillons à chaque fois. Ainsi afin de répondre à ce besoin plus important, j'ai commencé par choisir un panel de distributions possibles, à tester. On testerait donc, à l'aide justement de tests statistiques, si les échantillons peuvent suivre l'une ou plusieurs des lois choisis (en effet les tests statistiques peuvent être positifs pour plusieurs distributions certaines fois). On inférerait ensuite pour chaque échantillon de données les paramètres des distributions dont les tests effectués nous ont donné un résultat positif. Si les tests étaient négatifs pour une distribution, cela signifiait qu'inférer des paramètres n'aurait déjà pas beaucoup de sens, car l'échantillon n'avait que très peu de chance d'être représentatif de cette distribution. Enfin on terminerait cette démarche par la validation et le choix de la distribution la plus appropriée, si tant est qu'il y en a plusieurs qui correspondraient.

Distributions

Une fois les données en possession, donc les ratios d'efficience, la première étape eut été de trouver une distribution, donc dans un jargon plus mathématiques de réaliser l'inférence de distributions sur les données. Il existe en théorie une infinité de distributions, en fonction du type de distribution, en fonction des paramètres des distributions (et il existe très probablement en plus de nombreuses distributions encore non représentées ou non imaginées). Le fait est qu'il faut initialement se restreindre et choisir en fonction du domaine d'étude. Chaque fois que l'on souhaite faire une étude statistique, la première étape est de comprendre les données sur lesquelles on travaille et ensuite d'intérpreter les résultats, prendre des décisions en fonction de ces données.

Ici les données sont des ratios compris entre 0 et 1, elles sont donc continues. Nous allons donc déjà nous restreindre aux distributions continues au moins à valeurs dans l'intervale $[0, 1]$. Ma première intuition eut été que l'on prenne en considération les distributions connues à valeurs dans $[0, +\infty[$ et que même si l'on ne peut avoir que des ratios compris entre 0 et 1, on ne considèrerait que les restrictions de ces distributions à cet intervalle. Ainsi, les distributions que j'ai initialement pris en compte sont les suivantes : normale, log-normale, exponentielle, bêta, gamma, weibull. Ce sont des distributions connues et classiques, l'intérêt principal est le fait qu'elles ont déjà été étudiées et sont donc plus manipulables, notamment à travers le logiciel R utilisé. Pour autant, elles sont aussi déjà représentatives d'un grand nombre de tendances ou phénomènes.

— **Normale :** la distribution normale est très répandue et commune de par ses propriétés. Un des meilleurs tests élaboré à ce jour est le test de Shapiro-Wilk[5] qui permet de tester rapidement et avec fiabilité si un échantillon suit une loi normale.

— **Log-Normale :** la distribution log-normale est aussi très courante. Restreinte à l'intervalle $]0, +\infty[$, elle permet de représenter des populations ne pouvant accepter que des valeurs positives. On peut également tester facilement la log-normalité d'un échantillon à l'aide du même test, celui de Shapiro-Wilk : en effet une variable aléatoire X suit une loi log-normale $LN(\mu, \sigma^2)$ si la variable aléatoire $Y = \ln(X)$ suit une loi normale $N(\mu, \sigma^2)$. On peut donc réutiliser les tests de normalité.

— **Bêta :** le support de la distribution bêta est $[0, 1]$ ce qui semble parfaitement adapté pour des données comprises dans ce même intervalle. De plus cette famille de distributions est souvent utilisée pour ce type de données continues en pourcentage, donc comprises entre 0 et 1.

— **Gamma, weibull, exponentielle :** ces distributions à support dans $[0, 1]$ sont également envisageable. La forme de ces distributions peut également bien correspondre aux données que l'on possède : la densité présente souvent une croissance rapide, une atteinte du maximum et une décroissance plus faible ensuite.

De manière générale, les motivations qui m'ont poussé à choisir ces distributions à tester sont les suivantes :

— Le support sur lequel les données se trouvent, à savoir ici l'intervalle $[0, 1]$ ainsi que le type de données que l'on possède (continues ou discrètes).
— L'expérience et l'intuition. Ce sont des critères inquantifiables mais pourtant clés et importants dans de telles recherches.
— La facilité et la fiabilité des tests d'hypothèse statistique ainsi que des inférences que l'on peut réaliser sur ces distributions. Par exemple, la loi normale très étudiée et très connue possède plus d'une 30 de tests différents, dont certains sont très robustes. Ainsi la vérification et tester l'adéquation avec la loi normale ne pose pas de problèmes techniques.
— Certaines distributions sont au contraire inenvisageable car ne correspondent pas au domaine étudié. En effet celles-ci ont été écrites pour répondre à un besoin d'étude bien spécifique. Ce type de loi a rapidement été écarté car inimaginable et inadaptable dans de telles circonstances.
— Avec l'ensemble des échantillons en ma possession, j'ai également tracé le graphique de Cullen-Frey pour chacun d'entre eux afin d'avoir une idée global, certes approximative mais pas moins utile des distributions en jeu. Le graphique de Cullen-Frey représente l'aplatissement (*kurtosis* en anglais) en fonction du carré de l'asymétrie (*skewness* en anglais). Ce type de graphe ne permet pas de conclure si une distribution suit telle ou telle loi car les moments d'ordre 3 et 4 ne suffisent pas à caractériser une distribution (tout comme les moments d'ordre 1 et 2 que sont la moyenne et l'écart type). Mais il donne au moins un ordre d'idée des tendances en question.

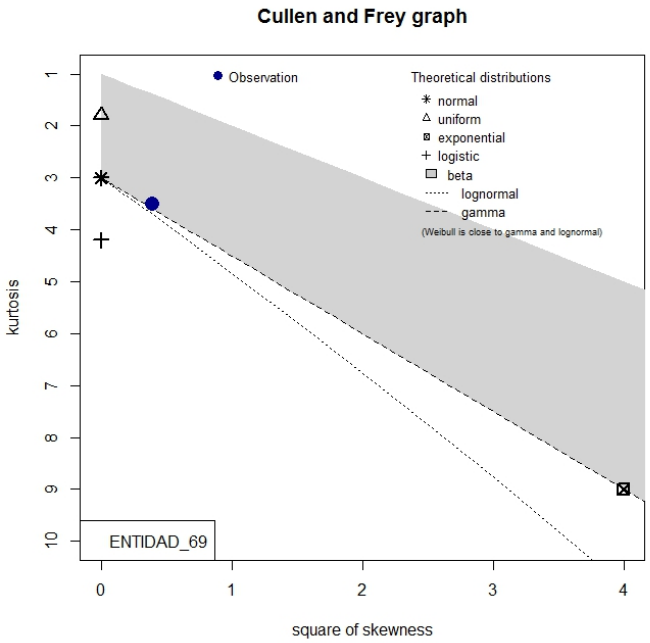


FIGURE 4 – Graphique de Cullen-Frey pour un jeu données (*Entidad 69*)

Méthodes d'inférence

La première étape consistait donc à identifier les possibles distributions. Maintenant le tout et de vérifier que cela correspond bien. Nous utilisons pour ceci des tests statistiques : le test de Shapiro-Wilk pour tester la normalité et la log-normalité (il suffit d'appliquer le test de Shapiro-Wilk sur le logarithme néperien de l'échantillon). On cite également les tests d'Anderson-Darling, le test de Kolmogorov-Smirnov, le test de Lilliefors. La performance de l'un par rapport à l'autre fait l'objet d'une étude complète[5] mais on peut retenir les informations suivantes :

— Les tests de Shapiro-Wilk et Anderson-Darling sont très fiables et performants mais ne permettent de tester que la normalité.

— Le test de Kolmogorov-Smirnov ne fait sens que lorsque que l'on connaît déjà les paramètres de la distribution testée. Si les paramètres ne sont pas connus et que l'on doit les estimer préalablement, on choisira alors le test de Lilliefors. Cependant le test de Kolmogorov-Smirnov permet aussi de tester d'autres types de distributions que la loi normale dès lors que l'on connaît la fonction de répartition.

Lors de cette étude, j'ai donc d'abord utiliser les tests statistiques afin de renforcer mes hypothèses concernant le fait que si oui ou non, un échantillon suivait l'une des lois citées auparavant. J'ai également utilisé les tracés des histogrammes ainsi que les QQ-plots. Pour rappel, un QQ-plot est un graphique comparant les quantiles de notre échantillon avec les quantiles de la loi supposée. Si sur le graphique, les points s'alignent globalement sur la droite affine $y = x$, alors l'échantillon de données a de fortes chances de bien suivre la loi supposée.

Une fois les vérifications appliquées, la seconde étape consistait à déterminer les paramètres des distributions en question. L'inférence est ici paramétrique car l'on souhaite avoir le maximum d'informations pour pouvoir déterminer, rappelons-le, nos seuils d'alerte sur les ratios d'efficience des entités. Afin de déterminer les paramètres j'ai utilisé diverses méthodes. Pour certaines lois, on connaît des estimateurs des paramètres, comme l'estimateur du maximum de vraisemblance : on définit la vraisemblance puis on résout un problème d'optimisation. On peut aussi estimer les paramètres à l'aide de la méthode des moments. Voici quelques exemples d'estimateurs de paramètres selon les lois :

— X suit une loi normale $N(\mu, \sigma^2)$:

$$\text{Estimateur de la moyenne : } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Estimateur sans biais de la variance : } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

— X suit une loi exponentielle $\varepsilon(\lambda)$:

$$\text{Estimateur du paramètre } \lambda : \overline{\lambda} = \frac{1}{\overline{x}} \text{ où } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

— X suit une loi bêta $\beta(a, b)$:

$$\text{Estimateur de } a : \overline{a} = \frac{E(X)(E(X) - E(X)^2 - V(X))}{V(X)}$$

$$\text{Estimateur de } b : \overline{b} = \frac{E(X) - 2E(X)^2 + E(X)^3 - V(X) + E(X)V(X)}{V(X)}$$

où $E(X)$ et $V(X)$ sont respectivement l'espérance et la variance de X dont on peut prendre les estimateurs $\overline{x}$ et s^2 également pour les calculer.

— X suit une loi gamma $\Gamma(k, \theta)$:

$$\text{Estimateur de } k : \overline{k} = \frac{E(X)^2}{V(X)}$$

$$\text{Estimateur de } \theta : \overline{\theta} = \frac{E(X)}{V(X)}$$

— X suit une loi log-normale $LN(\alpha, \beta)$, soit $Y = \ln(X) \sim N(\alpha, \beta)$:

$$\alpha = \mu_y \text{ et } \beta = \sigma_y^2$$

$$\mu_x = E(X) = e^{\mu_y + \frac{\sigma_y^2}{2}} \text{ et } CV_x = \frac{\sigma_x}{\mu_x} = \sqrt{e^{\sigma_y^2} - 1} \text{ le coefficient de variation de } X$$

$$\text{Estimateur de } \alpha : \overline{\alpha} = \ln(\frac{\mu_x}{\sqrt{1 + CV_x^2}})$$

$$\text{Estimateur de } \beta : \overline{\beta} = \ln(1 + CV_x^2)$$

— X suit une loi weibull à deux paramètres $W(k, \lambda)$. Les estimateurs $\overline{k}$ et $\overline{\lambda}$ de k et λ vérifient le système suivant :

$$\begin{cases} -\frac{\sum_{i=1}^{n} x_i^{\overline{k}} \ln(x_i)}{\sum_{i=1}^{n} x_i^{\overline{k}}} - \frac{1}{\overline{k}} - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} x_i^{\overline{k}} - \overline{\lambda} = 0 \end{cases}$$

Nous pouvons alors résoudre ces équations numériquement pour trouver k et λ à l'aide d'un algorithme de Newton ou de point fixe par exemple. J'ai notamment utilisé la fonction *uniroot()* de R afin d'y parvenir.

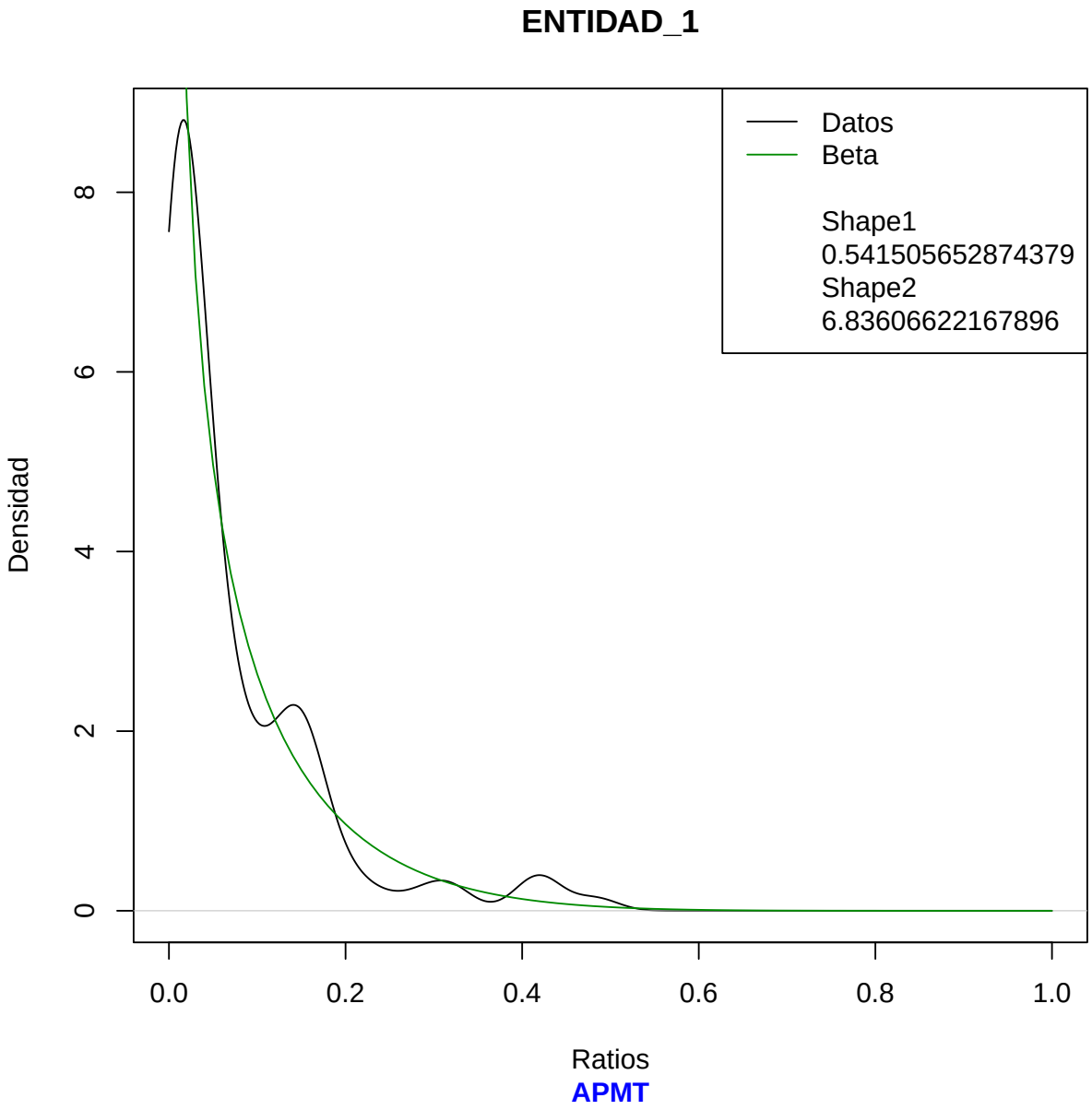
Cependant, afin d'affiner l'étude et parce que nous travaillons au sein du logiciel R, plusieurs extensions fournissent des fonctions déjà implémentées renseignant directement les paramètres. J'ai notamment utilisé le package *MASS* avec la fonction *fitdistr()* mais également le package *fitdistrplus* avec la fonction *fitdist()*.

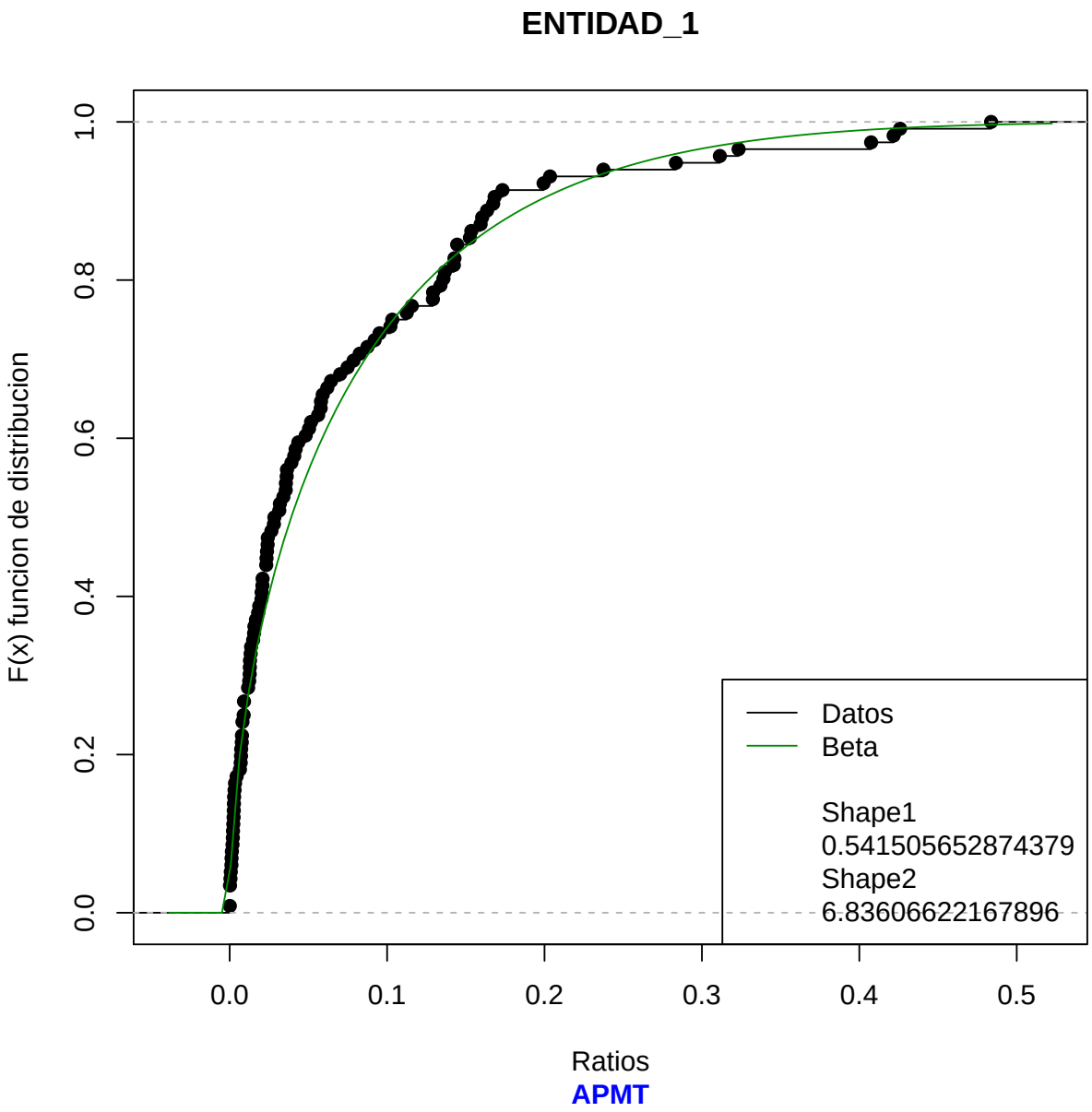
Ci-dessous des exemples de résultats :

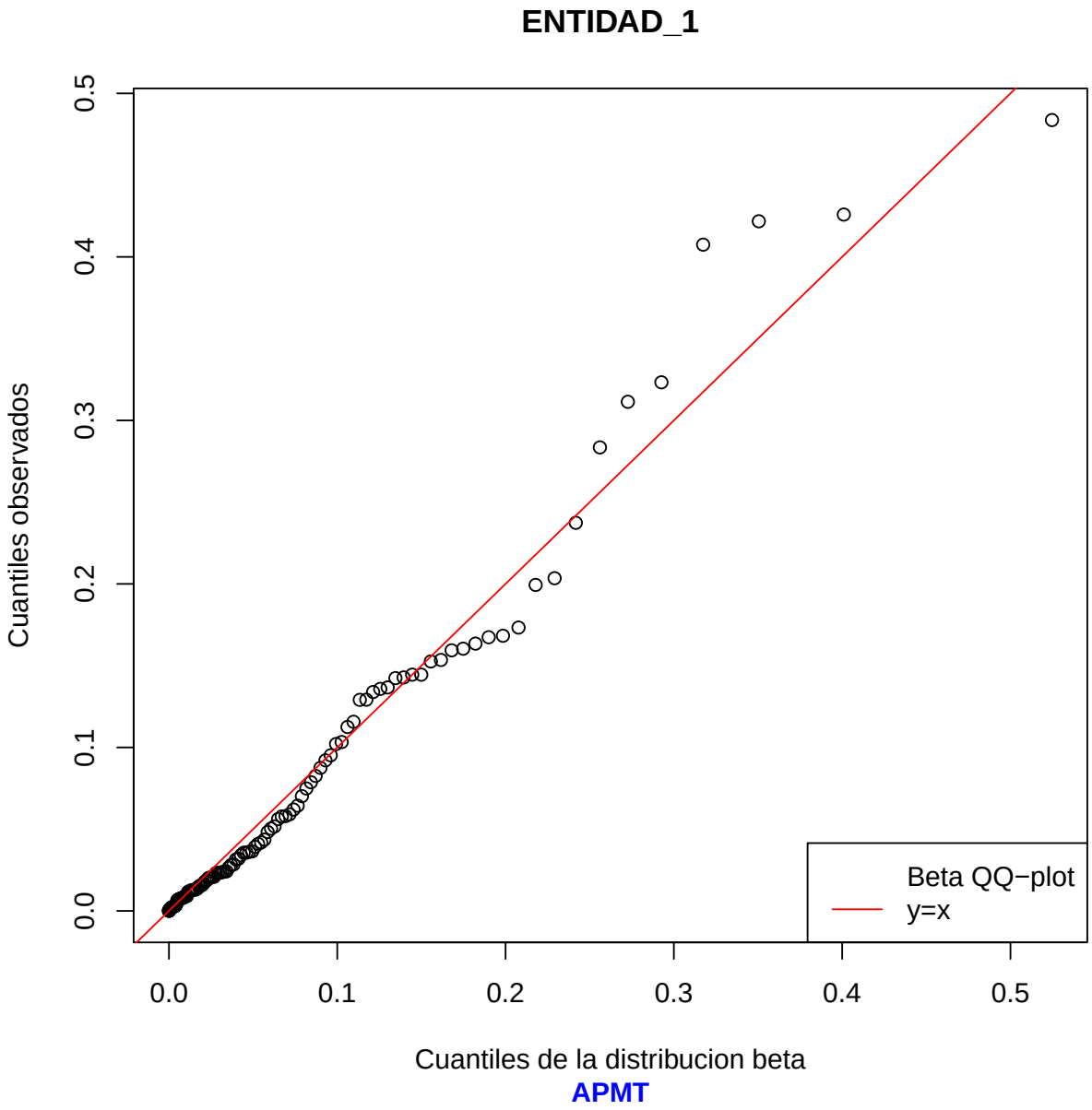
Page 20, 21, 22 : Le tracé des densités (estimation de densité par noyau) des données et d'une distribution supposée (bêta ici), le tracé de la fonction de répartition (*funcion de distribución* en espagnol), le QQ-plot. J'ai utilisé ici le package *MASS*. Les courbes se superposent, la fonction de répartition et le QQ-plot montrent certaines disparités néanmoins, notamment pour des valeurs de ratios élevées. Cela signifie donc qu'une distribution bêta semble adapté mais seulement pour les valeurs centrales et initiales. Il doit donc y avoir des valeurs anormales (pour une distribution bêta) dans les valeurs élevées de l'échantillon testé.

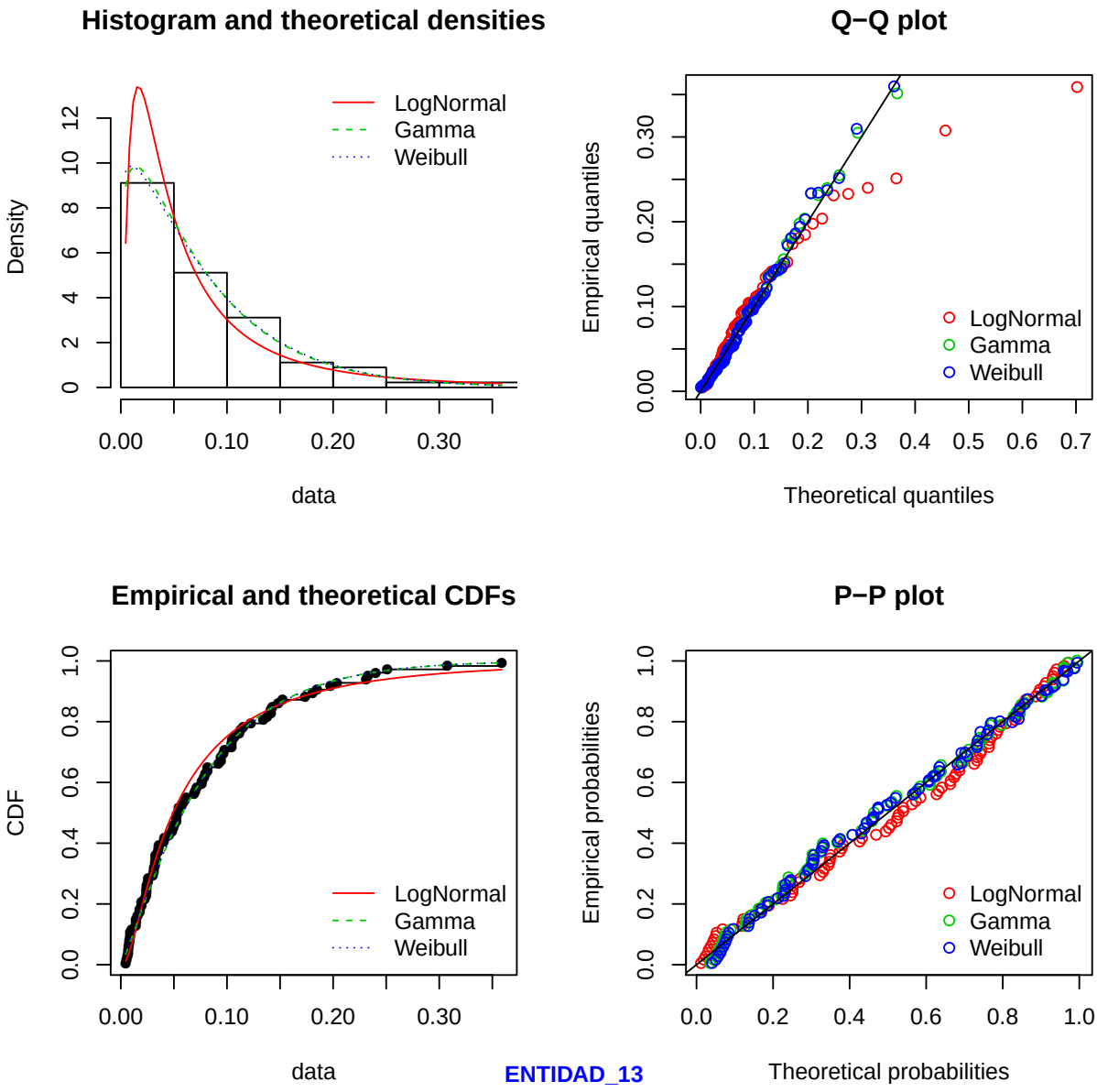
Page 23, 24 : Exemples de tracés similaires, où l'on compare les distributions à l'aide des QQ-plots et PP-plots également. Le QQ-plot (quantiles) renseigne sur l'adéquation d'une loi majoritairement plus sur le centre de la distribution tandis que le PP-plot (probabilités) nous donne des informations concernant les queues de distributions.

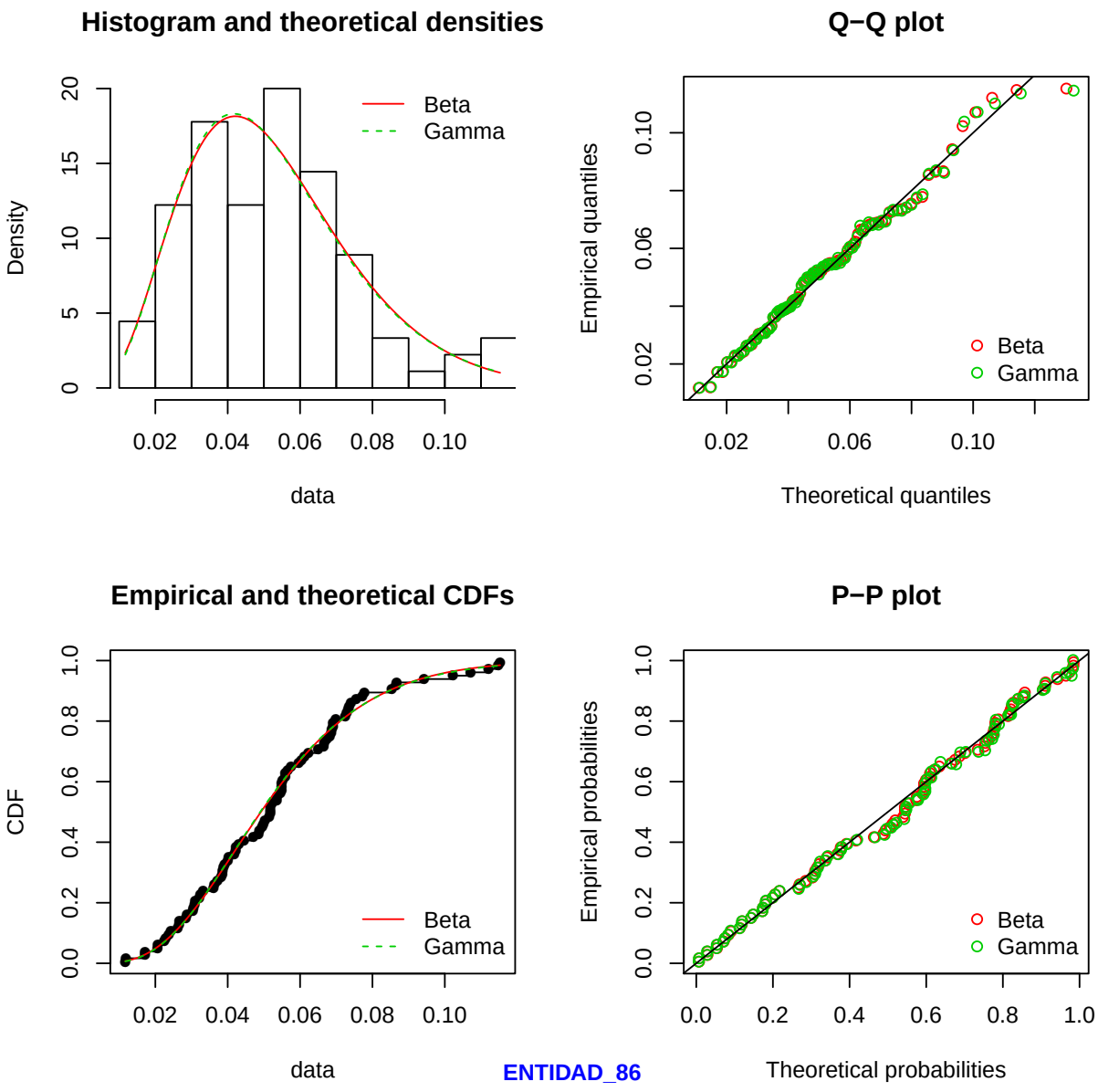
Page 25, 26, 27, 28 : J'ai également cherché à comparer les différentes méthodes d'inférence proposé par le package *fitdistrplus* sur une distribution bêta afin d'en choisir une qui serait idéal pour les échantillons de données que l'on étudie. 4 méthodes peuvent être utilisées : MLE (*maximum likehood estimation*) donc la méthode du maximum de vraisemblance, MGE (*maximum goodness of fit estimation*), MME (*moment matching estimation*), QME (*quantile matching estimation*). Le détail de ces méthodes est expliqué très clairement dans le document [3], nous n'allons donc pas nous attarder sur le fonctionnement de chacune de ces méthodes. Cependant, mon choix s'est porté sur la méthode MME au final, notamment pour sa sensibilité aux valeurs extrêmes. En effet, ne souhaitant pas considérer qu'il y ait de valeurs anormales dans l'échantillon, chaque valeur doit avoir un poids dans la démarche appliquée. Rappelons-le, les données sont des ratios, ce ne sont donc pas des mesures qui peuvent dépendre de la fiabilté d'un appareil les enregistrant, mais bien des valeurs toutes significatives. Ainsi même une valeur extrême (grande par exemple comme 0.99 a une importance car signifie que l'efficience de la liquidation de l'entité ce jour là fut très faible : autrement dit, presque aucun titres ou montant en espèces n'a été envoyé en règlement). Notons enfin que l'on peut également choisir l'algorithme d'optimisation lors du calcul des paramètres (quelque soit la méthode), mais cela n'a pas été étudié ici. Gardons seulement à l'esprit que l'inférence reste très délicate et beaucoup de conditions entrent en jeu.

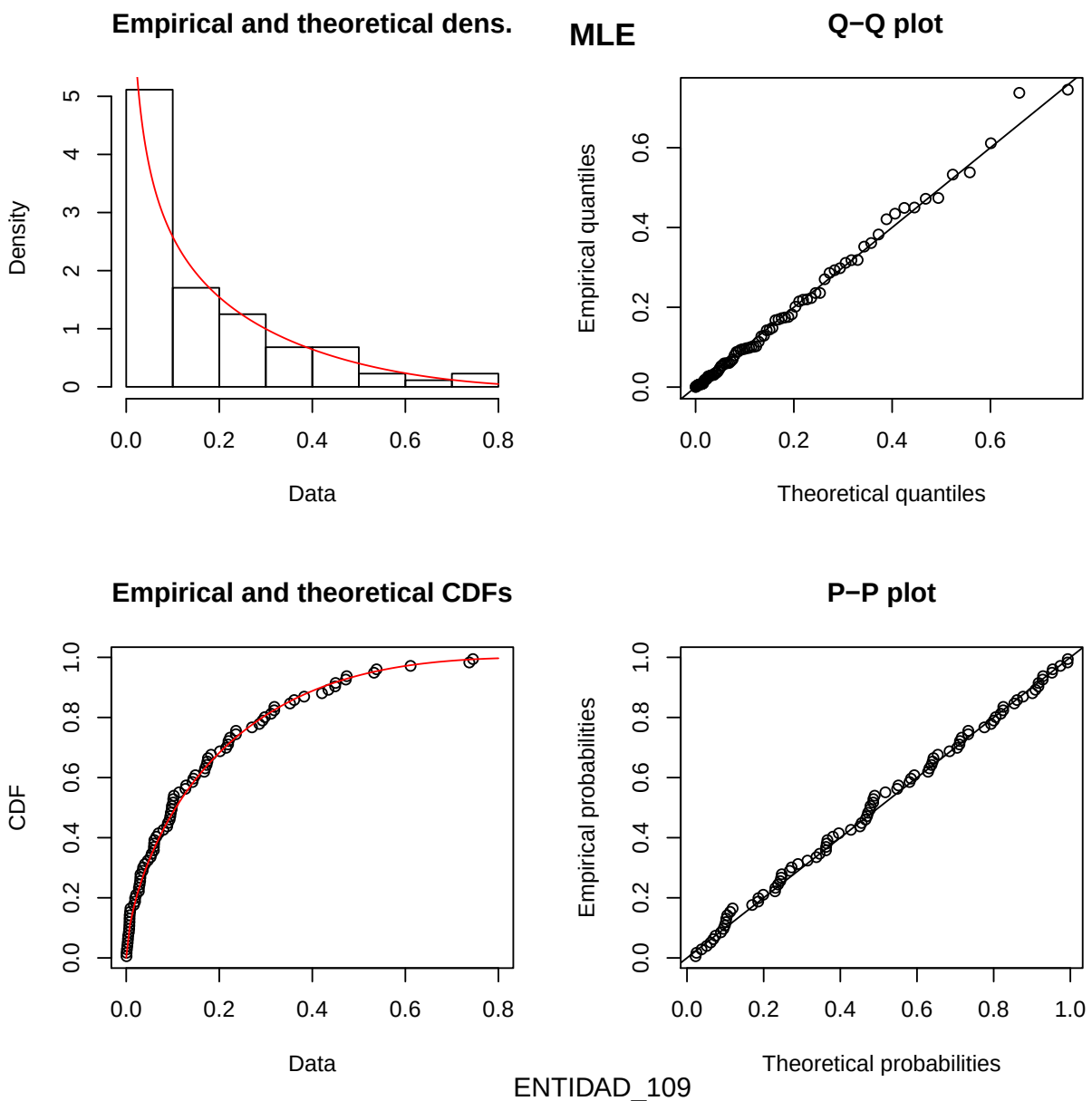

ENTIDAD_1
Datos
Beta
Shape1
0.541505652874379
Shape2
6.83606622167896
Densidad
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Ratios
APMT

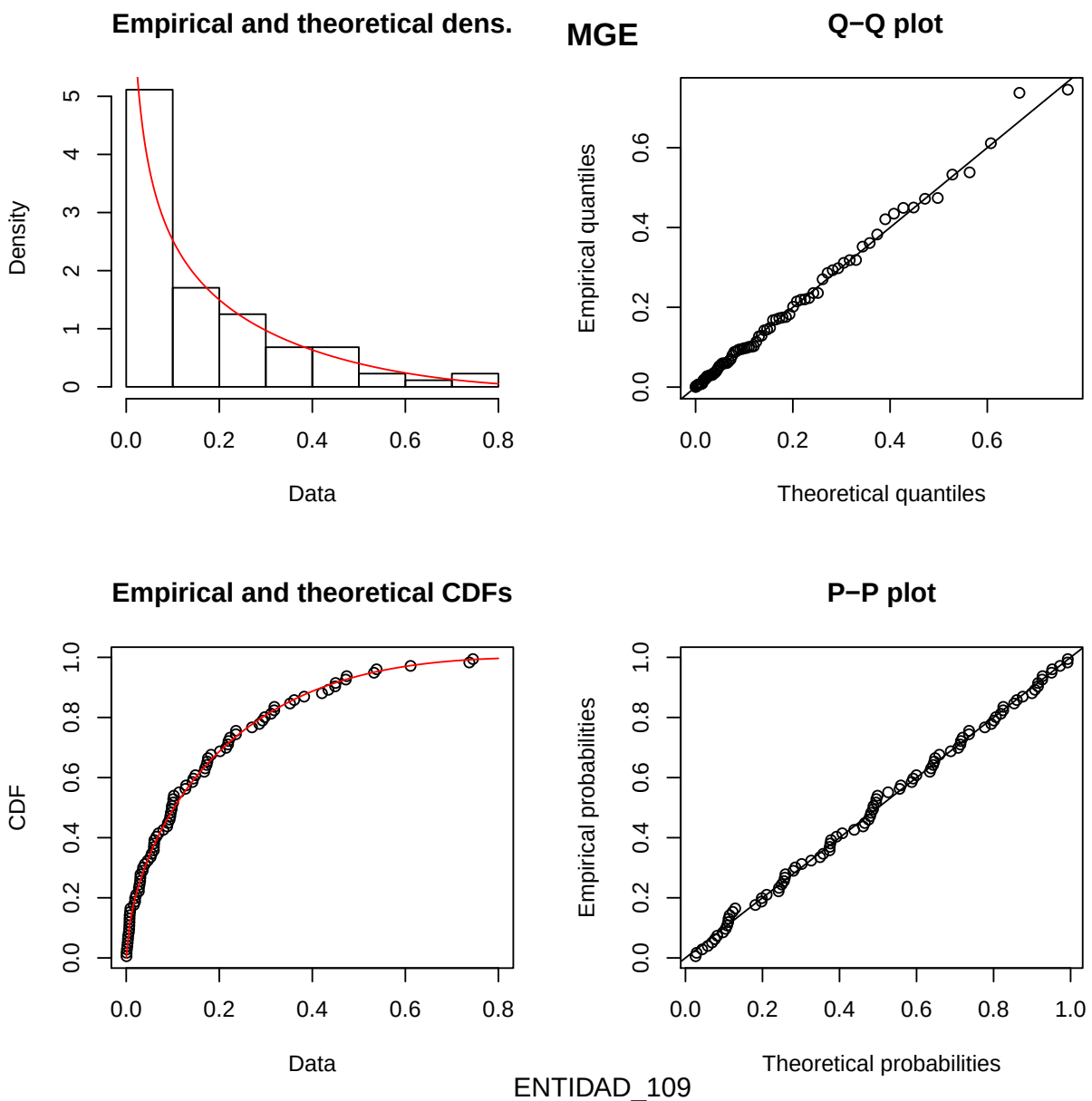

ENTIDAD_1
F(x) funcion de distribucion
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
Ratios
APMT
Datos
Beta
Shape1
0.541505652874379
Shape2
6.83606622167896

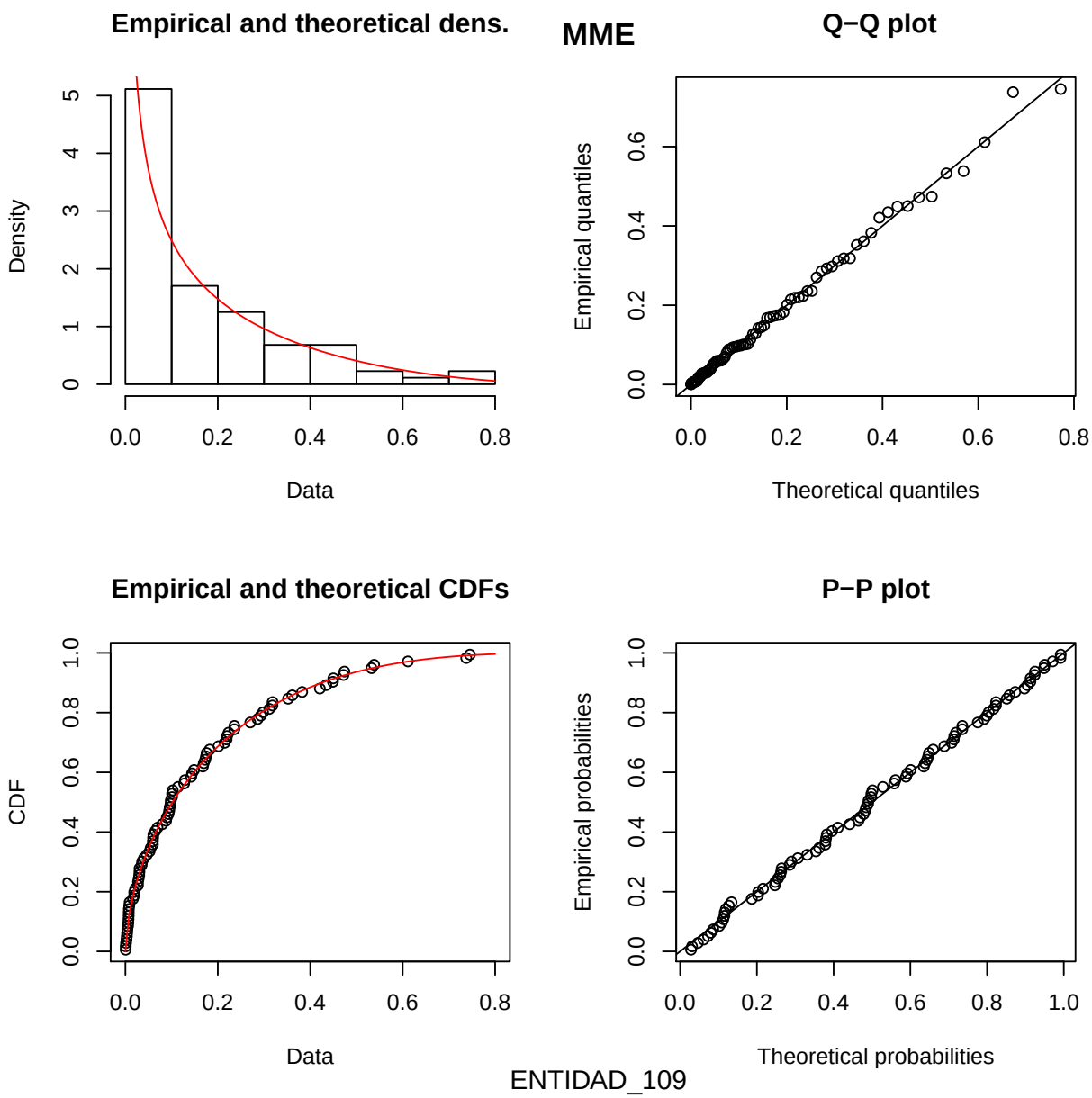

ENTIDAD_1
Cuantiles observados
0.0
0.1
0.2
0.3
0.4
0.5
Beta QQ−plot
y=x
Cuantiles de la distribucion beta
APMT

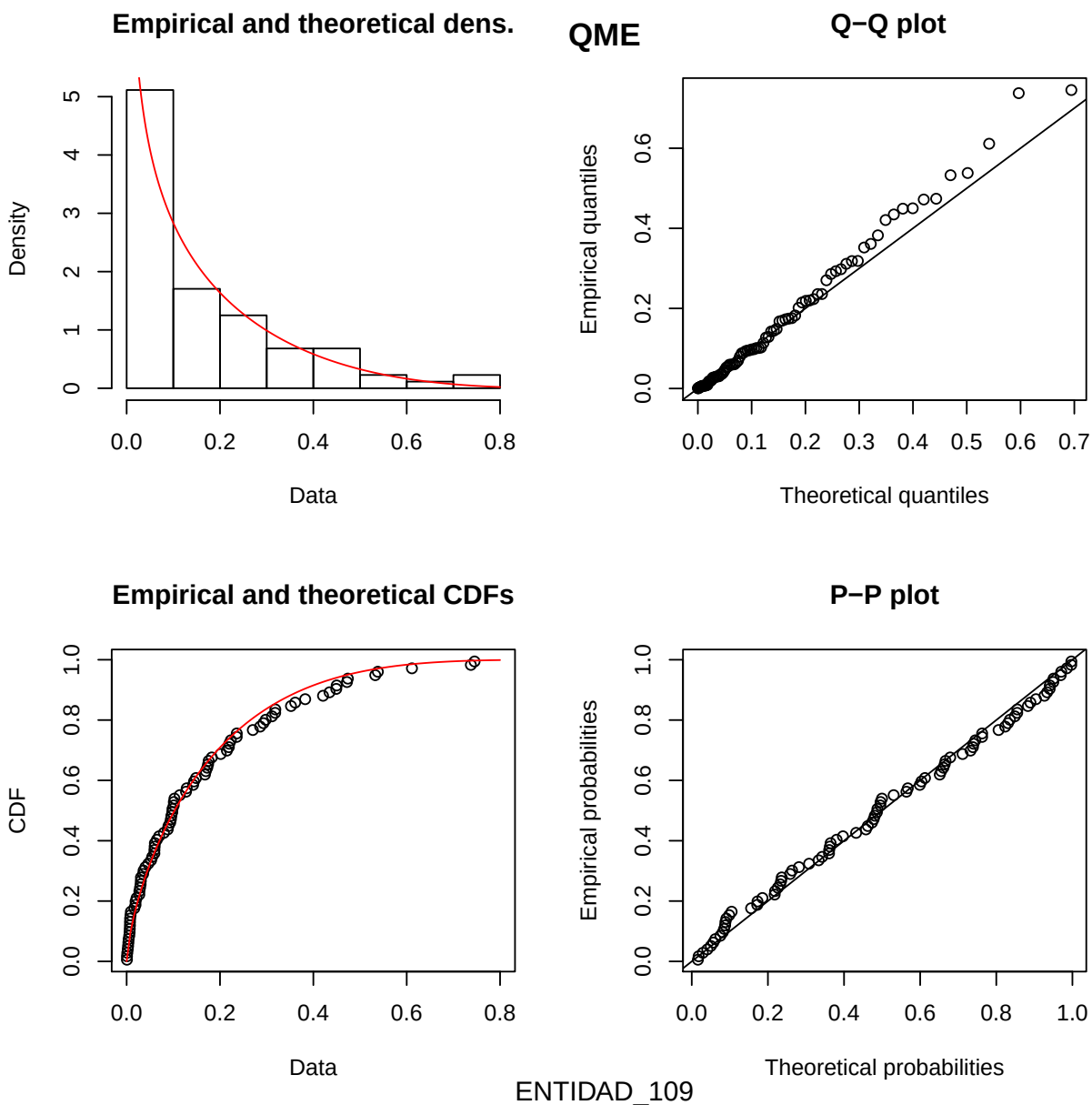

Histogram and theoretical densities
Density
data
LogNormal
Gamma
Weibull
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
P−P plot
Empirical probabilities
Theoretical probabilities
ENTIDAD_13


Histogram and theoretical densities
Density
data
Beta
Gamma
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
P−P plot
Empirical probabilities
Theoretical probabilities
ENTIDAD_86


MLE
Empirical and theoretical dens.
Density
Data
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P−P plot
Empirical probabilities
Theoretical probabilities
ENTIDAD_109


MGE
Empirical and theoretical dens.
Density
Data
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P−P plot
Empirical probabilities
Theoretical probabilities
ENTIDAD_109


MME
Empirical and theoretical dens.
Density
Data
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P−P plot
Empirical probabilities
Theoretical probabilities
ENTIDAD_109


QME
Empirical and theoretical dens.
Density
Data
Q−Q plot
Empirical quantiles
Theoretical quantiles
Empirical and theoretical CDFs
CDF
Data
P−P plot
Empirical probabilities
Theoretical probabilities

ENTIDAD_109

Choix du modèle approprié : goodness of fit

Maintenant que nous sommes parvenu à inférer les paramètres des distributions, à l'aide de R, nous devons étudier le fait que plusieurs modèles semblent parfois s'adapter ou correspondre aux échantillons de données en jeu. En effet, suite aux tests statistiques utilisés, plusieurs répondaient parfois positifs : l'échantillon suit une loi X et ce avec une erreur de première espèce α communément choisi, pour la majorité des tests d'hypothèse statistique, égale à 5%. Pour rappel, l'erreur de première espèce correspond à la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie. Dans notre cas, l'hypothèse nulle H_0 est par exemple : "*l'échantillon suit une loi bêta*".

Ainsi il nous faut un moyen de sélectionner la distribution la plus adéquate pour nos échantillons. La définition de "meilleure distribution" est dépendante de l'étude menée et de ce que l'on souhaite bien évidemment. L'idée est de mesurer une distance entre les distributions inférées et les échantillons correspondants. Il en existe plusieurs que l'on appelle *statistiques de qualité d'ajustement* mais on rencontre plus souvent le terme anglophone *goodness of fit statistics*. Pour les distributions continues que nous rencontrons ici, nous avons notamment les statistiques de Cramer-Von-Mises, Anderson-Darling et Kolmogorov-Smirnov qui sont utilisées, dans les packages de R cités plus haut. En particulier, nous considèrerons les deux premières sachant le fait que nous ne connaissons pas les paramètres des distributions concernées et que nous avons du utiliser des estimateurs. Ainsi la statistique de Kolmogorov Smirnov ne convient pas ici. Les différences sont complexes et dépendent de l'importance que l'on souhaite donner à telle ou telle partie d'une distribution (le centre, les queues de distributions ...). Comme nous n'avons pas d'informations supplémentaires, j'ai considéré la statistique d'Anderson-Darling afin de comparer la qualité des ajustements entre eux. Une des raisons principales de ce choix est le désir de prendre en considération autant les queues de distribution que la partie centrale[3]. Ainsi, la plus petite valeur de ces statistiques nous donne la distribution adéquate à considérer.

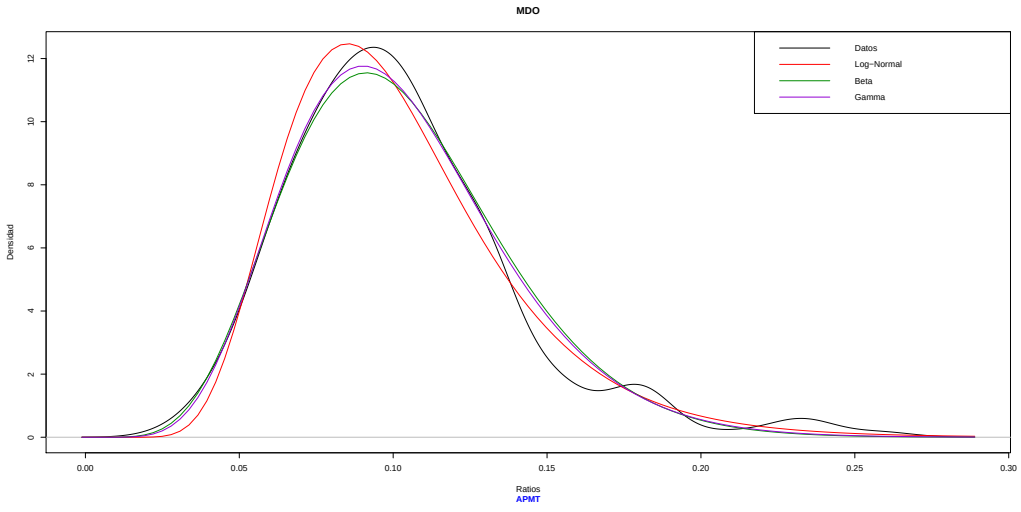


FIGURE 5 – Comparaison des distributions pour les données du marché

Fixation des seuils d'alerte

L'objectif était donc de fixer des seuils d'alerte concernant l'efficience dans la liquidation des titres et des montants en espèces des entités. Maintenant que nous possédons une distribution par entité, représentant le comportement de l'entité sur les 90 derniers jours (car les échantillons possèdent 90 valeurs), nous définissons alors le seuil d'anormalité à partir du quantile à 98% de la distribution. Une fois la distribution trouvée, nous calculons la valeur du quantile à 98% et définissons celle-ci comme le seuil d'alerte du jour de l'entité. Ensuite, si la valeur du ratio du jour est supérieure à ce seuil, nous considérons alors qu'il y a anormalité et nous le signalons. Si la valeur du jour est inférieure, nous déclarons qu'il n'y pas d'alerte.

Le calcul des valeurs des quantiles se fait simplement à l'aide de fonctions implémentées dans les packages utilisés de R.

La valeur 98% n'est bien évidemment pas choisie au hasard. Celle-ci dépends des données, de notre définition de l'anormalité dans notre cas et est l'aboutissement de nombreux tests effectués sur des données réelles, en conditions. L'objectif était d'avoir un seuil qui respecte un certain équilibre entre : avoir suffisament d'alertes pour ne pas rater d'éventuels problèmes mais également ne pas en avoir trop et être submergé d'alertes non ou peu significatives. Simplement, plus la valeur est faible, plus il y a de chances d'avoir d'alertes.

3.4 Problèmes rencontrés

Suite à toutes ces recherches effectuées, j'ai cependant été confronté à certains problèmes remettant en cause la démarche.

Le premier résidant dans la nature des données que l'on étudie. Ces dernières en effet présentent parfois de nombreux zéros, ou à l'inverse moins souvent mais quand même, plusieurs fois la valeur une. Les zéros au sein des échantillons sont donc des ratios d'inefficience nuls, cela signifie donc que tout les titres et montants en espèces ont été liquidé le jour correspondant et donc qu'il n'y a pas de problèmes à signaler. Au contraire, une valeur de un pour un ratio d'inefficience signifie qu'aucun titre ou montant en espèces n'a été réglé. Ceci pose problème dans la démarche expliquée précedemment car comme nous l'avons dit, les distributions inférées sont des distributions continues et lorsque l'on considère de telles distributions, on ne peut avoir deux fois une valeur identique au sein d'un échantillon censé représenter la distribution. Rappelons-le, pour X suit une loi de probabilité continue et pour x appartenant à l'intervalle sur lequel est définie la loi, on a $P(X = x) = 0$. Ainsi on ne peut utiliser les fonctions définies dans R pour inférer les distributions sur ces types d'échantillons (contenant plusieurs 0 ou 1). Le problème est que ces valeurs sont courantes. Afin d'y remédier, j'ai d'abord considéré le fait de ne pas prendre en compte les valeurs égales à 0 (celles égales à 1 étaient plus rares et posaient moins de problèmes). Ainsi je me retrouvais avec des échantillons intialement de 90 jours puis ensuite réduits, leur taille dépendant du nombre de 0. L'idée était que puisque les zéros ne sont pas représentatifs d'un quelconque problème de liquidation, on pouvait accepter de ne pas les considérer et de baser l'étude sur le reste des valeurs. Cependant ceci posait d'autres problèmes, à savoir le fait qu'on se retrouvait certaines fois avec des échantillons au final de taille vraiement trop petite (5, 10 , 13 valeurs non nulles par exemple). Du coup, il était impossible pour les fonctions de fonctionner correctement, car les échantillons dont on réalisait l'inférence ne contenaient pas assez de valeurs. On pouvait également se demander quelle était la pertinence de ne pas prendre en compte les valeurs nulles. Et que faire avec les échantillons comprenant de nombreux 1 ? Ainsi avec cette démarche, certaines entités avait une distribution inférée et d'autres non, ce qui n'était pas rigoureux et abouti.

Un autre problème était le choix de la taille de la période considérée. Celle-ci est fixé à 90 jours. En effet, il nous fallait une période suffisamment grande pour considérer assez de valeurs et pouvoir inférer correctement. De plus il fallait prendre en compte le fait que l'on supprimait les zéros des échantillons se qui réduisait encore la taille. Dans le même temps, la taille de la période considérée ne devait pas être trop grande également en raison d'aspects financiers et propres au fonctionnement du marché et de la liquidation/règlement : on retrouve au sein de BME diverses périodes durant lesquelles les comportements (dans le sens efficience de la liquidation des titres et montants en espèces) varient de manière significative. Nous aurions pu penser que prendre le maximum de données passées n'est que bénéfique pour exprimer et représenter le comportement d'une entité, cependant cela n'est pas vraiment adéquat de ce point de vue, et ne permet pas de prendre en compte le renouvellement des cycles au sein de BME. Ainsi, le choix de la période de 90 jours à considérer s'est fait en fonction de ces critères (pour information, l'ensemble des ratios que j'avais en ma possession pouvait remonter à un an voir plus).

3.5 Nouvelle piste d'étude

Suite à ces problèmes rencontrés, il a fallu trouver une solution pour y remédier et revoir la démarche établie alors. Les valeurs nulles jusqu'alors ignorées et mises de côté se devaient au final d'être prise en compte, en effet celles-ci représentent une absence de problème de liquidation, tout comme des valeurs de ratios très faibles (0,01 par exemple). Une idée suggérée alors par mon professeur de statistiques et probabilités de l'Université Autonome de Madrid, M.Antonio Cuevas, eut été de considérer non pas de simples distributions mais plutôt une distribution de mélanges (*mixture distributions* en anglais ou *mixtura de distribuciones* en espagnol).

Distribution de mélanges

La théorie des distributions de mélanges est complexe et est encore l'objet de nombreuses recherches. On présente par exemple plusieurs types de distribution de mélanges à savoir les mélanges finis et dénombrables ou au contraire les mélanges indénombrables.

Ici premièrement, nous avons restreint notre étude au choix de la famille de distributions bêta. En effet celles-ci étaient principalement les plus rencontrées dans la première démarche effectuée et présentent bien plus de concordance avec les données que les autres (notamment le support de la distribution $[0, 1]$ qui correspond parfaitement). Ainsi nous considérons donc la distribution de mélange suivante : une distribution dégénérée en 0 représentative des possibles plusieurs occurences de ratios nuls, une distribution dégénérée en 1 représentative des ratios égaux à 1 et une distribution bêta sur le reste du support, pour le reste des valeurs. Nous définissons alors la distribution suivante :

$$F(x,\theta) = \begin{cases} P_0, \text{ si } x = 0 \\ (1 - P_0 - P_1)\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \text{ si } x \in]0,1[\\ P_1, \text{ si } x = 1 \end{cases}$$

avec P_0 la probabilité d'obtenir un 0, P_1 la probabilité d'obtenir un 1. Nous pouvons interpréter ces valeurs comme les fréquences d'apparition des valeurs 0 et 1 sur la période de 90 jours considérée. Par exemple, si nous avons $P_0 = \frac{1}{2}$, cela signifie que en moyenne, nous avons la moitié des valeurs qui sont nuls. La variable θ correspond au vecteur des paramètres de notre distribution de mélange : $\theta = (\mu, \sigma, \nu, \tau)$. La fonction B est la fonction Bêta utilisée dans la définition des distributions bêta :

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

et nous pouvons exprimer les valeurs α, β, P_0 et P_1 en fonction de nos paramètres de la distribution de mélanges μ, σ, ν, τ.

$$\alpha = \frac{\mu(1-\sigma^2)}{\sigma^2}$$

$$\beta = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$$

$$P_0 = \frac{\nu}{1+\nu+\tau}$$

$$P_1 = \frac{\tau}{1+\nu+\tau}$$

Nous pouvons appeler cette distribution de mélange "zero-one inflated bêta distribution" à quatre paramètres μ, σ, ν, τ tels que : $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$, $\tau > 0$. Ainsi, inférer ce type de distribution sur nos données signifiera trouver la valeur du vecteur de paramètres θ et nous pourrons donc en déduire les probabilités P_0 et P_1 et la densité de la fonction avec les valeurs de α et β. Voici un exemple des résultats obtenus : j'ai utilisé à cette fin le package *gamlss* de R et divers sous-extensions afin d'obtenir les courbes et les résultats.

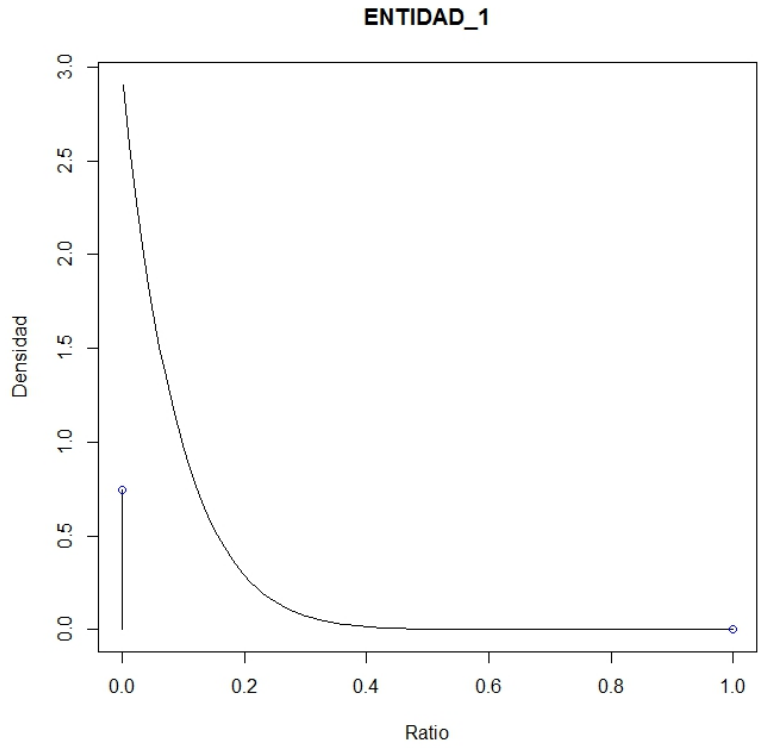


FIGURE 6 – Densité d'une distribution de mélange pour l'entité 1

On aperçoit bien ici la distribution dégénérée en 0 (P_0>0) ainsi que la distribution bêta sur le reste du support. Il n'y a pas de pic en 1, la probabilité P_1 est nulle ou très faible. Cette démarche s'applique aussi au cas où l'on n'a pas de valeurs en 0 ou 1 et donc les distributions dégénérées ne sont pas présentes :

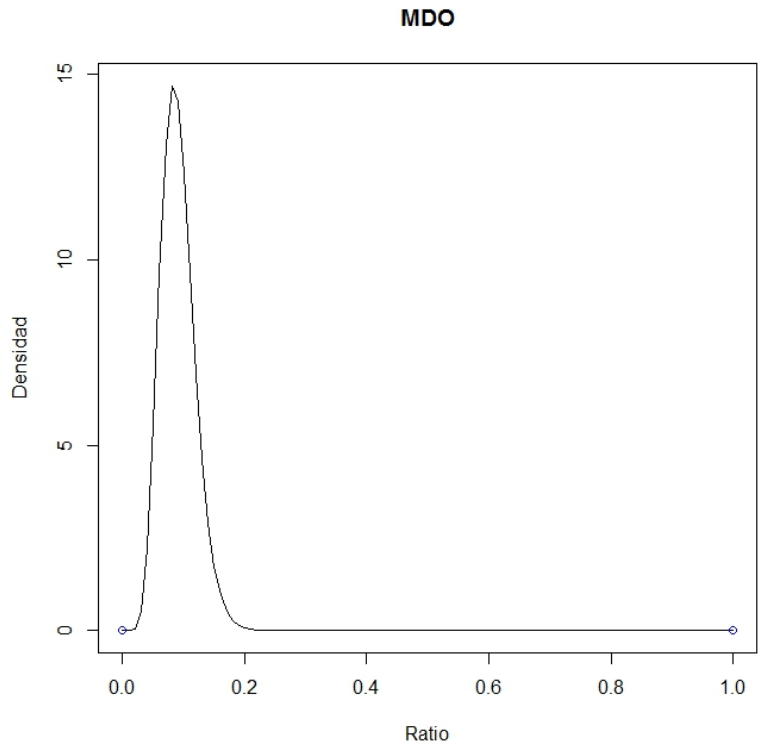


FIGURE 7 – Densité d'une distribution de mélange pour le marché

Fixation du seuil d'alerte

Nous utilisons dès lors la distribution de mélange trouvée pour définir nos seuils d'alerte. Ce type de distribution est très particulière et je n'ai pas trouvé de fonctions permettant de calculer directement les valeurs des quantiles de telles distributions. Il existe bien des fonctions quantiles dans le package *gamlss* mais remplissant des rôles totalement différents et concernant des cas préçis : voir *explanatory, response variable*.

Pour déterminer la valeur du quantile à 98%, j'ai alors procédé de la manière suivante : on peut générer un échantillon de taille N de la distribution de mélange une fois les paramètres θ en possession. L'idée et de générer un échantillon de taille N, de le trier par ordre croissant et de sélectionner alors la $(98\% \times N)^{eme}$ valeur qui correspondra au quantile à 98%. La taille de N sera alors fonction de la précision de la valeur du quantile trouvée. En effet plus N est grand, plus l'échantillon est de taille importante et représentatif de la distribution en question. Cependant, plus N est grand, plus le temps de calcul sera conséquent car le nombre de données traitées est important (beaucoup d'entité, beaucoup de jours pris en compte). Pour la procédure, j'ai choisi un N égal à 10000 qui donne un temps de calcul raisonnable ($<$ 10 min pour la globalité des fichiers

de données, rappelons que l'on veut utiliser la procédure chaque jour) et une précision suffisante. De plus un N dix fois plus grand n'apporte une différence qu'à partir de 10^{-4}. Exemple avec deux N différents sur un fichier de données :

N	Temps de calcul	Seuil
10000	1min 26s	0.0853
100000	3min 53s	0.0859

Suffisamment de données

Bien que nous pouvons maintenant inférer les distributions sur les échantillons, même en prenant en compte les valeurs extrêmes 0 et 1, nous rencontrons certaines fois des échantillons possédant vraiment trop de fois ces valeurs extrêmes. Que faire dans cette situation ? Le problème vient du fait qu'avec vraiment peu de valeurs non nuls, l'inférence d'une distribution de mélanges peut aussi échouer, par exemple si seulement 4 valeurs sont non nulles sur 90 au total. Et même si le programme donne un résultat, que penser d'une inférence d'une distribution bêta faite à partir de seulement 4 valeurs non nulles ? Ce questionnement nous amène vers un large problème : celui de déterminer un nombre minimum de valeurs à avoir pour réaliser une inférence de distributions. Cette question n'a pas de réponse exacte et dépends grandement du travail mené et peut extrêmement varier d'un cas à l'autre. L'important était ici de trouver une solution pertinente dans notre cas. Nous avons donc considéré qu'à partir de 10 valeurs non nulles, l'inférence pouvait être réaliser de manière fiable, en dessous, nous considérions que l'inférence ne ferait pas beaucoup de sens et nous baserions alors notre recherche du seuil différement. Un seuil fixé et prédeterminé dans ce cas serait utilisé alors. En effet, avoir beaucoup de valeurs de ratio nulles reflète l'inefficience dans la liquidation ou bien en fait l'inactivité de l'entité. Et c'est beaucoup plus souvent le cas de l'inactivité. Comme nous l'avons décrit au début, les ratios d'inefficience sont calculés à partir de différente valeurs et obtenir un zéro peut venir de deux façons : ou bien l'entité ne liquide pas tout ses titres et montants en espèces ou bien l'entité n'a tout simplement pas de titres ou montants en espèces à liquider et c'est ce second cas que l'on rencontrait majoritairement. Ainsi dans ce cas, un zéro n'est pas synonyme de problème particulier.

Dès lors nous utilisons comme seuil fixe celui du marché qui représente une référence pour l'ensemble des entités. Qui plus est, le marché est aussi tout le temps actif et est en relation direct avec l'ensemble des entités, d'où la justification d'avoir choisi un tel seuil. Le seuil d'alerte pour une entité possédant alors trop de ratios nuls pour utiliser les méthodes d'inférence sera alors défini comme le seuil d'alerte du marché.

Types d'alerte

Chaque entité a donc un seuil : propre ou celui du marché. L'idée est donc que lorsque la valeur du ratio du jour est supérieur à ce seuil, ceci témoignera d'un problème parce que la valeur serait alors anormalement grande par rapport aux valeurs précédentes de la période considérée. Seulement, ce n'est pas toujours le cas : en effet une entité peut très bien avoir beaucoup de données relativement faibles (< 0.05 par exemple) et par conséquent le seuil de l'entité sera faible également (0.04 par exemple). Ainsi dans ce cas, une valeur d'un ratio supérieure au seuil 0.04, disons 0.06, sera

immédiatement considérée comme problématique alors que ça ne l'est pas vraiment. Cela signifiera simplement que l'efficience de l'entité est moindre par rapport aux 90 derniers jours. Exemple :

Entité	Ratio du jour	Seuil
1	0.311	0.213
2	0.039	0.038

L'entité 1 a un ratio du jour de 0.311 plus grand que son seuil du jour 0.213. Nous souhaitons donc ici une alerte car 0.311 représente presque un tiers du volume (titres et espèces) non liquidé.

Au contraire, l'entité 2 a aussi un ratio du jour 0.039 supérieur à son seuil du jour 0.038 mais nous ne souhaitons pas signaler ceci comme un problème du même ordre que le cas précédent.

Pour ces raisons, nous avons donc établi un moyen de différencier les cas et d'avoir des alertes pertinentes. Nous définissons alors trois alertes d'importance croissante :

1. Le ratio du jour est supérieur au seuil du jour mais pas à celui du marché (le seuil fixé).
2. Le ratio du jour est supérieur au seuil du marché mais pas à celui de l'entité, le même jour.
3. Le ratio du jour est supérieur aux deux seuils du jour (celui de l'entité et celui du marché).

Explications :

L'alerte 1 est générée pour les entités dont les ratios sont faibles et signifie seulement que l'entité a moins bien liquidé, réglé le volume par rapport au 90 derniers jours.

L'alerte 2 se génère lorsque l'entité ne liquide pas au moins aussi bien que le marché mais mieux que ce qu'elle faisait par rapport aux jours précédents. Cela témoigne d'une possible amélioration, d'un possible retour à une liquidation correcte, c'est-à-dire au moins en dessous du seuil du marché.

L'alerte 3 est générée lorsqu'il y un problème important car dans ce cas, l'entité ne liquide pas aussi bien que les 90 derniers jours précédents mais également moins bien que le marché. D'où la priorité de signaler ces cas en premiers.

Exemple :

Entité	Ratio du jour	Seuil	Seuil du marché	Alerte
1	0.116	0.108	0.175	1
2	0.181	0.255	0.175	2
3	0.339	0.268	0.180	3

3.6 Améliorations

Ainsi s'effectue alors la procédure de contrôle des ratios d'efficience sur la liquidation du volume des entités. Des points restent toujours à éclaircir et à affiner, notamment des points de vue théorique et pratique :

- L'inférence final se fait globalement à l'aide du package *gamlss* de R. Si l'on souhaite être pointilleux, on pourrait alors étudier en détail le fonctionnement de chacune des fonctions et effectuer des améliorations à ce niveau là afin d'améliorer en un sens l'inférence.
- Le travail demandé a été testé et effectué sur plusieurs fichiers en conditions réelles afin de vérifier l'efficacité et la pertinence de la démarche choisie. Cependant, il peut toujours y avoir des nouveautés, des comportements imprévus ou qui n'étaient jamais apparus. Il ne faudrait donc pas hésiter à remettre en cause la démarche écrite si l'on s'aperçoit d'un changement majeur et à long terme dans le fonctionnement du marché, des entités et de la liquidation en finance.
- Certains paramètres sont toujours modifiables au gré du fonctionnement du marché, de la définition de l'anormalité et du fonctionnement des fonctions utilisées : nombre de données non nulles minimum pour inférer (ici fixé à 10), probabilité pour définir les seuils (quantile à 98%), taille de la période considérée (90 derniers jours).
- Les procédures d'automatisation des démarches ont leurs limites, surtout concernant une étude statistique qui dépends de nombreux paramètres. Il faut donc tenir en compte ceci. De plus le logiciel R présente des limites de capacités de calculs et d'incorporation des données. Vers la fin du projet, on m'a demandé d'adapter les codes écrits jusqu'alors à des cas encore plus riches, faisant intervenir plus de données, plus de comptes et d'entités. La démarche s'effectuait jusqu'à maintenant ainsi : import des données contenues dans des fichiers Excels, utilisation du code R, résultats en sortie dans un dossier. Cependant avec plus de données, cela se complique et l'on ne peut avoir des matrices ou vecteurs infiniment grands. Le temps de calcul serait également plus long. Ainsi on pourrait améliorer ceci en revoyant la méthode d'importation et de traitement des données dans R (importation avec fichier Access ?), ou bien complètement envisager un autre langage et logiciel ! (plus performant, non gratuit . . .)

4 Bilan

Au travers de ce stage à BME, j'ai pu apprendre de nombreuses choses. J'ai découvert au sein même du lieu les différents secteurs et postes d'activités du monde de la finance, mes connaissances générales en ont été renforcées. J'ai également pu me rendre compte du fonctionnement d'une telle société vis-à-vis des clients, partenaires ainsi que des difficultés rencontrées dans ce milieu. Au delà de l'entreprise elle-même, j'ai pu mettre en pratique une réelle démarche de travail et de recherche, en coopération avec mon maître de stage mais aussi avec les membres de l'équipe du PTI. Il y a eu un réel échange d'avis, d'opinions et de connaissances ce qui nous a permis de progresser correctement et d'apprendre pour chacun de nouvelles choses. J'ai beaucoup apprécié de travailler dans ces conditions à la fois professionnelles et détendus. Un point négatif que je peux néanmoins relever à travers ce stage est la considération de mes capacités. J'ai du me débrouiller globalement seul, étant mené à faire une réelle recherche statistique, je n'ai pas vraiment pu demander d'aide profonde à mes collègues ou bien même à mon responsable qui n'étaient pas suffisament qualifiés à certains moment dans le domaine de l'étude menée. J'étais en quelque sorte responsable des décisions menées dans le domaine mathématique et devait les accorder avec le monde de la finance. Je n'avais pas de validation ou d'appui me permettant d'être 100% certain que je prenais la bonne décision à tel ou tel moment. Mais cela a été un réel défi et me permettra d'aborder de futurs cas similaires avec beaucoup plus de sérénitude. Je suis en fait content d'avoir pu appliquer une démarche scientifique, réfléchie et responsable, ce qui est ce à quoi je suis formé depuis le début de toutes mes études. Enfin, le contexte de travail était encore plus particulier du fait que je me trouvais en Espagne. Je suis finalement très content d'y avoir pu pratiquer encore plus mon espagnol à l'oral comme à l'écrit et me sens désormais plus confiant dans l'exercice.

5 Annexe

5.1 Lexique, traductions, définitions et expressions idiomatiques

APMT : contra pago, correspond au terme *DvP* en anglais pour *Delivery Versus Payment* ou encore *contre paiement* en français.

BME : bolsas y mercados españoles, littéralement bourses et marchés espagnols.

C/V : compra/venta, littéralement achat/vente.

CCP : central counterparty, correspond au terme ECC en espagnol ou chambre de compensation en français.

DvP : delivery versus payment, correspond au terme *contra pago* en espagnol ou encore *contre paiement* en français.

ECC : entidad de contrapartida central, correspond au terme CCP en anglais ou chambre de compensation en français.

FREE : libre de pago, correspond au terme *FoP* en anglais pour *Free Of Payment* ou encore *franco de paiement* en français.

Mercado : le marché.

PTI : post trade interface, le département de contrôle, reçevant et fournissant l'informations au sein de BME.

Renta fija : pour *fixed income* ou *revenu fixe*. Type d'investissement ou contrat pour tout type d'actif financiers sous lequel l'émetteur est obligé d'effectuer les paiements en quantité et en temps préalablement déterminés. Par exemple si l'on acquiert un instrument financier à revenu fixe, on doit connaître les intérêts et les rendements dès lors que l'on possède le dit-instrument.

Renta variable : pour *variable income* ou *revenu variable*. Type d'investissement ou contrat pour tout type d'actif financiers sous lequel n'est pas garanti le retour du capital investi ni la rentabilité de l'actif concerné. Généralement, le mot fait référence aux actions (mercado de renta fija = marché des actions) mais ce ne sont pas les seuls ! Il y a aussi les fonds d'investissement ou encore les obligations convertibles par exemple.

Valores : les titres financiers.

Bibliographie

[1] Christophe Chesneau. *Sur l'adéquation à une loi de probabilté avec R*, Décembre 2016.

[2] J.W Davenport, J.C Dezbek, and R.J Hathaway. *Parameter estimation for finite mixture distributions*, 1988.

[3] Marie Laure Delignette-Muller and Christophe Dutang. *fitdistrplus : An R package for Fitting Distributions*, October 2014.

[4] Bill Huber. *Fitting distributions to data*, March 1999. Practical issues in the use of probabilistic risk assessment.

[5] Nornadiah Mohd Razali and Yap Bee Wah. *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests*, 2011.

[6] Ospina Raydonal. *A brief introduction to GAMLSS modeling.*

[7] Vito Ricci. *Fitting distributions with R*, February 2005.

[8] Bob Rigby and Mikis Stasinopoulos. *A flexible regression approach using GAMLSS in R*, June 2008.

[9] Mikis Stasinopoulos, Marco Enea, Robert A. Rigby, and Abu Hossain. *Inflated distributions on the interval [0,1]*, April 2017.