Risque de Données et Backtesting de la Value-at-Risk

Sylvain Benoit*

sous la direction de Christophe Hurlin

Université d'Orléans

Septembre 2010

Résumé

Les autorités de réglementation bancaire imposent aux banques de détenir un niveau minimum de fonds propres pour couvrir leur risque de marché. Ce montant est calculé par les banques elles-mêmes à partir de leur modèle interne d'évaluation du risque. Il est alors primordial pour les régulateurs et les banques de s'assurer de la validé de ces modèles. L'ensemble de ces techniques de validation est appelé le *backtesting*. Elles permettent de vérifier que les pertes observées *ex-post* sont en adéquation avec celles prévues. Or, ces modèles reposent sur de nombreuses hypothèses et sont soumis à différents types de risques : de modèle, d'estimation et surtout de données. Si jusqu'à présent, très peu d'études économétriques ont été consacrées à ce dernier type, cet article s'attache justement à mesurer l'impact du risque de données sur les procédures de *backtesting*, en observant les déformations de taille et de puissance de ces tests suite à une contamination.

Keywords : Profit-and-Loss, Risk Management, Backtesting, Value-at-Risk.

JEL classification : C58, C52, G28.

*Etudiant en deuxième année du Master ESA (Econométrie et Statistique Appliquée) à l'université d'Orléans, sylvain.benoit@univ-orleans.fr. Je tiens à exprimer de sincères remerciements à mon directeur de mémoire, Christophe Hurlin (Université d'Orléans), qui est à l'initiative de ce thème de recherche particulièrement intéressant. Ses conseils, sa patience et son appui ont contribué à la bonne marche de ce mémoire de master. Une pensée toute particulière est adressée à l'ensemble de l'équipe du Master ESA pour la qualité de leurs enseignements, m'offrant ainsi la possibilité de réaliser ce projet. Enfin, je remercie chaleureusement ma famille ainsi que mes amis pour leurs encouragements sans faille.

1 Introduction

Depuis plus de vingt ans, le Comité de Bâle impose des normes prudentielles dont l'objectif est d'assurer la pérennité du système bancaire. La réglementation (accords de Bâle I puis de Bâle II) impose alors aux banques ayant choisi de développer leur propre modèle interne, de calculer des prévisions de la *Value-at-Risk*[1] (VaR) afin de les comparer aux *Profit-and-Loss* (P&L) journaliers qu'elles observent. En théorie avec cette démarche, le régulateur espère que le montant du capital réglementaire associé au risque de marché soit le plus proche possible du véritable niveau de risque auquel font face les banques (Hirtle, 2003). Cela devrait ainsi limiter les violations de la VaR, c'est-à-dire les situations *out-of-sample* où la perte observée est supérieure au montant prédit de la VaR. Pour les banques, construire leur propre modèle interne d'évaluation des risques et de calcul de la VaR leur permet d'assurer la confidentialité de leur modèle. De plus, cette pratique entraine une non uniformité des approches de mesures de risques développées dans le secteur bancaire. Le principal enjeu pour le régulateur est alors d'évaluer la qualité de ces modèles internes. De nombreuses techniques statistiques de validation, nommé *backtesting*, permettent ainsi de vérifier l'adéquation des prévisions de la VaR. Ce contrôle est indispensable pour le régulateur puisqu'il a délégué aux banques le soin d'évaluer le niveau de risque qu'elles prennent. En fonction des résultats du *backtesting*, le régulateur peut classer les banques en trois zones distinctes (verte, jaune et rouge) et appliquer un facteur d'ajustement afin de traduire la valeur estimée de la VaR en un montant minimum de capital réglementaire permettant de couvrir le risque de marché. La capacité des procédures de *backtesting* à détecter une sur ou sous exposition au risque de marché d'une institution financière est donc primordial pour le régulateur afin de remplir sa mission, mais aussi pour le *risk manager* dans l'optique de contrôler en interne les risques pris par sa banque.

La multitude de modèles internes développés se justifie par l'existence de nombreuses méthodes de calcul de la VaR. Qu'elles soient paramétriques, historiques, ou non paramétriques, ces dernières délivrent des estimations de la VaR trop imprécises. Les montants estimés ne convergeant pas, le risque peut donc être parfois sur ou sous évalué. Ces différences résultent du fait que ces techniques de calcul reposent sur des hypothèses très diverses et critiquables. En effet, l'approche paramétrique suppose que la volatilité conditionnelle suit un modèle GARCH, que les innovations sont identiquement et indépendamment distribuées suivant une certaine loi, bien souvent la loi normale. Or ces suppositions peuvent

[1] La VaR est une estimation de la perte maximale qui ne devrait pas être dépassée pour un portefeuille donné, un horizon temporel donné, avec un niveau de confiance donné et sous des conditions normales de marché.

être fausses, tout comme l'ordre du modèle GARCH sélectionné. Un risque de spécification du modèle émerge. Afin d'éviter cet inconvénient, une approche historique peut être mise en place. Mais les prédictions de la VaR basées seulement sur une étude de l'historique récent ne répercute que trop tardivement les effets d'un changement de volatilité sur les marchés ou d'une rupture non anticipée de régime. Ce non ajustement entraine généralement un *cluster* de violations. Une autre alternative consiste alors à laisser parler les données de *Profit-and-Loss* (P&L). Une estimation non paramétrique de la fonction de densité des P&L est effectuée. Cette approche permet de ne pas spécifier de modèle, le risque de non ajustement se trouve proscrit puisque nous avons estimé la véritable fonction de distribution des rendements. Mais comment être certain de la qualité de cette estimation ? L'algorithme d'optimisation nécessaire aurait très bien pu s'arrêter sur un minimum local. Par ailleurs, les estimations des paramètres diffèrent selon le choix du *kernel* et surtout de la fenêtre de lissage. En conséquence, les modèles internes développés par les banques sont soumis à de nombreux risques, non évidents à première vue. Il est donc essentiel que les procédures de *backtesting* détectent ce type de mauvaise spécification de la VaR. Mais ces procédures sont elles aussi, soumises à des risques. Escanciano et Olmo (2008) ont montré que ces méthodes sont sujettes au risque de modèle engendré par une mauvaise spécification de la VaR mais aussi à un risque d'estimation. Le risque de modèle correspond à l'effet qu'entraine une mauvaise spécification de la distribution asymptotique des tests de *backtesting*. Ils montrent que les tests de couverture non conditionnelle, comme celui de Kupiec, sont affectés directement par ce risque de modèle tandis que les tests d'indépendance et de couverture conditionnelle ne le sont qu'indirectement au travers du risque d'estimation. Cela signifie que si la banque est en possession du bon modèle, c'est-à-dire celui qui est à l'origine des rendements qu'elles observent sur son portefeuille, mais qu'elle est contrainte d'estimer les paramètres de ce modèle et en particulier sur de petits échantillons, alors la banque est soumise au risque d'une mauvaise estimation des paramètres de son bon modèle et *in fine* a une mauvaises mesure de la VaR de son portefeuille. Mais plus la taille d'échantillon croit et plus le risque d'estimation se réduit. Toutefois, ces deux types de risques sur les procédures de backtesting ne semblent pas jouer un rôle majeur. En effet, les amendements des accords de Bâle, publiés par la *Bank for International Settlements* (BIS) en 1996, ont clairement identifé le risque de données comme le principal risque auxquels les banques sont confrontées.

Anticipant parfaitement le risque inévitable de contamination des données de P&L, la réglementation a édicté des règles claires à suivre concernant les données bancaires à utiliser pour valider les modèles internes. Ainsi, la banque doit porter une importance toute particulière à trois sources de contamination : les frais de gestion et les commissions, les revenus générés par l'*intraday trading* et les données historiques. Hendricks et Hirtle (1997)

ont évoqué ce problème en signalant que cette pratique de contamination pourrait, pour des raisons techniques, conduire à altérer les résultats du *backtesting* en sur ou sous rejetant les modèles testés. Mais, ils insistent sur le fait que la direction du décalage n'est pas claire. D'une part, l'inclusion dans les données de P&L des frais de gestion et des commissions, d'un montant toujours positif, tend à réduire le nombre d'exceptions puisque l'on décale vers la droite la distribution des P&L. D'autre part, l'impact des données d'*intraday trading* est lui moins évident. Ces revenus tantôt positifs, tantôt négatifs, rendent le sens du décalage de la distribution moins évident. Toujours est-il que ces montants augmentent vraisemblablement la volatilité des données journalières de P&L, entrainant ainsi la hausse de la probabilité d'une exception. En plus de la contamination causée par les frais de gestion et les commissions, et les revenus de l'*intraday trading*, cet article présente les données historiques comme une nouvelle source de pollution des P&L. Les procédures de *backtesting* doivent en effet utiliser des données rétrospectives pour construire les séquences de violations qu'elles étudient[2]. Etant donné que l'impact de ces données contaminées n'est pas clair, les banques ont la permission de les utiliser pour tester leur modèle, et peuvent, sur la base du volontariat déclarer dans leur rapport annuel quel type de données elles utilisent.

Pourtant en 2010, l'article de Frésard, Pérignon et Wilhelmson est le premier à mettre en évidence le fait que la plupart des banques de leur échantillon utilise des données contaminées pour tester la validité de leurs modèles. L'utilisation de données contaminées entraine un effet uniquement positif pour les banques en améliorant les performances de leurs modèles internes. Leur démarche s'appuie sur des informations spécifiques au *risk management* collectées à partir des rapports annuels des deux cents plus grosses banques commerciales[3] américaines et internationales, sur la période 2005-2008. Alors qu'elle devrait être proscrite, cette pratique est largement répandue à travers le secteur bancaire mondial, plus chez les banques américaines qu'européennes. Ils soulignent que la plupart des banques augmentent le taux de validation de leur modèle en utilisant ce type de données contaminées. Leur étude révèle que ces données polluées ont des effets majeurs sur les résultats du *backtesting*, entrainant une sur-validation des modèles internes d'évaluation du risque. Les montants de capitaux réglementaires ainsi calculés se trouvent réduits de 17% en moyenne. Ces auteurs montrent un renforcement de ce phénomène avec la dernière crise financière de 2008 et signalent enfin, à travers diverses simulations de Monte-Carlo, que les méthodes de *backtesting* employées sont très sensibles à la pollution des données.

Mais comment évaluer la déformation des procédures de *backtesting* ? Ancré précisément

[2]Les rendements de P&L à comparer aux prévisions *ex-ante* de la VaR doivent être calculés en utilisant les pondérations des actifs de la veille (données rétrospectives) et non du jour (données historiques).

[3]Classement obtenu à partir du montant de l'actif total (exprimé en dollars américains) de chaque banque à la fin de l'année fiscale de 2006.

autour de cet enjeu, notre article se donne pour objectif d'évaluer l'impact de l'utilisation des P&L contaminées sur ces procédures. Etudier cette sensibilité oblige à observer les déformations de taille et de puissance entrainées par ces données polluées sur quelques tests usuels de validation de la VaR. Cette démarche permet ainsi de détecter quels tests sont les plus sensibles, et à quels phénomènes. Dans le but de s'affranchir de tout risque d'hypothèses, de spécification et d'estimation, trois études de Monte-Carlo sont menées, soit une pour chaque type de contamination, et ce, dans l'optique d'en observer leurs impacts respectifs. Concrètement, nous simulons sous l'hypothèse nulle un processus générant des rendements non pollués sur lesquels nous calculons les prévisions adéquates de la VaR. Puis, nous appliquons les tests de *backtesting* à ces données non polluées. Ensuite, nous polluons volontairement nos rendements et les comparons aux prévisions de la VaR précédentes. Ainsi, les déformations de tailles empiriques des tests de *backtesting* faisant suite à cette pollution sont mises en valeur. La démarche est similaire pour étudier les différences de puissance, à une exception près : puisque nous simulons sous l'hypothèse alternative, les prévisions de la VaR réalisées sur les rendements non pollués sont inexactes. Nous choisissons en effet un mauvais modèle de calcul de la VaR car nous voulons observer la fréquence de rejet de l'hypothèse nulle, que ce soit avec ou sans pollution des rendements. Dès lors, nous comparons les différences de taille et de puissance observées, et détectons quantitativement quel type de pollution entraîne le plus de déformations sur les résultats du *backtesting*.

Selon les conditions imposées par le régulateur, il s'avère que le sens de la déformation des procédures ne fait plus aucun doute puisque deux des trois types de pollution tendent à diminuer les taux de rejet de l'hypothèse nulle. Cela signifie que des modèles sont validés alors qu'ils ne devraient pas l'être. En revanche, les déformations de taille obtenue sur grand échantillon varient d'un type de contamination à l'autre. De plus, il faut distinguer les contaminations qui sont exclusivement positives de celles qui sont soit positives, soit négatives. Ainsi, l'impact des données polluées par les frais de gestion et les commissions est beaucoup plus important que celui provoqué par les rendements historiques. Enfin, il faut distinguer les techniques de validation de la VaR en fonction de leurs caractéristiques puisque les résultats observés diffèrent assez sensiblement d'une procédure à l'autre. Même si elles sont toutes affectées pour une taille d'échantillon de 250 et un niveau de risque de 1%, les tests de couverture conditionnelle basés directement sur l'occurence des violations sont les plus déformés par le risque de données.

Le plan de notre article est le suivant. La section 2 revient sur la définition des trois différentes sources de pollution des données de P&L. La section 3 développe les techniques de *backtesting* employées dans l'article afin de mettre en évidence leurs déformations propre à l'emploi de chaque type de données contaminées. La section 4 présente trois expérimentations

de Monte-Carlo réalisées et discute des principaux résultats concernant les déformations de taille et de puissance empiriques observées. Enfin, la dernière section résume notre étude et conclut sur la nécessité de n'utiliser que des données non contaminées.

2 Les sources de contamination des pertes et profits

Bien que l'emploi de ce type de données ne soit pas recommandé par la réglementation (*Basel Committee on Banking Supervision*, BIS, Janvier 1996), plusieurs phénomènes poussent tout de même les institutions financières à les utiliser. Parmi ceux-ci, la concurrence bancaire internationale joue un rôle moteur. Dans cette optique, l'objectif de la banque est de limiter au maximum l'appréciation de ses fonds propres réglementaires, même en période de fortes turbulences, afin de ne pas être trop pénalisée. Ceci pourrait expliquer en partie pourquoi ce phénomène a été plus présent pendant la dernière crise financière. Par ailleurs, si une banque utilise des données polluées, elle possède, toutes choses égales par ailleurs, un avantage sur ses concurrents puisqu'elle bénéficierait d'un effet de levier plus grand en raison de son plus faible montant de capital réglementaire. Aux yeux des investisseurs cette banque deviendrait alors plus attractive. Or, dans le contexte actuel de globalisation financière, il est fort probable que les autres banques aient un comportement mimétique afin de ne pas souffrir de ce désavantage.

Enfin, cette pratique pourrait être la conséquence d'une mauvaise séparation entre le *risk management* et le *front office*. Si la VaR est utilisée pour contrôler ou rémunérer la prise de risque des traders, alors certains d'entre eux pourraient sous-estimer son montant en appliquant une technique non adéquate. En fournissant ensuite au risque manager des P&L contaminées, ils seraient quasi certains de valider ce mauvais modèle d'évaluation du risque de marché. De ce fait, les revenus réalisés par ces traders seraient associés à un niveau de risque plus faible, d'où une meilleure valorisation de son travail. Il existe donc de multiples motivations conduisant à l'utilisation de données contaminées.

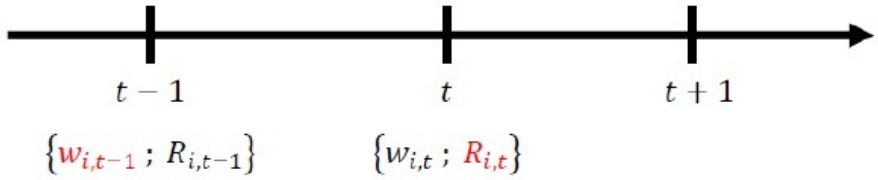
La suite de cette section définit chaque source de pollution pour en faire ressortir leurs spécificités.

2.1 L'utilisation des rendements historiques

Afin de valider l'approche mise en place par le *risk management*, la VaR *ex-ante* doit être comparée aux rendements *ex-post* du portefeuille, obtenu en considérant une logique de poids constants. Il faut donc vérifier que le poids des actifs dans le portefeuille ne varie pas au cours de la journée. Le premier élément de contamination repose sur cette dimension statique

de la VaR. En effet, la VaR ne tient compte que des modifications possibles de la valeur du portefeuille compte tenu des fluctuations des cours des actifs et/ou des possibles changements de taux au cours de la journée. L'institution financière calcule ainsi ses prévisions *ex-ante* de la VaR ($VaR_{t+1/t}$) sachant les positions qu'elle détient à la date t (w_t). Or, ces positions peuvent changer d'une journée à l'autre. Ainsi, la bonne pratique consisterait à comparer les données de P&L rétrospectives ($P\&L_t^{CH}$) aux prévisions de la VaR, au lieu de prendre les P&L historiques ($P\&L_t^{H}$) qui incorporent les modifications du portefeuille de la journée. Les P&L rétrospectifs mesurent le changement de valeur du portefeuille en tenant compte uniquement de l'évolution des prix, les pondérations de chaque actifs étant ceux de la date $t-1$. Ce sont par définition les P&L sans contamination.

Soit $w_{i,t}$ et $R_{i,t}$, respectivement le poids et le rendement de l'actif i à la date t. Le rendement du portefeuille rétrospectif à la date t est donc égal à $P\&L_t^{CH} : R_{p,t} = w_{i,t-1}.R_{i,t}$, tandis que le rendement du portefeuille historique est égal à $P\&L_t^{H} : \tilde{R}_{p,t} = w_{i,t}.R_{i,t}$. La figure ci-dessous permet de visualiser les poids et les rendements à utiliser :



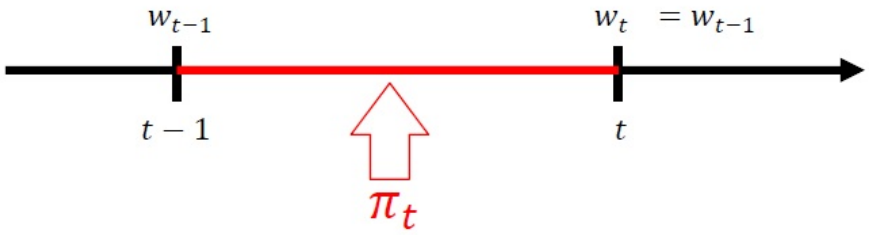
Les rendements rétrospectifs calculés à la date t sont ensuite comparés à la prévision de la VaR faite en $t-1$, qui est égale en théorie à l'opposé du percentile d'ordre α de la distribution des $P\&L$. Les tests de *backtesting* doivent donc être menés à partir de ces données. Or, il est fort probable que certaines banques utilisent les données historiques pour les comparer aux prévisions de la VaR. Cette pratique est valide uniquement lorsque le poids de chaque actif reste constant dans le temps, sinon nous allons à l'encontre de la définition de la VaR.

2.2 L'intraday trading

La seconde source de pollution provient des revenus *intraday*. Les portefeuilles détenus par les banques ne sont pas statiques mais dynamiques. En effet, les traders font varier à chaque instant les positions qu'ils détiennent en fonction des stratégies d'investissement appliquées. Concrètement, ils achètent ou vendent les actifs de leur choix compte tenu de leurs anticipations, faisant ainsi fluctuer les pondérations de chaque actif de leur portefeuille. Par exemple, considérons un portefeuille constitué de 100 actifs. On impose au trader de détenir le même nombre d'actions au début et à la fin de la journée. Malgré cette restriction, et si l'on suppose qu'il n'est soumis à aucune contrainte de liquidité, le trader conserve la

possibilité de spéculer à la hausse ou à la baisse sur l'évolution des prix de l'actif en achetant ou vendant ce titre.

Ces revenus, notés π_t, se définissent donc comme la somme des gains et des pertes générés au cours d'une journée par l'*intraday trading*. Les P&L *intraday* ($P\&L_t^I$) mesurent alors le changement de valeur du portefeuille en incorporant aux P&L du portefeuille, les revenus générés par l'*intraday trading*, c'est-à-dire des arbitrages et techniques de trading que les opérateurs appliquent quotidiennnement à leur portefeuille. Ces rendements pollués sont égaux à $P\&L_t^I = P\&L_t^{CH} + \pi_t$. Ces recettes ou pertes totales de la journée ne sont pas liées aux mouvements de prix sur lequel est calculé la VaR puisque l'horizon de temps est différent. Or, la VaR tient compte du mouvement global quotidien et non de chaque mouvement haussier ou baissier qui a lieu tout au long de celle-ci. La figure suivante illustre cette source de pollution :



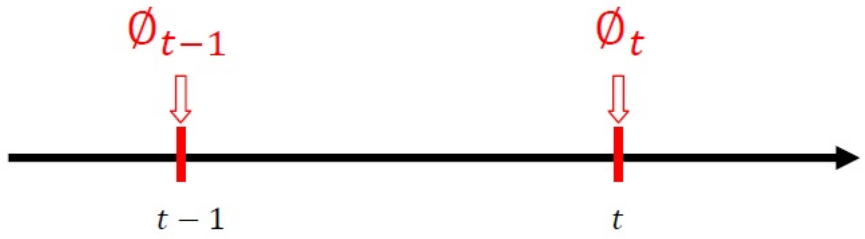
Soit w_t le poids de l'actif composant le portefeuille à la date t. Les pondérations des actifs dans le portefeuille sont identiques d'une période à l'autre, les rendements rétrospectifs et historiques sont donc identiques. Le montant de pollution *intraday* est obtenu au cours de la journée t puis incorporer aux P&L à la fin de chaque date t, et peut-être négatif, positif ou nul.

2.3 Les frais de gestion et les commissions

Chaque jour les institutions financières facturent à leurs clients un grand nombre de frais, ces derniers constituent la troisième source de pollution. Citons par exemple, les frais d'investissement et/ou les frais de courtage, appelés les commissions. Ils sont prélevés par la banque dès qu'un client achète une part d'un fond ou dès que ce dernier souhaite effectuer une opération sur les marchés boursiers. En général, les frais à l'entrée sont fixes tandis que les frais de courtage (facturés par le *broker* lors de la passation de l'ordre boursier) peuvent être proportionnels au montant de la transaction effectuée. En plus de ces commissions, viennent s'ajouter les frais de gestion (indirects ou directs) que font payer les banques à leurs clients en rémunération de la gestion d'OPCVM ou d'un compte titres. Globalement,

ces frais et commissions représentent un montant non négligeable de la valeur totale du portefeuille détenu par la banque. Ils constituent donc une source importante de pollution des données.

Ces frais de gestion et commissions peuvent également être perçus comme un montant proportionnel au volume d'activité ou à la liquidité de la journée. En effet, plus les volumes échangés au cours d'une journée sont importants, plus le montant de ces frais sera élevé. Suivant les banques, ceux-ci se retrouvent ou non dans les données de P&L de la journée. Ces revenus, notés ϕ_t, sont strictement positifs et leur montant varie d'une banque à l'autre. Ils décalent donc la distribution des P&L vers la droite. Les P&L frais de gestion et commissions ($P\&L_t^{F\&C}$) mesurent le changement de valeur du portefeuille en incorporant ces frais perçus, soit $P\&L_t^{F\&C} = P\&L_t^{CH} + \phi_t$. Ces frais sont contemporains aux revenus provenant de l'*intraday trading* et sont également incorporés aux P&L à chaque date t, comme l'indique la figure ci-dessous. En revanche, le montant de ces frais varie d'une date à l'autre mais leur valeur est toujours strictement positive.



3 Les techniques de validation de la Value-at-Risk

La logique de Bâle II est de laisser les banques construire et leur propre modèle interne de risque pour évaluer leur risque de marché, au travers le calcul de la VaR. Suivant cette logique de délégation, ces VaR sont ensuite validées par le régulateur en utilisant des procédures de backtesting.

Etant donné que la véritable valeur de la VaR n'est pas observable *ex-post*, les techniques usuelles de validation des prévisions (*Mean Square Error, Mean Absolute Percentage Error...*) ne peuvent pas être utilisées. Par ailleurs, il n'existe pas de proxy de la VaR auquel nous pouvons comparer nos prévisions. Dès lors, des procédures de *backtesting* ont été développées, cherchant toutes à tester l'hypothèse nulle selon laquelle la séquence de violations des prévisions de la VaR est valide. Alors, le modèle interne d'évaluation du risque de marché est adéquat. Les procédures de *backtesting* sont séparées en deux catégories, la *density forecast evaluation* et celle de l'*event probability forecast evaluation*. Les tests appliqués

dans cet article appartiennent à la seconde catégorie et sont séparés en trois grandes approches. La première basée sur l'étude de la fréquence des violations, la seconde sur l'analyse des durées entre violations, et la troisième sur une approche multidimensionnelle.

Il existe un grand nombre de techniques de prévisions de la VaR. C'est pourquoi les procédures de validation sont généralement de type *model free* afin qu'elles puissent être appliquées à n'importe quel modèle interne. Les prévisions de la VaR aux rendements observés du portefeuille en construisant la séquence de violations suivante :

$$I_t(\alpha) = \begin{cases} 1 & \text{si } r_t < -VaR_{t|t-1}(\alpha) \\ 0 & \text{sinon} \end{cases} . \tag{1}$$

Cette séquence doit respecter deux conditions afin d'être validée :

1. L'hypothèse de couverture non conditionnelle : chaque jour la probabilité d'avoir une violation doit être exactement égale au taux de couverture α :

$$\Pr\left[I_t(\alpha) = 1\right] = E\left[I_t(\alpha)\right] = \alpha. \tag{2}$$

2. L'hypothèse d'indépendance : les violations de la séquence issues d'un même taux de couverture doivent être indépendamment distribuées. Ce n'est pas parce que l'on observe une violation aujourd'hui, qu'il y en aura une demain. Tout cluster de violations est ainsi évité.

Les techniques de validation employées par la suite s'attachent à vérifier ces deux conditions simultanément en construisant un test joint afin de tester cette hypothèse de couverture conditionnelle. Seule la statistique de Pérignon et Smith (2008) teste l'hypothèse de couverture non conditionnelle. Il existe de nombreuses techniques de validation très diverses, c'est pourquoi nous avons sélectionné un large panel de statistiques de tests afin de comparer les déformations de taille et de puissance au sein de chaque grande segmentation de ces techniques.

3.1 Tests fondés sur la fréquence des violations

Les tests de *backtesting* détaillés dans la suite de cette sous section, ne s'intéressent qu'aux occurences de violations. La seule information utilisée est donc la survenue ou non d'une violation. Le test LR_{CC} de Christoffersen (1998) propose ainsi de modéliser le processus $I_t(\alpha)$ de violations par une chaine de Markov, ayant comme matrice de probabilités de transition la matrice suivante :

$$\Pi = \begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}, \tag{3}$$

où $\pi_{ij} = \Pr\left[I_t(\alpha) = j \mid I_{t-1}(\alpha) = i\right]$. Cette chaîne de Markov permet de modéliser une éventuelle dépendance temporelle dans la séquence $I_t(\alpha)$ mais uniquement à l'ordre 1. On cherche ainsi à savoir si la probabilité d'observer une violation à la date t est bien indépendante de l'état en $t-1$ en vérifiant que ces probabilités sont bien égales à α. L'hypothèse nulle est donc exprimée par la matrice suivante :

$$H_0 : \Pi = \Pi_\alpha = \begin{pmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{pmatrix}. \tag{4}$$

Christoffersen propose une statistique de type *Likelihood Ratio* (LR) :

$$\begin{aligned} LR_{CC} &= -2\{\ln L\left[\Pi_\alpha, I_1(\alpha), ..., I_T(\alpha)\right] \\ &\quad - \ln L\left[\hat{\Pi}, I_1(\alpha), ..., I_T(\alpha)\right]\} \xrightarrow[T\to\infty]{L} \chi^2_{(2)}, \end{aligned} \tag{5}$$

où $\hat{\Pi}$ est l'estimateur du maximum de vraisemblance de la matrice de transition, sous l'hypothèse alternative de dépendance, et où $\ln L\left[\Pi, I_1(\alpha), ..., I_T(\alpha)\right]$ désigne la log-vraisemblance des violations $I_t(\alpha)$ associées à une matrice de transition Π. Cette statistique permettant de tester l'hypothèse nulle de couverture conditionnelle souffre de deux limites majeures : d'une part il ne teste que la dépendance temporelle au premier ordre et d'autre part la chaîne de Markov ne permet pas de prendre en compte le rôle d'autres variables. En effet, seule la série des violations est utilisée dans le but de détecter une possible dépendance des exceptions entre elles.

Le test d'Engle et Manganelli (2004) corrige ces deux aspects. Leur test repose sur le processus de violation $I_t(\alpha)$ mais centré sur α, tel que $Hit_t(\alpha) = I_t(\alpha) - \alpha$. Dès lors, ils définissent un modèle de régression linéaire liant les violations centrées courantes $Hit_t(\alpha)$ aux violations centrées passées $Hit_{t-k}(\alpha)$. Soit le modèle de régression linéaire suivant :

$$\begin{aligned} Hit_t(\alpha) &= \delta + \sum_{k=1}^{K} \beta_k Hit_{t-k}(\alpha) + \sum_{k=1}^{K} \gamma_k \\ &\quad g\left(Hit_{t-k}(\alpha), Hit_{t-k-1}(\alpha), ..., z_{t-k}, z_{t-k-1}, ...\right) + \varepsilon_t, \end{aligned} \tag{6}$$

où les résidus ε_t satisfont :

$$\varepsilon_t = \begin{cases} 1-\alpha & \text{avec une probabilité } \alpha \\ -\alpha & \text{avec une probabilité } 1-\alpha \end{cases}, \tag{7}$$

et où $g(...)$ désigne une fonction des violations passées et de variables z_{t-k}, comme les valeurs passées de la VaR. L'hypothèse nulle est donc définie par :

$$H_0 : \delta = \beta_k = \gamma_k = 0, \qquad \forall k = 1, ..., K. \tag{8}$$

Sous l'hypothèse alternative, la séquence des prévisions de la VaR n'est pas correcte puisque les régresseurs sont corrélés avec la variable dépendante. La statistique de test construite par Engle et Manganelli est une statistique de Wald reposant sur la normalité asymptotique de l'estimateur des MCO. Si l'on note $\Psi = (\delta \ \beta_1 \ ...\beta_k \ \gamma_1 \ ...\gamma_K)'$ le vecteur des paramètres $2K+1$ de ce modèle et Z la matrice des variables explicatives du modèle, alors sous l'hypothèse nulle, la statistique de test vérifie :

$$DQ_{CC} = \frac{\widehat{\Psi}' Z' Z \widehat{\Psi}}{\alpha(1-\alpha)} \xrightarrow[T\to\infty]{L} \chi^2_{(2K+1)}. \tag{9}$$

3.2 Tests fondés sur les durées

Dans cette sous section, les tests employés prennent en compte la durée entre deux violations successives. Afin de palier la faible puissance des tests de validation notamment sur de petites tailles d'échantillon, Christoffersen et Pelletier (2004) ont développé un test fondé sur cette approche. L'idée étant que la durée entre deux violations successives ne devrait pas avoir de mémoire, autrement dit, de récurrence systématique. Ainsi la moyenne entre deux violations devrait être égale à $1/\alpha$. Soit $D_i = t_i - t_{i-1}$ la durée entre la $i^{ème}$ violation et la précédente, pour $i = 1, \ldots, T$. Ces durées sont associées à une variable de censure, valant 1 si la durée est censurée et 0 sinon. Sous l'hypothèse nulle d'indépendance des violations, la durée suit une loi exponentielle de paramètre α et de densité :

$$f(d;\alpha) = \alpha \exp(-\alpha d) \qquad \forall d \in \mathbb{N}^*. \tag{10}$$

Sous l'hypothèse alternative, ils supposent que la durée suit une loi de Weibull de paramètre d'échelle b, de paramètre de centrage a, et de densité égale à :

$$g(d;b;a) = a^b b d^{b-1} \exp\left[-(ad)^b\right]. \tag{11}$$

Cette hypothèse permet de proposer un test simple de couverture conditionnelle tel que :

$$H_0 : b = 1 \qquad a = \alpha. \tag{12}$$

Le test associé à cette hypothèse est une statistique de test de type ratio de vraisemblance, telle que :

$$LR_{CC}^{durée} = 2\left(ll_{nc} - ll_c\right) \xrightarrow[T\to\infty]{L} \chi^2_{(2)}, \tag{13}$$

où ll_c et ll_{nc} désignent respectivement la log-vraisemblance de la loi exponentielle (modèle contraint) et la log-vraisemblance de la loi de Weibull (modèle non contraint). Toutefois, Haas (2005) a mis en évidence qu'une approximation continue par une loi géométrique des durées entre deux violations pouvait avoir de lourdes conséquences sur les propriétés à distance finie de ces tests de validation.

L'approche *Generalized Methods of Moments* (GMM) développée par Candelon, Colletaz, Hurlin et Tokpavi (2009) tient compte de cette remarque en utilisant une distribution discrète pour modéliser les durées entre deux violations successives. En plus de cette nouveauté, ils développent un test plus puissant, basé sur la méthodologie détaillée auparavant par les travaux de Bontemps (2006), et Bontemps et Meddahi (2004, 2005). Ces articles se sont intéressés aux distributions discrètes de la famille de Ord comme la distribution de Poisson, Binomial, Pascal et Hypergéométrique, auxquelles ils ont associés des polynômes orthonormaux. Ainsi, ils utilisent ces polynômes orthonormaux associés à la distribution géométrique comme des conditions de moments pour définir les conditions d'orthogonalité de leur test GMM. Sous l'hypothèse nulle de couverture conditionnelle, les durées, produites à partir de la séquence des violations, sont identiquement et indépendamment distribuées et suivent une distribution géométrique avec une probabilité de succès égale au taux de couverture α. On a alors l'hypothèse nulle suivante :

$$H_{0,CC} : E\left[M_j(d_i;\alpha)\right] = 0 \qquad j = \{1, ..., p\}, \tag{14}$$

où p est le nombre de conditions de moments utilisés, et $M_j(d_i;\alpha)$ représente les polynômes orthonormaux associés à la distribution géométrique ayant une probabilité de succès égale à α pour toutes durées entières strictement positives. Etant donné que les polynômes orthonormaux ont une matrice asymptotique de variance covariance connue, nous pouvons écrire facilement la matrice des poids optimaux de la J_{stat} puisqu'elle est égale à la matrice identité. Sous l'hypothèse nulle, la statistique de test associée à p polynômes orthonormaux est définie telle que :

$$J_{CC}(p) = \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(d_i;\alpha)\right)' \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(d_i;\alpha)\right) \xrightarrow[N\to\infty]{d} \chi^2_{(p)}, \tag{15}$$

avec $M(d_i;\alpha)$ un vecteur de dimension (p,1) dont les composants sont les polynômes orthonormaux $M_j(d_i;\alpha)$ pour $j = 1,\ldots,p$ et α est le taux de couverture.

3.3 Tests multinomiaux

La principale limite aux tests précédents basés sur les durées est qu'ils supposent la réalisation d'au moins une violation dans la série pour le test de Christofersen et Pelletier (2004), voire même de deux violations minimum pour le test GMM puisqu'il utilise uniquement des durées non censurées. Or, pour de petites tailles d'échantillon et un taux de couverture faible, comme c'est le cas pour la norme imposée par le régulateur (prévisions de la VaR sur 250 jours au seuil de risque de 1%), il est alors probable que l'on obtienne aucune violation de la VaR. D'où l'absence de durées et l'impossibilité de calculer les deux tests statistiques précédents. Pour être sûr d'obtenir au moins une violation même sur petit échantillon, il suffit d'augmenter le seuil de risque. C'est pourquoi, les tests de *backesting* développé à présent sont dits multivariés puisqu'ils sont calculés en utilisant plusieurs taux de couverture α. Hurlin et Tokpavi (2007) proposent une statistique de Portmanteau Multivarié dans le but de tester simultanément l'absence d'autocorrélation dans le vecteur des séquences de la variable *Hit* obtenu à des seuils de couverture différents. La quantité d'information utilisée est alors plus importante que pour les tests fondés sur un unique niveau de risque, d'où la plus grande puissance de ce test. Techniquement, il faut construire le vecteur $Hit_t(\theta_i)$ de dimension (T,m) où T est la taille de l'échantillon et m correspond au nombre de seuils de couverture différents employés. Sous l'hypothèse nulle, ce vecteur doit respecter la condition suivante :

$$\begin{aligned} H_0 \quad : \quad & E\left[Hit_t(\theta_i) Hit_{t-k}(\theta_j)'\right] = 0 \\ & \forall k = 1,...,K, \qquad \forall\,(\theta_i,\theta_j) \in \Theta = \{\theta_i,...,\theta_m\}. \end{aligned} \tag{16}$$

Pour implémenter cette statistique, il faut ensuite définir la matrice des variances covariances empiriques de ce vecteur Hit_t, que l'on nomme $\hat{C}_k$:

$$\hat{C}_k = (\hat{c}_{ijk}) = \sum_{t=k+1}^{T} Hit_t Hit'_{t-k} \qquad \forall k \in \mathbb{N}^*. \tag{17}$$

Soit la matrice $\hat{R}_k = D\hat{C}_k D$ avec D la matrice diagonale contenant les écarts-types associés au processus univarié des $Hit_t(\theta_i)$, définis par $\sqrt{\hat{c}_{ijk}}$ pour $i = 1, \ldots, m$. Dès lors, la statistique de test sous l'hypothèse nulle établie est égale à :

$$Q_m(K) = T \sum_{k=1}^{K} \left(vec\widehat{R}_k \right)' \left(\widehat{R}_0^{\ -1} \otimes \widehat{R}_0^{\ -1} \right) \left(vec\widehat{R}_k \right) \underset{T\to\infty}{\overset{L}{\longrightarrow}} \chi^2_{(K\ m^2)}. \tag{18}$$

Le dernier test employé est également un test multivarié qui est la généralisation multinomiale du test de *backtesting* sur la couverture non conditionnelle développé par Kupiec (1995). Pérignon et Smith (2008) se concentrent toujours sur la distribution des P&L mais avec plusieurs niveaux de couverture au cas où ce test rejetterait un modèle mal spécifié qui aurait pourtant été validé par le test univarié. Plus d'informations concernant la forme de la queue gauche de la distribution sont ainsi récupérées, permettant d'améliorer la puissance du test. Soit K différents taux de couverture classés par ordre décroissant, $p_1 > p_2 > ... > p_K$. Les K prévisions associées de la VaR sont alors rangées par ordre croissant $VaR_{t+1|t}(p_1) < VaR_{t+1|t}(p_2) < ... < VaR_{t+1|t}(p_K)$. Désormais, nous associons à ces K prévisions de la VaR, une variable indicatrice J définie de la manière suivante :

$$J_{i,t+1} = \begin{cases} 1 & \text{si } -VaR_{t+1|t}(p_{i+1}) < R_{t+1} \leq -VaR_{t+1|t}(p_i) \\ 0 & \text{sinon} \end{cases}, \tag{19}$$

pour $i = 1, ..., K$. Matrice à laquelle nous concaténons horizontalement la variable $J_{0,t+1} = \prod_{i=1}^{K}(1 - J_{i,t+1})$. La variable $J_{i,t+1}$ est une variable aléatoire suivant une loi de Bernoulli valant 1 avec une probabilité de $\theta_i = p_i - p_{i+1}$ quand $i \succeq 0$ et $J_{0,t+1} = 1$ avec une probabilité $\theta_0 = 1 - p_1$. Soit θ le vecteur de dimension K des paramètres suivants $\theta = (\theta_1, ...\theta_K)'$, $n_i = \sum_{t=1}^{T} J_{i,t}$ et $\hat{\theta}_i$ l'estimateur du maximun de vraisemblance de θ avec i éléments donnés par $\hat{\theta}_i = (1/T)\sum_{t=1}^{T} J_{i,t}$. Alors sous l'hypothèse nulle, les auteurs proposent une statistique de type LR définie par :

$$LR_{MUC} = 2\left(\ln\left[\frac{\left(l - l'\hat{\theta}\right)}{(l - l'\theta)^{n_0}} \right] + \sum_{i=1}^{K} \ln\left(\frac{\hat{\theta}_i}{\theta_i} \right)^{n_i} \right) \underset{T\to\infty}{\overset{L}{\longrightarrow}} \chi^2_{(K)}. \tag{20}$$

Si $K = 1$, on retrouve le test de Kupiec (1995) de couverture non conditionnelle.

Par conséquent, ce large panel de tests, mis en application dans les simulations Monte-Carlo de la section suivante, permet de comparer les distorsions de taille et de puissance de chacun des tests afin de savoir comment réagissent les différentes catégories de tests au risque de données.

4 Monte-Carlo design

Cette section se propose de quantifier les effets de la contamination des données sur les tests d'évaluation de la VaR. Jusqu'à présent la contamination a été modélisée dans son ensemble, c'est-à-dire sans distinguer les différentes caractéristiques de chaque source de pollution. L'innovation de notre étude est donc de générer des données de P&L pour chaque type de pollution. Après avoir expliquer la méthodologie, nous commenterons les résultats de nos trois expérimentations en comparant les tailles empiriques et les puissances empiriques obtenues sur les données avec et sans contamination, pour ainsi mettre en avant les déformations de taille et puissance engendrées par les données contaminées sur les tests de validation.

La méthodologie commune à nos trois expérimentations consiste à générer $10,000$ itérations d'un DGP simulant des P&L. Le choix du DGP permet de construire les véritables prévisions de la VaR puisque nous connaissons les paramètres du processus, nous simulons ainsi sous l'hypothèse nulle des tests de *backtesting*. Pour simuler sous l'hypothèse alternative, nous calculons de mauvaises prévisions de la VaR en utilisant une VaR non paramétrique de type *Historical Simulations*. En simulant nos propres données de P&L, nous nous affranchissons de tout risque de spécification, d'estimation et d'ajustement. Le risque de données, polluant nos données de P&L est développé dans les sous sections suivantes.

Pour chaque simulation, nous calculons les statistiques de tests de *backtesting*. La taille empirique correspond alors à la fréquence de rejet de l'hypothèse de couverture conditionnelle (ou de couverture non conditionnelle pour le test de Pérignon et Smith) observée sur ces simulations, lorsque nous sommes sous l'hypothèse nulle, c'est-à-dire les situations pour lesquelles le modèle interne de prévisions de la VaR est adéquat. Pour chaque test statistique de validation implémenté dans nos expériences, cette fréquence de rejet devrait être proche de la taille nominale, c'est-à-dire de l'erreur de type I, qui est fixée à 5%. Cette taille empirique devrait tendre vers la taille nominale au fur et à mesure que la dimension de l'échantillon des données simulées tend vers l'infini. La puissance empirique correspond à la fréquence de rejet de l'hypothèse nulle lorsque nous sommes sous l'hypothèse alternative, c'est-à-dire les situations pour lesquelles le modèle de prévisions de la VaR n'est pas adéquat. Le niveau de significativité des tests étant toujours fixé à 5%, la puissance empirique devrait tendre vers 95%. Mais les tests de validation sont peu puissants, on ne devrait donc pas atteindre ce seuil.

Les dimensions utilisées d'échantillon, notées T, sont au nombre de six : 250, 500, 750, 1000, 1250 et 1500. Les tests sont appliqués à deux séries distinctes de violations, car calculées pour deux taux de couverture distincts (1% et 5%). Pour les tests multivariés, deux séries

de violations sont donc utilisées simultanément. Le nombre de polynômes orthonormaux sélectionnés pour appliquer le test GMM est de cinq. C'est également le chiffre utilisé pour calculer l'ordre de l'autocorrélation pour le $Q_m(K)$ test. Une dernière précision, étant donné que certains tests ne peuvent pas être calculés s'il n'y a pas de violations, nous avons uniquement travaillé sur les séries générant au moins deux violations. Toutefois, cela n'empêche pas le fait que la valeur de la statistique de test de Christoffersen et Pelletier ne soit pas déterminée pour chaque échantillon en raison de la non convergence de l'estimateur sous l'alternative.

4.1 La contamination historique

4.1.1 Méthodologie

Afin de prendre en compte cette pollution, imaginons un portefeuille multivarié constitué de N actifs. Soit R_t les rendements simulés de chacun des actifs :

$$\underset{(N,1)}{R_t} = \underset{(N,1)}{\varepsilon_t} = \underset{(N,N)}{\sqrt{H_t}}\ \underset{(N,1)}{z_t} \qquad \text{où} \qquad z_t \overset{L}{\underset{i.i.d.}{\sim}} N(0, I_N). \tag{21}$$

Le processus ε_t, qui équivaut à celui des rendements, satisfait une représentation du modèle BEKK(p,q,K) proposé par Baba, Engle, Kraft et Kroner. La composante H_t correspond donc à la volatilité conditionnelle, les matrices A_{ik} et G_{ik} sont carrées, de taille (N, N). La matrice $\Omega = C_0'C_0$ est symétrique définie positive de taille (N, N). Pour s'assurer que H_t soit définie positive, il faut que les matrices H_{t-i} pour tout $i = 1, ..., p$ le soient également.

$$H_t = C_0'C_0 + \sum_{k=1}^{K}\sum_{i=1}^{q} A_{ik}'\varepsilon_{t-i}\varepsilon_{t-i}'A_{ik} + \sum_{k=1}^{K}\sum_{i=1}^{p} G_{ik}'H_{t-i}G_{ik}. \tag{22}$$

Le choix de cette modélisation nous permet de construire deux portefeuilles distincts, celui formant les P&L historiques (données contaminées) et l'autre constitué des P&L rétrospectifs (données non contaminées). On a alors $P\&L^H : \widetilde{R}p_t = w_{i,t}R_{i,t}$ et $P\&L^{CH} : Rp_t = w_{i,t-1}\ R_{it}$ où $w_{i,t}$ est le poids de l'actif i à la date t. Les poids de chaque actif varient à chaque date t et correspondent à la construction d'un portefeuille à variance minimale. Les poids à la date t sont donc ceux qui minimisent la volatilité conditionnelle du portefeuille à la date $t-1$. Les rendements de nos deux portefeuilles vont ensuite être systématiquement comparés aux prévisions de la VaR réalisées sur les rendements rétrospectifs :

$$VaR_{Rp_{t|t-1}} = -\Phi^{-1}(\alpha)\sqrt{w_{i,t-1}'H_t w_{i,t-1}}. \tag{23}$$

Les paramètres sont repris de l'article de Tse (2000) et le nombre d'actifs choisi est de trois. Les matrice Ω, A et G sont paramétrées comme suit :

$$\Omega = \begin{pmatrix} 0.8 & 0.2 & 0.2 \\ 0.2 & 0.8 & 0.2 \\ 0.2 & 0.2 & 0.8 \end{pmatrix}, \; A = \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0 & 0.3 & 0.1 \\ 0 & 0 & 0.3 \end{pmatrix}, \; G = \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0 & 0.5 & 0.2 \\ 0 & 0 & 0.5 \end{pmatrix}.$$

La figure 1 permet de comparer sur un échantillon de 250 jours les P&L rétrospectifs avec les P&L historiques afin de mieux en apprécier les quelques différences. Bien que le minimum et le maximum des rendements pollués soient supérieurs à ceux des rendements non pollués, nous remarquons que la distribution des données historiques a une variance légèrement plus faible que celle des données rétrospectives, ce qui est attendu puisque les poids utilisés sont ceux qui minimisent la variance en $t-1$. Son asymétrie à gauche est moins marquée, en revanche ses queues de distribution sont plus épaisses. Enfin, la moyenne de cette série historique est plus faible que celle de la série rétrospective. Les figures 2 et 3 délivrent une représentation graphique des prévisions de la VaR associées aux rendements non contaminés. Ces prévisions sont ensuite comparées aux rendements rétrospectifs et historiques générés par le modèle pour ainsi en construire la séquence des violations de la VaR pour chaque type de données. Le nombre d'exceptions observé sur les données rétrospectives est de 7 contre 6 avec les données historiques.

Les tests de validation sont ensuite calculés pour chaque simulation de ce processus. Deux séries distinctes de violations de la VaR sont ainsi construites, celle avec les données sans contamination et l'autre avec les données contaminées.

4.1.2 Résultats

Cette sous section présente les résultats empiriques de notre expérimentation. Nous présentons tout d'abord quelques caractéristiques des séries générées, puis nous analysons les séries de violations. Enfin, nous commentons les tailles et les puissances de nos tests de *backtesting*.

Comme nous l'a fait remarquer la figure 1, les données rétrospectives et historiques simulées sont très proches. L'étude de leurs moments réalisée avec le tableau 1, nous signale que ces séries sont centrées autour de 0. La moyenne des données polluées est sensiblement égale à celle des données propres. Leurs variances sont proches de 1, mais comme attendu la variance moyenne des séries historiques est plus faible que celle des séries rétrospectives.

Tab. 1 – Moyenne des moments empiriques des séries de P&L générés

	Moyenne	Variance	Skewness	Kurtosis
$T = 250$	-0.0102	0.9622	-0.0546	3.1776
	(-0.0099)	(0.9518)	(-0.0565)	(3.1710)
$T = 500$	-0.0012	0.9337	-0.0068	3.1229
	(-0.0011)	(0.9239)	(-0.0068)	(3.1153)
$T = 750$	-0.0007	0.9309	-0.0013	3.1183
	(-0.0006)	(0.9210)	(-0.0016)	(3.1107)
$T = 1000$	0.0000	0.9295	-0.0012	3.1218
	(0.0000)	(0.9196)	(-0.0013)	(3.1137)
$T = 1250$	-0.0004	0.9296	-0.0008	3.1226
	(-0.0004)	(0.9196)	(-0.0011)	(3.1139)
$T = 1500$	-0.0004	0.9300	-0.0011	3.1242
	(-0.0004)	(0.9201)	(-0.0012)	(3.1167)

Notes : pour chaque simulation, nous en calculons les quatre premiers moments. Le tableau récapitule les moyennes de ces moments pour chaque type de données et chaque taille d'échantillon. Les chiffres entre parenthèses correspondent à ceux provenant des données historiques tandis que les autres proviennent des données rétrospectives.

Il faut en voir la conséquence de la minimisation de la variance sur les P&L historiques. La distribution de nos séries est symétrique et légérement leptokurtique. Les queues de distributions de nos séries rétrospectives sont systématiquement plus épaisses que celles de nos séries historiques. Les caractéristiques de nos données de P&L provenant de notre processus générateur de données (équations 21 et 22) sont bien en adéquation avec les résutats attendus.

Le plus faible nombre de violations obtenu avec les données historiques s'explique par l'optimisation réalisée en $t - 1$ afin de déterminer les poids du portefeuille à la date t. Cette minimisation de la volatilité conditionnelle tend donc à diminuer le nombre d'exceptions. Bien que la différence moyenne de violations entre nos deux séries soit toujours et assez largement inférieure à un, il s'avère que la moyenne des violations issue des séries rétrospectives est statistiquement différente de celle obtenue avec les séries historiques, d'après le test de Student. Par ailleurs, le test non paramétrique de Kolmogorov-Smirnov, rejette toujours l'hypothèse nulle selon laquelle la distribution du nombre de violations observées sur les données historiques est la même que celle obtenue avec les données rétrospectives. Le tableau 2 nous permet d'affirmer que ce processus génère bien un nombre de violations différent selon le

Tab. 2 – Nombre de violations observé pour différentes prévisions de la VaR

	VaR 1%	VaR 5%	VaR HS 1%	VaR HS 5%
$T = 250$	3.4150	13.4943	3.2103	13.4099
	(3.3506)	(13.3117)	(3.1555)	(13.1808)
$T = 500$	5.2282	25.2593	4.9579	24.9813
	(5.1157)	(24.8796)	(4.8115)	(24.5390)
$T = 750$	7.5479	37.614	7.3397	37.4352
	(7.3502)	(37.0119)	(7.1039)	(36.7885)
$T = 1000$	9.9726	49.9825	9.7518	49.8937
	(9.7336)	(49.2050)	(9.4424)	(49.0632)
$T = 1250$	12.4836	62.5708	12.1944	62.3122
	(12.1571)	(61.5886)	(11.8179)	(61.2947)
$T = 1500$	15.0078	75.0999	14.6048	74.8241
	(14.6298)	(73.9900)	(14.1666)	(73.5266)

Notes : pour chaque simulation, nous en calculons le nombre de violations sur la série rétrospective et sur la série historique avec les bonnes et les mauvaises prévisions de la VaR. Puis nous en reportons le nombre moyen pour chaque taille d'échantillon et données rétrospectives mais aussi historiques (chiffres entre parenthèses).

type de données utilisées. Sachant que la prévision de la VaR est valide, nous retrouvons de manière quasi systématique, avec les données rétrospectives, une moyenne de violations parfaitement égale au niveau de risque, et cela pour chaque taille d'échantillon. En revanche, avec les données historiques, ce n'est pas toujours le cas. Cette moyenne de violations est toujours inférieure à celle calculée avec les données rétrospectives. En moyenne sur nos $10,000$ simulations, les données historiques ont moins d'exceptions que les données rétrospectives. Cette baisse s'étend en moyenne de -1.89% à -2.62% avec les bonnes prévisions de la VaR, et entre -1.70% et -3% pour les violations obtenues avec les prévisions de la VaR de type *Historical Simulation*. Par ailleurs, les nombres moyens de violations observées en comparant les P&L avec les prévisions de la VaR HS sont systématiquement inférieurs à ceux obtenus avec les prévisions paramétriques de la VaR. Ce résultat est valide pour tous les taux de couverture que l'on utilise des données propres ou contaminées.

Le tableau 3 permet de comparer les tailles empiriques des tests usuels de *backtesting* appliqués sur les données non polluées et polluées. Avant de commenter les déformations de tailles existantes entre nos séries, il faut tout d'abord s'assurer que les résultats obtenus sont en adéquation avec ceux des autres publications. Nous commentons donc les tailles empiri-

TAB. 3 – Tailles empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0308	0.0945	0.1650	0.0923	0.0596	0.0623	0.0017	0.0135	0.1283	0.0397
	(0.0290)	(0.0905)	(0.1555)	(0.0824)	(0.0590)	(0.0636)	(0.0018)	(0.0145)	(0.1230)	(0.0370)
$T = 500$	0.0208	0.0838	0.1539	0.0603	0.0775	0.0669	0.0057	0.0244	0.1212	0.0365
	(0.0191)	(0.0859)	(0.1393)	(0.0541)	(0.0779)	(0.0692)	(0.0052)	(0.0289)	(0.1163)	(0.0370)
$T = 750$	0.0284	0.1044	0.1314	0.0608	0.0893	0.0767	0.0210	0.0308	0.1161	0.0513
	(0.0262)	(0.1040)	(0.1154)	(0.0461)	(0.0939)	(0.0801)	(0.0233)	(00327)	(0.1144)	(0.0485)
$T = 1000$	0.0389	0.1081	0.1109	0.0564	0.0872	0.0784	0.0295	0.0315	0.1125	0.0536
	(0.0372)	(0.1064)	(0.0969)	(0.0441)	(0.0913)	(0.0848)	(0.0338)	(0.0356)	(0.1071)	(0.0511)
$T = 1250$	0.0394	0.1286	0.0898	0.0553	0.0785	0.0844	0.0336	0.0341	0.1009	0.0522
	(0.0349)	(0.1328)	(0.0761)	(0.0466)	(0.0831)	(0.0907)	(0.0345)	(0.0384)	(0.1010)	(0.0540)
$T = 1500$	0.0432	0.1091	0.0891	0.0564	0.0708	0.0933	0.0305	0.0329	0.0979	0.0512
	(0.0390)	(0.1099)	(0.0703)	(0.0436)	(0.0771)	(0.0955)	(0.0364)	(0.0345)	(0.0887)	(0.0478)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle BEKK(1,1,1) avec des perturbations normales sur chacun des trois actifs. A chaque date la VaR correspondante est calculée avec le même modèle BEKK et satisfait l'hypothèse de couverture conditionnelle. La taille empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la taille empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la taille empirique calculée à partir des rendements historiques. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

TAB. 4 – Puissances empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2\,(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0503	0.1188	0.2970	0.2544	0.0516	0.0312	0.0046	0.0175	0.2646	0.0174
	(0.0468)	(0.1159)	(0.2844)	(0.2369)	(0.0476)	(0.0321)	(0.0044)	(0.0198)	(0.2564)	(0.0201)
$T = 500$	0.0347	0.1073	0.3422	0.2604	0.0694	0.0190	0.0075	0.0310	0.3205	0.0029
	(0.0334)	(0.1073)	(0.3257)	(0.2401)	(0.0742)	(0.0221)	(0.0092)	(0.0355)	(0.3063)	(0.0057)
$T = 750$	0.0520	0.1335	0.4095	0.3272	0.0502	0.0125	0.0092	0.0315	0.3470	0.0018
	(0.0562)	(0.1384)	(0.3894)	(0.3168)	(0.0583)	(0.0163)	(0.0133)	(0.0340)	(0.3398)	(0.0034)
$T = 1000$	0.0640	0.1488	0.4218	0.4043	0.0426	0.0137	0.0113	0.0344	0.3808	0.0013
	(0.0665)	(0.1518)	(0.4067)	(0.3908)	(0.0521)	(0.0146)	(0.0161)	(0.0390)	(0.3664)	(0.0026)
$T = 1250$	0.0958	0.1709	0.4438	0.4774	0.0418	0.0124	0.0115	0.0358	0.3881	0.0009
	(0.0914)	(0.1712)	(0.4263)	(0.4612)	(0.0471)	(0.0138)	(0.0173)	(0.0426)	(0.3703)	(0.0031)
$T = 1500$	0.1076	0.1947	0.4681	0.5642	0.0375	0.0134	0.0093	0.0370	0.4109	0.0004
	(0.1035)	(0.1956)	(0.4428)	(0.5402)	(0.0457)	(0.0151)	(0.0126)	(0.0456)	(0.3921)	(0.0018)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle BEKK(1,1,1) avec des perturbations normales sur chacun des trois actifs. A chaque date la VaR HS est simplement le quantile non conditionnelle des 250 précédentes observations journalières des P&L. La puissance empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la puissance empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la puissance empirique calculée à partir des rendements historiques. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2\,(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

ques obtenues avec les données propres. Le test LR_{CC} est *well sized* au niveau de couverture de 1% mais *oversized* à 5%, le $LR_{CC}^{durée}$ est *oversized* pour les deux taux de couverture. La $J_{CC}(5)$ est *undersized* à 1% et 5%. Ces résultats confirment ceux délivrés par l'article de Candelon, Colletaz, Hurlin et Tokpavi (2010). Le test DQ_{CC} de Engle et Manganelli (2004) est *well sized* à 5% mais *oversized* à 1%. La statistique de test $Q_2(5)$ est ici *oversized* tout comme dans l'article de Hurlin et Tokpavi (2007). Enfin, l'évolution de la taille empirique observée sur le test LR_{MUC_2} est la même que dans l'article de Pérignon et Smith (2008) où le test est *well sized* puisque plus la taille de l'échantillon croit plus la taille empirique tend à être proche de la taille nominale. En comparant les différences de tailles empiriques obtenues avec les données rétrospectives avec celles calculées avec les données historiques, nous remarquons des différences systématiques. De manière générale, bien que dans l'ensemble les différences ne soient pas amples, nous constatons une baisse des taux de rejet pour les tests fondés sur la fréquence des violations et les tests multinomiaux. Au contraire des tests fondés sur les durées, où nous assistons à une hausse des taux de rejet. Les différences de tailles sur le test de Christoffersen (1998) sont très marquées au seuil de risque de 1% mais peu évidentes à 5%. Le sous-rejet des violations obtenues avec les données historiques est donc évident uniquement à 1%. Ce n'est pas la cas pour le test de Engle et Manganelli (2004), où les déformations sont très amples et ce pour n'importe quel taux de couverture. Les taux de rejet sont également plus faibles pour le test de Hurlin et Tokpavi (2007) et celui de Pérignon et Smith (2008) même si leur amplitude est assez faible. Les tests sur les durées ont une réaction atypique puisqu'ils rejettent plus souvent l'hypothèse nulle avec les données historiques qu'avec les données rétrospectives. La différence est une nouvelle fois minime mais croissante avec la taille de l'échantillon. Le test de Christoffersen et Pelletier (2004) est encore plus *oversized* tandis que la J_{stat} tendrait à être *well-sized*.

Le tableau 4 regroupe les puissances empiriques des tests de *backtesting* obtenues sur les données rétrospectives et historiques en utilisant des prévisions de la VaR de type *Historical Simulation*. La littérature actuelle sur le sujet met en évidence le fait que la plupart des tests de *backtesting* sont peu puissants. Nos résultats nous le confirment, la meilleure puissance obtenue est de 56,42% avec le test de Engle et Manganelli (2004) sur données rétrospectives, au seuil de risque de 5% et pour une taille d'échantillon de 1500. Le test multinomial de Pérignon et Smith est très peu puissant lorsque l'on utilise une VaR HS, le même résultat est obtenue dans l'article de référence. Les prévisions de VaR HS n'entrainent pas une baisse du nombre de violations assez significative pour rejeter de manière plus pronocée l'hypothèse de couverture non conditionnelle. Le test $Q_2(5)$ ainsi que les tests basés sur la fréquence des violations voient leur puissance se réduire avec les données contaminées. La différence de puissance est de plus en plus ample au fur et à mesure que la taille de l'échantillon croit.

Le manque de puissance est donc renforcé pour les tests fondés sur les violations et les tests multinomiaux si l'on utilise des données historiques pour valider les prévisions de la VaR. Ce n'est pas le cas pour les tests fondés sur les durées, leur puissance reste largement inférieure à 10% pour chaque taille de l'échantillon et tout type de données. Toutefois, la puissance du test $LR_{CC}^{durée}$ et $J_{CC}(5)$ augmente lorsque l'on utilise des données contaminées. Nos résultats ont la particularité d'être très différents d'un type de procédures de validation à l'autre. Ainsi, nous pouvons séparer ces derniers en deux groupes. Le premier regroupe les tests fondés sur la fréquence des violations et les test multinomiaux. Valider un modèle de prévisions de la VaR avec ce type de procédures et des données polluées par l'utilisation des poids historiques conduit globalement à des tailles et des puissances empiriques plus faibles pour chaque seuil de risque. Les modèles de prévisions de la VaR qu'ils soient bon ou non tendent donc à être moins rejetés si l'on utilise des données contaminées au lieu de données propres. Le second groupe est formé par les tests fondés sur les durées, les déformations de tailles et de puissance empiriques sont alors dans le sens inverse de ce que l'on observe sur le groupe précédent. L'emploi de données polluées entraîne donc une légère augmentation de la taille et de la puissance de ces tests bien que le nombre de violations observées diminue avec les P&L contaminées. Le résultat le plus surprenant concerne le test de Candelon et consorts qui devrait être bien plus puissant que ce que l'on observe dans nos simulations.

Notre prochaine expérimentation s'intéresse à la pollution produite par les activités de *trading intraday* et à leurs conséquences sur les procédures de validation.

4.2 Les revenus de l'intraday trading

4.2.1 Méthodologie

Modéliser ce type de pollution relève du challenge méthodologique. D'une part, il faut mettre en place une technique d'allocation d'actifs pour capturer cette contamination. D'autre part, agréger nos données *intraday* pour en construire des rendements quotidiens n'est pas aisé, puisque cela modifie le modèle suivi par nos données. Drost et Nijman (1993) ont mis en évidence la déformation des processus GARCH suite à une agrégation temporelle, et plus précisemment lors du passage de données journalières à des données hebdomadaires. Toutefois, le passage de données *intraday* à des données quotidiennes n'étant pas évoqué dans cet article, nous devons assumer le fait que le modèle suivi par les données de haute fréquence (journalières) demeure inconnu.

Deux approches similaires sont utilisées pour générer des données d'ultra haute fréquence (*intraday*). Ces données peuvent être simulées soit à intervalle de temps constant, soit à intervalle irrégulier comme c'est le cas dans la réalité puisque chaque changement de cours

correspond à l'enregistrement d'une transaction. Bien que la première technique soit soumise à quelques critiques, notamment la perte d'informations, nous avons tout de même décider dans une optique de simplification de l'utiliser, laissant ainsi de côté la seconde modélisation proposée par Engle (2000). En effet, l'idée principale de cette sous-section est de réussir à intégrer une technique de *trading* permettant de polluer nos données et non pas de simuler des données parfaitement fidèle à la réalité. L'article de Giot (2000) étudie d'ailleurs ces deux types de données d'ultra haute fréquence, sur un même actif, afin d'en étudier l'autocorrélation et la volatilité. Sur les données espacées à intervalle régulier de 5 minutes, entre Septembre 1996 et Novembre 1996 (soit 13 semaines), il estime notamment un simple modèle GARCH(1,1) sur les rendements désaisonnalisés d'ultra haute fréquence de l'action IBM. Ces rendements sont calculés chaque jour à partir des échanges dits réguliers, c'est-à-dire intervenant entre 9h30 et 16h. Nous reprenons alors la paramétrisation issue de son estimation afin de générer 78 $(12 * 6,5)$ rendements intraday. Ce chiffre correspond au nombre de cotations à intervalle de 5 minutes qu'il y a dans l'intervalle des échanges réguliers. Le DGP se résume ainsi :

$$R_t = 0.033 - 0.04\ R_{t-1} + \varepsilon_t \quad \text{où} \quad \varepsilon_t = \sqrt{H_t}\ z_t \quad \text{et} \quad z_t \backsim N(0,1), \tag{24}$$

$$H_t = 0.02 - 0.068\ \left(\varepsilon_{t-1}\right)^2 + 0.912\ H_{t-1}. \tag{25}$$

Les rendements suivent un processus AR(1), le coefficient associé à ce terme est négatif. Cela signifie que l'on a une légère autocorrélation négative à cette fréquence correspondant à un effet de rappel dans les rendements. Nous transformons ces données en prix que l'on note P_i pour $i = 1, 2, ..., 78$, puis nous en calculons un rendement logarithmique non pollué correspondant à notre $P\&L^{CH}$ à l'aide de l'équation 26 :

$$R_t = \log\left(\frac{P_{78}}{P_1}\right). \tag{26}$$

Une fois nos données transformées en prix, nous utilisons une analyse chartiste très simple afin de calculer les revenus de *trading* réalisés au cours de la journée. Ces revenus proviennent d'une technique élémentaire de prise de position, faisant suite à la lecture de deux moyennes mobiles d'ordre différent appliquées sur les prix : une courte d'ordre 3 (tous les quarts d'heure) et l'autre longue d'ordre 12 (toutes les heures). Si la moyenne mobile courte coupe à la hausse la moyenne mobile longue, alors le trader décide d'acheter l'action, et il vend lorsque la moyenne mobile courte coupe à la baisse la moyenne mobile longue. Une position est donc prise à chaque fois que ces moyennes mobiles se coupent. Le cadran supérieur de

la figure 4 permet de visualiser cette technique sur 5 jours de cotation. Pour ne pas polluer ces revenus par d'autres sources de contamination, nous générons nos rendements pour un portefeuille constitué uniquement d'un seul actif dont le poids en début et fin de journée de *trading* est égal à l'unité. Cela n'empêche en rien la capacité du trader à faire fluctuer le poids de cet actif tout au long de la journée en fonction de sa stratégie de prise de position. Toutefois, il se peut qu'à la fin de la journée, le nombre de croisements des moyennes mobiles soit impair. Cela signifie donc que le portefeuille n'est pas constitué d'un unique actif. Soit il y en deux actifs, soit il y en a aucun dans le portefeuille. Une position de rééquilibrage est donc appliquée dans ces deux cas de figure à chaque fin de journée afin d'obtenir un portefeuille formé d'un seul et unique actif. Les gains ou les pertes réalisés sur la journée sont ensuite additionnés et constituent le montant de la pollution quotidienne noté π_t. Le rendement pollué, c'est-à-dire $P\&L_t^I$, est donc calculé en utilisant l'équation 27 :

$$R_t^C = \log\left(\frac{P_{78} + \pi_t}{P_1}\right) \tag{27}$$

Nous répétons T fois cette simulation, où T correspond à la taille de l'échantillon sélectionnée, afin d'obtenir $T*78$ rendements ultra haute fréquence et ainsi calculer T rendements journaliers propres mais aussi T rendements journaliers contaminés par les revenus de l'*intraday trading*. Le cadran inférieur de la figure 4 représente une évolution quotidienne des montants de la pollution *intraday* sur 250 jours, noté π_t, par un diagramme en bâtons. On remarque que cette pollution n'est pas uniquement positive ou négative. Cette caractéristique est très importante puisqu'elle ne nous permet pas de situer la distribution des données de $P\&L_t^I$ par rapport à celle des données non polluées. Mais déterminer avec précision le modèle suivi par ces rendements quotidiens n'est pas possible. En effet, si nous appliquons la formule de Drost et Nijman (1993) afin de tenter de trouver le modèle GARCH correspondant à ces données quotidiennes et bien nous trouvons des coefficients ARCH et GARCH très proche de 0. Ce résultat est tout à fait normal et l'absence d'autocorrélation dans les carrés des rendement journaliers nous confirme qu'un modèle GARCH ne peut pas modéliser ces rendements journaliers. Dès lors, nous assumons le fait que nous ne connaissons pas le modèle suivi par nos rendements journaliers. La démarche adoptée afin de déterminer une vraisemblable bonne prévision de la VaR est d'utiliser les prévisions de la VaR faites pour le lendemain matin 9h30 sachant que nous sommes la veille à 16h.

$$VaR_{R_{(79*t)|(78*t)}} = -\left(0.033 - 0.04R_{(78*t)} + \Phi^{-1}(\alpha)\sqrt{H_{(79*t)}}\right) \qquad \forall t = 1, ..., T. \tag{28}$$

Afin d'obtenir une approximation des prévisions de la VaR quotidiennes, nous transformons les prévisions *intraday* de la VaR comme l'explicite l'équation 29 :

$$VaR_{R_{t|t-1}} = \sqrt{78}\left[VaR_{R_{(79*t)|(78*t)}}\right] \qquad \forall t = 1, ..., T. \tag{29}$$

Une fois ces T prévisions journalières de la VaR obtenues, nous les comparons aux $P\&L_t^{CH}$ et aux $P\&L_t^I$. La figure 5 représente l'évolution sur 250 jours des rendements non pollués ainsi que les approximations des prévisions de la VaR, le cadran du bas contient les violations associées à ce jeu de données. La figure 6 est identique à la précédente à l'exception du fait que l'on s'intéresse au rendements polluées par les revenus *intraday*. Sur cet exemple, nous observons plus de 1% de violations, sur les $P\&L_t^{CH}$ nous observons 5 violations dont deux consécutives tandis que nous avons 8 violations sur les $P\&L_t^I$. Ce type de pollution soit négatif soit positif augmente ici le nombre de violations. Au regard des statistiques descriptives présentent sur les deux figures, cela est la conséquence d'une augmentation de la variance mais surtout d'une explosion de la kurtosis. En plus d'une technique de trading sans aucun doute trop élémentaire puisqu'elle ne permet pas de générer des revenus régulièrement positifs, dans notre exemple la pollution moyenne sur l'échantillon est de -0.0552, ce qui est particulièrement faible, il faut être prudent avec les interprétations de la sous section suivante puisque nous devons tenir compte d'un risque de modèle puisque nos prévisions de la VaR utilisées pour étudier la puissance ne sont que des approximations de leur véritable valeur. En revanche l'étude de la puissance est réalisée sans aucun problème.

4.2.2 Résultats

Il est évident que la déformation entrainée par ces gains et pertes journalières modifie énormément les caractéristiques des rendements, mais quel est l'impact de l'utilisation de ces données polluées sur les procédures de *backtesting* de la VaR.

Cette expérimentation produit des résultats très surprenants qui doivent être expliqués en détail. La table 5 contient les moyennes par taille d'échantillon des quatre premiers moments empiriques. En moyenne les rendements journaliers qu'ils soient contaminés ou non sont centrés sur 0. Les $P\&L^I$ ont une légère asymétrie à droite. La variance des données non polluées est plus faible que celle des données contaminées par les revenus intraday, il en est de même pour la kurtosis mais la différence entre ces deux types de données est légèrement plus marquée pour la kurtosis que pour la variance. Il n'y a pas d'évolution en fonction de la taille d'échantillon, les résultats sont donc similaires pour toutes les tailles d'échantillon. Les déformations de tailles empiriques ne vont sans doute pas aller dans le même sens que pour la pollution historique puisque le nombre de violations observé sur les données

contaminées par les revenus *intraday* est largement supérieur à celui obtenu sur les données propres. Le tableau 6 résume ce nombre d'exceptions en reportant les chiffres moyens de violations par types de données, taille d'échantillon mais aussi types de prévisions de la VaR avec ses différents niveaux de risque. Dans tous les cas de figures, le nombre moyen de violations obtenu par taille d'échantillon sur les $P\&L^I$ est très largement supérieur à celui observé sur les $P\&L_t^{CH}$. Il faut également souligner le fait que les approximations des prévisions de la VaR conduisent à plus de violations en moyenne qu'avec les prévisions de type *Historical Simulation*. Ce résultat va conduire à un taux de rejet extrêmement élevé lorsque nous allons commenter les déformations de tailles empiriques.

TAB. 5 – Moyenne des moments empiriques des séries de P&L générés

	Moyenne	Variance	Skewness	Kurtosis
$T = 250$	-0.0007	0.0079	-0.0510	3.8750
	(-0.0025)	(0.0115)	(0.2666)	(5.1687)
$T = 500$	-0.0001	0.0077	-0.0014	3.8471
	(-0.0020)	(0.0113)	(0.3186)	(5.2503)
$T = 750$	0.0000	0.0077	-0.0006	3.8690
	(-0.0019)	(0.0113)	(0.3233)	(5.2836)
$T = 1000$	0.0000	0.0077	-0.0004	3.8767
	(-0.0019)	(0.0113)	(0.3204)	(5.3009)
$T = 1250$	0.0000	0.0077	0.0006	3.8813
	(-0.0019)	(0.0113)	(0.3218)	(5.3463)
$T = 1500$	0.0000	0.0077	0.0003	3.8778
	(-0.0018)	(0.0113)	(0.3258)	(5.3199)

Notes : pour chaque simulation, nous en calculons les quatre premiers moments. Le tableau récapitule les moyennes de ces moments pour chaque type de données et chaque taille d'échantillon. Les chiffres entre parenthèses correspondent à ceux provenant des données contaminées par les revenus *intraday* tandis que les autres proviennent des données non polluées.

Le tableau 7 récapitule les tailles empiriques calculées à partir des données non contaminées et contaminées. L'adéquation des tailles empiriques de chaque test statistique observées sur nos données n'est pas conforme à ce que l'on observe dans la littérature. Concernant les $P\&L^{CH}$, l'ensemble des tests de validation appliqué apparait *oversized*, seul le test de Candelon et consorts sur les durées est largement *undersized* pour tous les niveaux de risques mais uniquement sur les rendements non contaminés. En effet, avec les $P\&L^I$, l'intégralité

des tests sont *oversized* et la taille empirique de ces tests tend vers l'unité au fur et à mesure que la taille d'échantillon augmente. Ce phénomène était prévisible puisque le nombre de violations observées sur ces données contaminées est très important, les tests de validation le détecte et rejette l'hypothèse nulle même si celle-ci est vrai. Il y a tellement de violations que l'on viole de manière quasi systématique l'hypothèse de couverture non conditionnelle. Dans ces simulations, nous assistons à un sur-rejet du modèle de VaR proposé que ce soit avec les données contaminées ou non.

TAB. 6 – Nombre de violations observé pour différentes prévisions de la VaR

	VaR 1%	VaR 5%	VaR HS 1%	VaR HS 5%
$T = 250$	4.0087	13.8270	3.0252	13.1061
	(8.2434)	(19.9734)	(5.9306)	(18.8489)
$T = 500$	6.8596	26.4722	4.8436	24.8809
	(15.7902)	(39.2996)	(11.1197)	(36.9118)
$T = 750$	10.1955	39.5656	7.1926	37.3411
	(23.5754)	(58.8948)	(16.5842)	(55.4006)
$T = 1000$	13.5854	52.7757	9.5944	49.7829
	(31.4833)	(78.4931)	(22.0601)	(73.7930)
$T = 1250$	16.9973	65.9576	11.9810	62.2623
	(39.3123)	(98.2404)	(27.6561)	(92.2650)
$T = 1500$	20.4004	79.1618	14.3494	74.6223
	(47.0595)	(117.7163)	(33.0836)	(110.5466)

Notes : pour chaque simulation, nous en calculons le nombre de violations sur la série rétrospective et sur la série historique avec les bonnes et les mauvaises prévisions de la VaR. Puis nous en reportons le nombre moyen pour chaque taille d'échantillon et données non polluées mais aussi contaminées par les revenus *intraday* (chiffres entre parenthèses).

L'analyse se poursuit par la description des résultats du tableau 8 contenant la puissance empirique de nos tests de *backtesting* par taille d'échantillon et pour différents taux de couverture. Sur les données non polluées, la faible puissance des tests ressort parfaitement, nous sommes donc incapable de rejeter ces mauvaises prévisions de la VaR de type *Historical Simulation*. Seul le test de Engle et Manganelli (2004) a une puissance largement supérieure à tous les autres tests, il est suivi par le test de Hurlin et Tokpavi (2007). Une nouvelle fois il faut s'interroger sur la faible puissance de la J_{stat}. Par ailleurs, remarquons que ces prévisions de la VaR ne sont pas rejetées par le test multinomial de la couverture non conditionnelle puisque le nombre de violations obtenu est dans ce cas assez proche du niveau de risque. En

TAB. 7 – Tailles empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2\,(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0670	0.1055	0.2457	0.1042	0.0821	0.0673	0.0089	0.0127	0.1265	0.0786
	(0.6473)	(0.4852)	(0.8067)	(0.4893)	(0.5537)	(0.3272)	(0.3084)	(0.0783)	(0.1313)	(0.6830)
$T = 500$	0.0690	0.1044	0.2729	0.0799	0.0939	0.0691	0.0111	0.0194	0.1136	0.0984
	(0.8821)	(0.7130)	(0.9332)	(0.6620)	(0.8402)	(0.5850)	(0.6155)	(0.2279)	(0.1468)	(0.9056)
$T = 750$	0.1118	0.1145	0.2537	0.0805	0.1068	0.0747	0.0155	0.0230	0.1078	0.1291
	(0.9778)	(0.8636)	(0.9817)	(0.8028)	(0.9543)	(0.7790)	(0.8238)	(0.4204)	(0.1671)	(0.9810)
$T = 1000$	0.1347	0.1256	0.2454	0.0820	0.1199	0.0898	0.0203	0.0225	0.0916	0.1490
	(0.9934)	(0.9346)	(0.9938)	(0.8863)	(0.9897)	(0.8868)	(0.9340)	(0.6075)	(0.1888)	(0.9959)
$T = 1250$	0.1686	0.1371	0.2549	0.0822	0.1402	0.0942	0.0278	0.0215	0.0943	0.1784
	(0.9986)	(0.9762)	(0.9986)	(0.9471)	(0.9976)	(0.9570)	(0.9801)	(0.7620)	(0.2165)	(0.9991)
$T = 1500$	0.2081	0.1279	0.2638	0.0795	0.1546	0.1066	0.0289	0.0213	0.0872	0.1979
	(1.0000)	(0.9914)	(0.9999)	(0.9737)	(0.9997)	(0.9822)	(0.9938)	(0.8563)	(0.2349)	(1.0000)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle UHF. La taille empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la taille empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la taille empirique calculée à partir des rendements contaminés par les revenus de l'*intraday trading*. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2\,(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

TAB. 8 – Puissances empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2\,(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0278	0.0792	0.1995	0.1727	0.0455	0.0404	0.0017	0.0152	0.1663	0.0116
	(0.3383)	(0.4108)	(0.5662)	(0.4699)	(0.2565)	(0.2447)	(0.1257)	(0.0655)	(0.1708)	(0.4386)
$T = 500$	0.0174	0.0737	0.2270	0.1697	0.0712	0.0253	0.0045	0.0226	0.1696	0.0027
	(0.4958)	(0.6002)	(0.7284)	(0.6317)	(0.4235)	(0.4233)	(0.2411)	(0.1493)	(0.2000)	(0.6600)
$T = 750$	0.0240	0.0734	0.2708	0.2065	0.0469	0.0196	0.0066	0.0204	0.1659	0.0010
	(0.6757)	(0.7639)	(0.8202)	(0.7757)	(0.5802)	(0.6167)	(0.3639)	(0.2807)	(0.2205)	(0.8366)
$T = 1000$	0.0289	0.0800	0.2753	0.2525	0.0448	0.0223	0.0050	0.0224	0.1578	0.0012
	(0.7828)	(0.8643)	(0.8889)	(0.8702)	(0.7094)	(0.7534)	(0.4874)	(0.4209)	(0.2396)	(0.0171)
$T = 1250$	0.0303	0.0904	0.2802	0.3026	0.0377	0.0238	0.0056	0.0202	0.1515	0.0003
	(0.8637)	(0.9282)	(0.9369)	(0.9309)	(0.8069)	(0.8534)	(0.6141)	(0.5598)	(0.2755)	(0.9633)
$T = 1500$	0.0347	0.0833	0.2922	0.3762	0.0367	0.0259	0.0045	0.0198	0.1543	0.0003
	(0.9210)	(0.9590)	(0.9630)	(0.9591)	(0.8708)	(0.9144)	(0.7065)	(0.6826)	(0.3045)	(0.9834)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle UHF. La puissance empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la puissance empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la puissance empirique calculée à partir des rendements contaminés par les revenus de l'*intraday trading*. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2\,(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

revanche, avec les données contaminées par les revenus de l'*intraday trading* la puissance empirique de nos tests tend une nouvelle fois vers un, mais de manière moins fulgurante qu'avec les approximations de nos prévisions de la VaR. La raison étant que les prévisions de type HS entraînent moins de violations que nos autres prévisions de VaR qui sont censées être les meilleures.

La faiblesse de nos résultats s'expliquent par deux phénomènes. D'une part, le modèle de prévisions de la VaR n'est pas le bon puisque nous ne connaissons pas le modèle suivi par les rendements journaliers que nous calculons. Il nous faut alors trouver des régles d'aggrégation temporelle des modèles ARCH-GARCH afin de passer d'un modèle ultra haute fréquence à un modèle de haute fréquence. D'autre part, les revenus de trading produits ne sont pas toujours positifs. Ce dernier point est sans aucun doute la principale cause de la non mise en évidence du fait que les données contaminées par les revenus *intraday* sont à l'origine d'un sous rejet des modèles internes proposés par les banques lorsqu'on les valide par ce type de données. En effet, la technique de prises de positions mise en application permet de générer des revenus que très légèrement supérieurs à 0 en moyenne. Ces derniers ne sont donc pas strictement positifs. Or, dans la réalité, il est fort probable que les techniques appliquées soient beaucoup plus efficaces que celle proposée, permettant ainsi de générer un gain journalier beaucoup plus important. Les techniques de trading algorithmique et quantitatif sont sans doute á explorer et appliquer. Par ailleurs, le fait que notre portefeuille soit constitué d'un seul actif, limite considérablement les opportunités d'arbitrage et de prise de positions qui nous sont offertes. D'où l'obtention de revenus *intraday* plus faibles que dans la pratique. Ainsi, des modifications sont à apporter sur ces deux points afin de parfaitement mettre en évidence les déformations de tailles empiriques causées par les données contaminées par les revenus de l'*intraday trading*.

4.3 Frais de gestion et commissions

4.3.1 Méthodologie

La modélisation de la pollution provenant des frais de gestion et des commissions est réalisée à partir d'une *Mixture of Distribution Hypothesis* (MDH). Cette technique nous permet de générer une distribution jointe rendements – volumes, sachant l'arrivée d'une variable latente symbolisant bien souvent un flux d'information, noté K_t. Ces modélisations ont été appliquées dans le but de mesurer la corrélation positive existante entre la volatilité des rendements et les volumes échangés quotidiennement, en fonction des annonces économiques. Nous reprenons ainsi l'approche développée par Andersen (1996). Il propose un modèle MDH modifié, se basant sur ceux mis au point par Harris (1986 et 1987). Son objectif est de s'af-

franchir des limites du modèle de son prédécesseur en supposant que les volumes détrendés, noté $\tilde{V}_t$, suivent une loi de Poisson et non une loi normale. De fait, cela entraine la stricte positivité des volumes échangés. Il prend également en compte une composante de bruit, en introduisant un terme constant noté $m0$. Le coefficient $m1$ est un facteur de proportionnalité indiquant de combien le volume fluctue suite à une information. Son modèle se résume comme suit :

$$\begin{cases} R_t|K_t \sim N(\overline{r}, K_t) \iff & R_t = \overline{r} + z_t\sqrt{K_t} \quad \text{avec} \quad z_t \sim N(0,1) \\ \tilde{V}_t|K_t \sim c\ Po\left(m_0 + m_1 K_t\right) & \end{cases}, \tag{30}$$

où K_t est un processus de volatilité stochastique de type *Stochastic Auto-Regressive Volatility* (SARV) définit par $\sqrt{K_t} = \omega + \beta\sqrt{K_{t-1}} + \alpha\sqrt{K_{t-1}}u_t$, avec $u_t = |v_t| / E|v_t|$ et $v_t \sim GED_{w(r)}(0,1)$. Les paramètres sélectionnés pour notre génération de processus proviennent d'une estimation du modèle ci-dessus sur les rendements et les volumes de l'action IBM, réalisée sur une période allant de Janvier 1973 à Décembre 1991. Ils sont récapitulés dans le tableau 9 ci-dessous.

Tab. 9 – Paramétrisation utilisée

$\bar{r}$	c	m_0	m_1	ω	α	β	*scale parameter*	*shape parameter*
0.012	0.041	15.853658	4.1707317	0.31964	0.269	0.517	1.379	1.13

Afin de ne prendre en compte que la pollution issue des frais de gestion et des commissions, nous générons des données pour un portefeuille constitué d'un seul actif et dont le poids reste constant tout au long de l'expérimentation. Les rendements obtenus sont ensuite manipulés afin de les transformer en prix, noté P_t. En effet, notre moyen de contamination est d'incorporer à la valeur quotidienne du portefeuille, un montant de frais et commissions proportionnel au volume échangé de la journée. Nous avons décidé que ce montant correspondrait à 10% du volume détrendé échangé de la journée. En valeur, cela revient à inclure une somme comprise entre 0.02% et 0.2% de la valeur quotidienne du portefeuille. Nous y remarquons également les violations des prévisions de la VaR associées aux données non contaminées. Une fois le montant du portefeuille contaminé, nous en recalculons un rendement pollué pour obtenir le $P\&L^{F\&C}$ qui est égal à :

$$R_t^C = \frac{P_t + \phi_t}{P_{t-1}}. \tag{31}$$

Ensuite, les deux séries de rendements produites par ce DGP sont systématiquement comparées aux prévisions suivantes de la VaR :

$$VaR_{R_{t|t-1}} = -\overline{r} - \Phi^{-1}(\alpha)\sqrt{K_t}. \tag{32}$$

La figure 7 représente simultanément les rendements non pollués ainsi que les prévisions de la VaR qui lui sont associées afin d'en mettre en évidence les violations par l'intermédiaire de petits cercles. Le graphique du bas de cette figure comprend les mêmes éléments sauf que l'on est sur les rendements contaminés par les frais de gestion et les commissions. Nous remarquons que le nombre de violations est plus faible sur les données polluées, elles sont au nombre de 4 contre 6 pour les données non contaminés sur notre période d'étude. La principale raison de ce phénomène provient du fait que la moyenne des rendements de la série polluée est plus élevée que celle des données non contaminées par les frais de gestion et les commissions. La figure 8 permet de mieux apprécier cette pollution quotidienne en représentant l'évolution des volumes échangés sur la période ainsi que le montant de la pollution que l'on y associe. Ces derniers sont bien toujours positifs comme c'est le cas en pratique.

4.3.2 Résultats

Nous exposons dans un premier temps les caractéristiques des séries générées, puis dans un second temps nous analysons le nombre de violations produites faisant suite à la comparaison des données de P&L propres et contaminées avec les bonnes prévisions de la VaR mais aussi avec les prévisions de la VaR HS. Enfin, dans un troisième temps, nous commentons les tailles et les puissances de nos tests de *backtesting*.

Le tableau 10 récapitule la moyenne obtenue sur les quatre premiers moments empiriques avec les données propres et les données contaminées par les frais de gestion et les commissions. Les séries de $P\&L^{CH}$ ont bien une moyenne qui tend à se rapprocher de $\bar{r}$ tandis que celle des données contaminées est proche de 0.1. Cette différence est en adéquation avec ce que nous attendons puisque ce type de pollution est toujours positive, il est donc normal que le rendement de la série polluée soit supérieur à celui de la série non contaminée. La variance n'est pas très différente d'une série à l'autre et reste stable avec la taille d'échantillon. La distribution de nos séries est symétrique et leptokurtique, bien que l'on observe un léger décalage à droite de la distribution des $P\&L^{F\&C}$. Les queues de distributions de nos séries propres et contaminées sont donc plus épaisses que celles d'une loi normale mais la différence d'excès de kurtosis entre nos deux types de séries est vraiment infime.
Etant donné que la distribution des $P\&L^{F\&C}$ est décalée vers la droite comparée à celle des

Tab. 10 – Moyenne des moments empiriques des séries de P&L générés

	Moyenne	Variance	Skewness	Kurtosis
$T = 250$	$-0,0046$	$2,0505$	$-0,1092$	$5,4646$
	$(0,0853)$	$(2,0526)$	$(-0,0947)$	$(5,4598)$
$T = 500$	$0,0094$	$1,9538$	$-0,0108$	$5,5214$
	$(0,0991)$	$(1,9562)$	$(0,0045)$	$(5,5207)$
$T = 750$	$0,0120$	$1,9408$	$0,0040$	$5,6718$
	$(0,1017)$	$(1,9432)$	$(0,0199)$	$(5,6714)$
$T = 1000$	$0,0121$	$1,9453$	$0,0045$	$5,8144$
	$(0,1026)$	$(1,9479)$	$(0,0209)$	$(5,8156)$
$T = 1250$	$0,0127$	$1,9461$	$-0,0010$	$5,9015$
	$(0,1040)$	$(1,9488)$	$(0,0161)$	$(5,8972)$
$T = 1500$	$0,0118$	$1,9413$	$0,0012$	$5,9452$
	$(0,1045)$	$(1,9442)$	$(0,0184)$	$(5,9438)$

Notes : pour chaque simulation, nous en calculons les quatre premiers moments. Le tableau récapitule les moyennes de ces moments pour chaque type de données et chaque taille d'échantillon. Les chiffres entre parenthèses correspondent à ceux provenant des données contaminées par les frais de gestion et les commissions tandis que les autres proviennent des données non polluées.

$P\&L^{CH}$, nous nous attendons à obtenir moins de violations de la VaR avec nos données contaminées. Le tableau 11 confirme cette attente. En moyenne, le nombre de violations obtenu sur les données polluées est largement plus faible qu'avec les données propres. L'écart de violations se creuse au fur et à mesure que la taille d'échantillon croit. Le test de rang de Wilcoxon d'égalité de la médiane entre nos séries polluées et non polluées est systématiquement rejeté. Cette pollution a donc un impact significatif sur le nombre de violations produit avec les données contaminées. Il faut une nouvelle fois remarquer que le nombre de violations moyen observé avec les prévisions de la VaR HS est inférieur à celui obtenu avec les bonnes prévisions de la VaR et ce pour chaque de type de données et pour toutes tailles d'échantillon. Le tableau 12 permet de comparer les tailles empiriques des tests usuels de *backtesting* appliqués sur les données non polluées et polluées. Comme avec les données rétrospectives, les tests employés ont des tailles empiriques conforment à ce que l'on observe dans de précédentes études. Le test LR_{CC} n'est que très légèrement *undersized* au seuil de risque de 1% tandis qu'il est *oversized* au niveau de couverture de 5%. Le test DQ_{CC} de Engle et Manganelli (2004) est *well sized* à 5% mais *oversized* à 1%. Le test $LR_{CC}^{durée}$ de Christoffersen et Pelletier (2004)

TAB. 11 – Nombre de violations observé pour différentes prévisions de la VaR

	VaR 1%	VaR 5%	VaR HS 1%	VaR HS 5%
$T = 250$	$3,4484$	$13,4771$	$3,2225$	$13,5394$
	$(2,9986)$	$(11,7253)$	$(2,9615)$	$(12,1919)$
$T = 500$	$5,2915$	$25,2974$	$5,0067$	$25,1205$
	$(4,3991)$	$(21,7377)$	$(4,5142)$	$(22,5139)$
$T = 750$	$7,5740$	$37,5707$	$7,3702$	$37,5034$
	$(6,2508)$	$(32,2784)$	$(6,6108)$	$(33,5594)$
$T = 1000$	$9,9978$	$50,0164$	$9,8503$	$49,9622$
	$(8,2381)$	$(42,9491)$	$(8,8416)$	$(44,7035)$
$T = 1250$	$12,4563$	$62,4409$	$12,3241$	$62,4513$
	$(10,2410)$	$(53,5002)$	$(11,0617)$	$(55,8016)$
$T = 1500$	$15,0026$	$74,9716$	$14,7185$	$74,9008$
	$(12,2798)$	$(64,1651)$	$(13,1802)$	$(66,8063)$

Notes : pour chaque simulation, nous en calculons le nombre de violations sur la série propre et sur la série polluée, avec les bonnes et les mauvaises prévisions de la VaR. Puis nous en reportons le nombre moyen pour chaque taille d'échantillon et données non polluées mais aussi contaminées par les frais de gestion et les commissions (chiffres entre parenthèses).

est toujours *oversized* et cela pour chaque seuil, contrairement à J-stat qui est *undersized* . Le test multivarié $Q_2\,(5)$ de Hurlin et Tokpavi (2007) est *oversized* alors que le test LR_{MUC_2} de Pérignon et Smith (2008) est *well sized*. Pour interpréter les déformations de tailles, nous comparons la taille empirique obtenue avec les $P\&L^{CH}$ à celle acquise avec les $P\&L^{F\&C}$. Une nouvelle fois, les tests de *backtesting* peuvent être séparés en deux catégories. En effet, les tests fondés sur les durées et les tests multinomiaux sont soumis à de fortes distorsions de tailles. Ces tests déjà *oversized*, voient leur taille empirique augmenter de manière fulgurante avec ce type de pollution. Ainsi, la J-stat qui est pourtant *undersized* et le test LR_{MUC_2} qui lui est *well sized* deviennent largement *oversized*. Le plus faible nombre de violations est donc détecté par ce type de procédures statistiques. Ceci est particulièrement vrai lorsque la taille d'échantillon augmente. En revanche, nous remarquons qu'au niveau de risque de 1% et pour une taille d'échantillon de 250, alors la taille empirique observée sur les $P\&L^{F\&C}$ est plus faible que celle obtenue sur les $P\&L^{CH}$ sauf pour le test multinomial d'Hurlin et Tokpavi (2007). Par ailleurs, l'impact sur les tests fondés sur l'occurence des violations est moins uniforme. Le test de Christoffersen au seuil de risque de 1% et jusqu'à une taille d'échantillon de 750 à une taille empirique plus faible avec les données contaminées. Toutefois, cette taille

TAB. 12 – Tailles empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2\,(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0302	0.0881	0.1825	0.0929	0.0617	0.0655	0,0018	0.0153	0.1173	0.0431
	(0.0205)	(0.0670)	(0.1363)	(0.0577)	(0.0505)	(0.0743)	(0,0006)	(0.0305)	(0.1246)	(0.0355)
$T = 500$	0.0214	0.0880	0.1620	0.0705	0.0827	0.0643	0,0071	0.0303	0.1294	0.0376
	(0.0146)	(0.0941)	(0.1152)	(0.0429)	(0.0983)	(0.1102)	(0,0117)	(0.0762)	(0.1429)	(0.0577)
$T = 750$	0.0246	0.1001	0.1251	0.0572	0.0863	0.0704	0,0199	0.0298	0.1115	0.0489
	(0.0159)	(0.1473)	(0.0987)	(0.0372)	(0.1374)	(0.1371)	(0,0390)	(0.0933)	(0.1271)	(0.0985)
$T = 1000$	0.0373	0.1087	0.1043	0.0564	0.0827	0.0836	0,0271	0.0347	0.1066	0.0550
	(0.0493)	(0.1861)	(0.0846)	(0.0449)	(0.1548)	(0.1759)	(0,0599)	(0.1140)	(0.1222)	(0.1380)
$T = 1250$	0.0372	0.1345	0.0942	0.0512	0.0788	0.0805	0,0317	0.0355	0.1002	0.0545
	(0.0466)	(0.2449)	(0.0758)	(0.0496)	(0.1510)	(0.2028)	(0,0756)	(0.1341)	(0.1232)	(0.1631)
$T = 1500$	0.0392	0.1124	0.0849	0.0535	0.0716	0.0973	0,0324	0.0341	0.1027	0.0506
	(0.0588)	(0.2397)	(0.0681)	(0.0612)	(0.1484)	(0.2357)	(0,0851)	(0.1482)	(0.1202)	(0.1898)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle MDH. A chaque date la VaR correspondante est calculée avec le même modèle MDH et satisfait l'hypothèse de couverture conditionnelle. La taille empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la taille empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la taille empirique calculée à partir des rendements contaminées par les frais de gestion et les commissions. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2\,(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

TAB. 13 – Puissances empiriques des tests de backtesting

	LR_{CC}		DQ_{CC}		$LR_{CC}^{durée}$		$J_{CC}(5)$		$Q_2(5)$	LR_{MUC_2}
$\alpha =$	1%	5%	1%	5%	1%	5%	1%	5%	—	—
$T = 250$	0.0524	0.1238	0.3306	0.2917	0.0513	0.0311	0.0107	0.0346	0.3550	0.0215
	(0.0455)	(0.1048)	(0.2938)	(0.2318)	(0.0458)	(0.0399)	(0.0072)	(0.0473)	(0.3563)	(0.0165)
$T = 500$	0.0411	0.1301	0.3948	0.3181	0.0779	0.0287	0.0155	0.0556	0.4569	0.0044
	(0.0386)	(0.1355)	(0.3460)	(0.2588)	(0.0885)	(0.0514)	(0.0137)	(0.0915)	(0.4533)	(0.0107)
$T = 750$	0.0623	0.1514	0.4524	0.3945	0.0702	0.0283	0.0208	0.0679	0.4968	0.0026
	(0.0609)	(0.1660)	(0.4046)	(0.3325)	(0.0861)	(0.0606)	(0.0245)	(0.1209)	(0.5045)	(0.0138)
$T = 1000$	0.0779	0.1729	0.4840	0.4771	0.0618	0.0355	0.0256	0.0826	0.5509	0.0016
	(0.0690)	(0.1985)	(0.4464)	(0.4146)	(0.0808)	(0.0748)	(0.0293)	(0.1489)	(0.5660)	(0.0101)
$T = 1250$	0.1060	0.1970	0.4930	0.5617	0.0652	0.0412	0.0263	0.0944	0.5806	0.0010
	(0.0956)	(0.2393)	(0.4458)	(0.5100)	(0.0818)	(0.0985)	(0.0326)	(0.1767)	(0.5991)	(0.0152)
$T = 1500$	0.1354	0.2214	0.5287	0.6438	0.0641	0.0449	0.0272	0.0998	0.6209	0.0007
	(0.1071)	(0.2782)	(0.4679)	(0.5890)	(0.0750)	(0.1215)	(0.0348)	(0.2048)	(0.6265)	(0.0171)

Notes : pour chaque simulation, la distribution des pertes et profits du portefeuille est générée par un modèle MDH. A chaque date la VaR HS est simplement le quantile non conditionnelle des 250 précédentes observations journalières des P&L. La puissance empirique des tests correspond à la fréquence de rejet de l'hypothèse nulle obtenue avec $10,000$ simulations. T représente la taille de l'échantillon de la VaR. La taille nominale est de 5%. Le chiffre du haut représente la puissance empirique calculée à partir des rendements rétrospectifs, donc non pollués. Tandis que le chiffre entre parenthèse représente la puissance empirique calculée à partir des rendements contaminés par les frais de gestion et les commissions. La statistique de test de Christoffersen est abréviée par LR_{CC}, celle de Engle et Manganelli par DQ_{CC}, celle de Christoffersen et Pelletier par $LR_{CC}^{durée}$, celle de Candelon et consorts par $J_{CC}(5)$, celle d'Hurlin et Tokpavi par $Q_2(5)$ tandis que celle de Pérignon et Smith l'est par LR_{MUC_2}.

empirique devient plus élevée que celle obtenue avec les données propres pour les tailles d'échantillon supérieures ou égales à 1000. Ce test semble devenir *well sized* à 1% avec les $P\&L^{F\&C}$. Au niveau de risque de 5%, l'écart de taille empirique devient de plus en plus ample avec la taille d'échantillon. Le test devient alors largement *oversized*. Enfin, le test de Engle et Manganelli (2004) est le seul à réagir différemment. Les données contaminées sont ici associées à une taille empirique quasi systématiquement plus faible qu'avec les données non polluées réduisant ainsi le fait que ce test soit *oversized*. Ce test tendrait alors à moins rejeter un modèle interne de la banque si l'on utilise des données contaminées par les frais de gestion et les commissions afin de le valider.

Le tableau 13 regroupe les puissances empiriques des tests de *backtesting* obtenues sur les $P\&L^{CH}$ et les $P\&L^{F\&C}$. Les prévisions de la VaR utilisées sont de type *Historical Simulation*. Les tests doivent donc détecter ce mauvais modèle.

Nos résultats mettent ici en évidence une séparation des tests de validation en deux groupes. Les tests basés sur les durées et les tests multinomiaux peuvent être rapprochés même si l'ampleur de la puissance empirique de ces tests est bien différente. En effet, la statistique $J_{CC}(5)$ est de nouveau peu puissante alors que l'article de référence met en avant le fait que cette statistique soit plus puissante que les autres tests. Mais concernant les comparaisons entre les différents types de données, il s'avère que la puissance est plus élevée avec les $P\&L^{F\&C}$ qu'avec les données propres. Ce qui signifie que l'on rejette plus souvent le modèle interne erroné avec ce type de données polluées. Toutefois, ces différences sont très marginales. A noter que le test multinomial de Hurlin et Tokpavi (2007) est très puissant dans nos simulations. Le plus faible nombre de violations obtenu avec les données contaminées ne semble pas affecter le test de couverture non conditionnelle de Pérignon et Smith (2008) puisque sa puissance reste très faible. Cependant, comme avec les tailles empiriques, ce groupe de tests voit sa puissance empirique aller à contre courant des conclusions énoncées précédemment au niveau de risque de 1% et pour une taille d'échantillon de 250. Dans cette situation particulière presque tous les tests ont alors une puissance plus petite avec les données contaminées qu'avec les données propres. La différence la plus flagrante est à mettre à l'actif du second groupe, c'est-à-dire celui des tests fondés sur la fréquence des violations, et en particulier du test de Engle et Manganelli (2004). La contamination tend alors à réduire le taux de rejet de l'hypothèse nulle alors qu'elle devrait l'être. Seul le test LR_{CC} au seuil de risque de 5% voit sa puissance croître avec les données polluées.

Par conséquent, il est évident que des déformations de taille très importantes apparaissent lorsque l'on utilise ce type de contamination. L'attente d'un sous-rejet des prévisions de la VaR est confirmée dans la configuration demandée par la réglementation et cela pour tous les tests à l'exception du test de Hurlin et Tokpavi (2007). Mais en présence de données

contaminées, plus le taux de couverture augmente, plus la fréquence de rejet de l'hypothèse nulle grimpe. Il en est de même lorsque la taille de l'échantillon augmente. Cela signifie que le décalage vers la droite de la distribution des données de $P\&L_t^{F\&C}$ par rapport aux données de $P\&L_t^{CH}$ a des conséquences sur les résultats de *backtesting* en augmentant la fréquence de rejet sur grand échantillon, à l'exception du test de Engle et Manganelli (2004), mais en la diminuant sur petit échantillon et en particulier au seuil réglementaire. Les distorsions de puissances empiriques sont très diverses, mais comme pour la taille empirique, la puissance des tests de validation devient encore plus faible en présence de données contaminées dans la configuration imposée par le régulateur. Le décalage certain vers la droite de la distribution des $P\&L_t^{F\&C}$, aussi léger soit-il, est sans aucun doute beaucoup plus marqué sur grand échantillon. Ainsi, les tests de *backtesting* peuvent mieux détecter ce phénomène et *in fine* rejeter l'hypothèse nulle. Au contraire, sur petit échantillon et avec un faible seuil de risque, le décalage de la distribution n'est pas assez prononcé pour que les tests de *backtesting* s'en rendent compte.

5 Conclusion

Cette étude des déformations des tailles et des puissances empiriques des principaux tests de *backtesting* met clairement en évidence la sensibilité de ces procédures concernant le risque de données. Elle renforce l'idée selon laquelle les banques ont un avantage à valider leur modèle interne de risque de marché en employant des données contaminées. Bien que déconseillée par le régulateur, cette pratique courante reste tolérée. Jusqu'ici, connaître avec précision l'impact de données contaminées sur ces procédures demeurait ambigu. Parce qu'elle s'appuie justement sur cet enjeu inédit, notre étude empirique démontre que les données polluées, par les rendements historiques ou les frais de gestion et les commissions (ces derniers ont le plus d'impacts dans notre article car ils décalent de manière significative la distribution des P&L vers la droite), dévient de manière non négligeable les résultats du *backtesting* en sur-validant de bons modèles mais aussi en validant davantage de modèles qui ne devraient l'être. En effet, les tailles empiriques calculées sur les données contaminées sont bien souvent plus faibles que celles obtenues sur les données propres, en particulier dans la configuration imposée par le régulateur. Ceci traduirait le fait que les tests de backtesting tendraient à être moins *oversized* lorsque l'on utilise des données contaminées. Il en est de même pour les puissances, c'est justement ce cas qui est très avantageux pour les banques. En effet, la probabilité qu'elles valident leur modèle interne alors que ce dernier n'est pas adéquat est plus élevée. Toutefois, chaque catégorie de tests a une déformation qui lui est propre tant au niveau de l'ampleur que du sens de la déformation asymptotiquement. Les

plus déformés sont les tests fondés sur la fréquence des violations, viennent ensuite les tests multinomiaux tandis que les tests basés sur les durées résistent le mieux à ce type de contamination puisque leurs tailles et puissances empiriques sont plus élevées en présence de ces données polluées. Ils sont donc plus à même de détecter ce type de contamination et ainsi rejeter plus souvent les modèles internes car ils sont artificiellement validés par l'emploi de ce type de données. Il ne fait pour nous aucun doute qu'il en serait de même pour les données contaminées par les revenus de l'*intraday trading* si notre modélisation était plus efficiente.

Puisque les prévisions de la VaR se trouvent de ce fait moins souvent rejetées par les tests de validation, cela signifie que le modèle interne développé par les banques n'est pas rejeté, alors qu'il aurait pu l'être si les banques avaient employé des données non contaminées. Cette technique possède l'avantage de permettre à la banque de détenir moins de fonds propres que ce qu'elle devrait, si elle suivait scrupuleusement la réglementation. Il n'est donc pas surprenant que cette pratique se soit accentuée au cours de la dernière crise puisqu'elle permet de limiter l'appréciation du capital réglementaire au titre du risque de marché, et ainsi limiter l'effet procyclique des accords de Bâle II qui demande plus de capitaux en période de crise alors que c'est justement à ce moment là qu'il est le plus difficile d'en conserver. Cependant, une inadéquation de ce capital minimum par rapport au risque de marché peut avoir des conséquences dramatiques sur la sphère financière toute entière, et *in fine* sur la sphère réelle. De plus, il faut porter une attention toute particulière aux résultats observés dans le cadre imposé par la réglementation, c'est-à-dire au seuil de risque de 1% et une taille d'échantillon de 250 observations. Sous ces conditions, la quasi totalité des tests de *backtesting* ont une puissance et une taille empiriques plus faibles avec les données polluées qu'avec les données propres, or le comportement asymptotique de certaines de ces procédures va dans le sens inverse. Cette méthodologie réglementaire est vraisemblablement trop étroite puisqu'elle ne permet pas d'apprécier un modèle de risque de marché sur un long horizon permettant de prendre en compte davantage de violations et ainsi apporter plus d'informations aux tests de validation.

Pour conclure, cette étude prouve que la qualité des données utilisées pour le *backtesting* est primordiale puisque des données contaminées permettent de valider plus souvent les modèles de prévisions de la VaR. Désormais, cette qualité de données se pose donc comme un enjeu fondamental. Encore plus que le risque de modèle et d'estimation qui ont un faible impact sur les conclusions des tests de *backtesting*, le choix des données devrait donc susciter les préoccupations de toutes les parties : les banques qui devront tenir compte de ce risque de données, et le régulateur qui devra être plus attentif à cette pratique.

Références

[1] ANDERSEN, T. G. (1996), "Return Volatility and Trading Volume : An Information Flow Interpretation of Stochastic Volatility", *The Journal of Finance*, Vol. 51, N° 1, p. 169-204.

[2] BAUWENS, L., LAURENT, S., et ROMBOUTS, J.V.K (2003), "Multivariate Garch Models : A Survey", Core Discussion Paper.

[3] BASEL COMMITTEE ON BANKING SUPERVISION (1996), "Supervisory Framework for the Use of Backtesting in Conjunction with the Internal Models Approach to Market Risk Capital Requirements, Bank for International Settlements".

[4] BONTEMPS, C. (2006), "Moment-based tests for discrete distributions", Working paper.

[5] BONTEMPS, C., et MEDDAHI, N. (2005), "Testing normality : A GMM approach", *Journal of Econometrics*, 124, p. 149-186.

[6] BONTEMPS, C., et MEDDAHI, N. (2006), "Testing distributional assumptions : A GMM approach", Working paper.

[7] CAMPBELL, S. D. (2005), "A Review of Backtesting and Backtesting Procedures", *Finance and Economics Discussion Series Divisions of Research & Statistics and Monetary Affairs Federal Reserve Board*, Washington, D.C.

[8] CANDELON, B., COLLETAZ, G., HURLIN, C., et TOKPAVI, S. (2010), "Backtesting Value-at-Risk : a GMM duration-based test", *Journal of Financial Econometrics*, p. 1-30.

[9] CHRISTOFFERSEN, P. F. (1998), "Evaluating interval forecasts", *International Economic Review*, 39, p. 841-862.

[10] CHRISTOFFERSEN, P. F., et PELLETIER, D. (2004), "Backtesting Value-at-Risk : A Duration-Based Approach", *Journal of Financial Econometrics*, 2, 1, p. 84-108.

[11] DOWD, K. (2005), *Measuring Market Risk, Second Edition*, Wiley Finance.

[12] DROST, F. C., et NIJMAN, T. E. (1993), "Temporal Aggregation of Garch Processes", *The Econometrics Society*, Vol. 61, 4, p. 909-927.

[13] ENGLE, R. F. (2000), "The Econometrics of Ultra-High-Frequency Data", *The Econometric Society*, Vol. 68, 1, p. 1-22.

[14] ENGLE, R. F., et KRONER, K. F. (1995), "Multivariate simultaneous generalized ARCH", *Econometric Theory*, 11, p. 122-150.

[15] ENGLE, R. F., et MANGANELLI, S. (2004), "CAViaR : Conditional autoregressive Value-at-Risk by regression quantiles", *Journal of Business and Economic Statistics* 22, p. 367-381.

[16] ESCANCIANO, J. C. and OLMO J. (2008), "Robust Backtesting Tests for Value-at-Risk Models", Working Paper, Dept. Economics, Indiana University.

[17] ESCANCIANO, J. C. and OLMO J. (2009), "Backtesting Parametric Value-at-Risk with Estimation Risk", Center for Applied Economics and Policy Research, 28 Working Paper, forthcoming in *Journal of Business and Economic Statistics*.

[18] FRANCQ, C., et ZAKOIAN, J. M. (2009), *Modèles GARCH, Structure, Inférence Statistique et Applications Financières*, Ed. Economica, collection « économie et statistique avancées ».

[19] FRESARD, L., PERIGNON, C., et WILHELMSOM, A. (2010), "The Pernicious Effects of Contaminated Data in Risk Management", Working paper.

[20] GIOT, P. (1999), "Time transformations, intraday data and volatility models", Core Discussion Paper.

[21] GUILLAUME, D. M., PICTET, O. V., et DACOROGNA, M. M. (1995), "On the intra-daily performance of GARCH processes".

[22] HARRIS, L. (1986), "Cross-security tests of the mixture of distributions hypothesis", *Journal of Financial and Quantitative Analysis* 21, p. 39-46.

[23] HARRIS, L. (1987), "Transaction data tests of the mixture of distributions hypothesis", *Journal of Financial and Quantitative Analysis* 22, p. 127-141.

[24] HENDRICKS, D., et HIRTLE, B. (1997), "Bank Capital Requirements for Market Risk : The Internal Models Approach", *Federal Reserve Bank of New York Economics Policy Review*, December, p. 1-12.

[25] HIRTLE, B. (2003), "What Market Risk Capital Reporting Tells us about Bank Risk", *Federal Reserve Bank of New York Economics Policy Review*, September, p. 37-54.

[26] HURLIN, C. et TOKPAVI, S. (2007), "Bactesting Value-at-Risk Accuracy : A Simple New Test", *Journal of Risk*, Vol. 9, 2, p. 19-37.

[27] HURLIN, C. et TOKPAVI, S. (2008), "Une Evaluation des Procédures de Backtesting : Tout va pour le Mieux dans le Meilleur des Mondes", *Finance*, vol 29(1), p. 53-80.

[28] KUPIEC, P. H. (1995), "Techniques for Verifying the Accuracy of Risk Measurement Models", *Journal of Derivatives*, 3, p. 73-84.

[29] PERIGNON, C., et SMITH, D. R. (2008), "A New Approach to Comparing VaR Estimation Methods", *Journal of Derivatives*.

[30] STULZ, R. (2009), "Six Ways Companies Mismanage Risk", *Harvard Business Review*, 87, p. 86-94.

[31] TSE, Y. K. (2000), "A test for constant correlations in a multivariate GARCH model", *Journal of Econometrics*, p. 107-127.

FIG. 1 – P&L rétrospectifs et historiques.

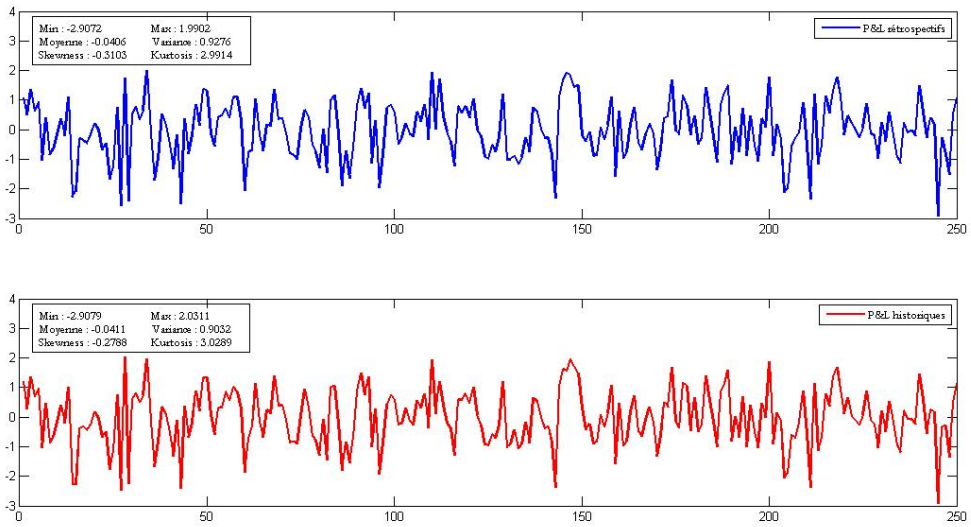


FIG. 2 – P&L rétrospectifs et prévisions de la VaR avec sa séquence de violations.

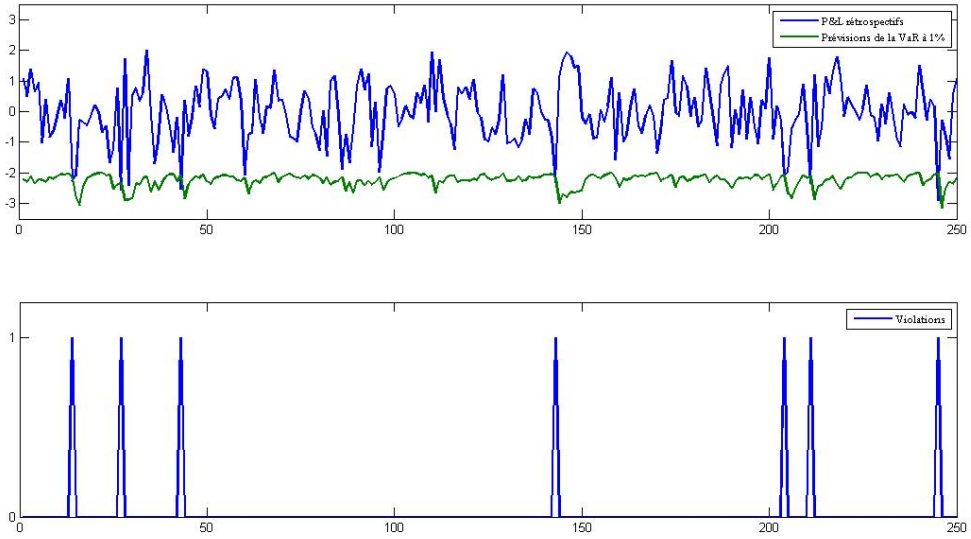


FIG. 3 – P&L historiques et prévisions de la VaR avec sa séquence de violations.

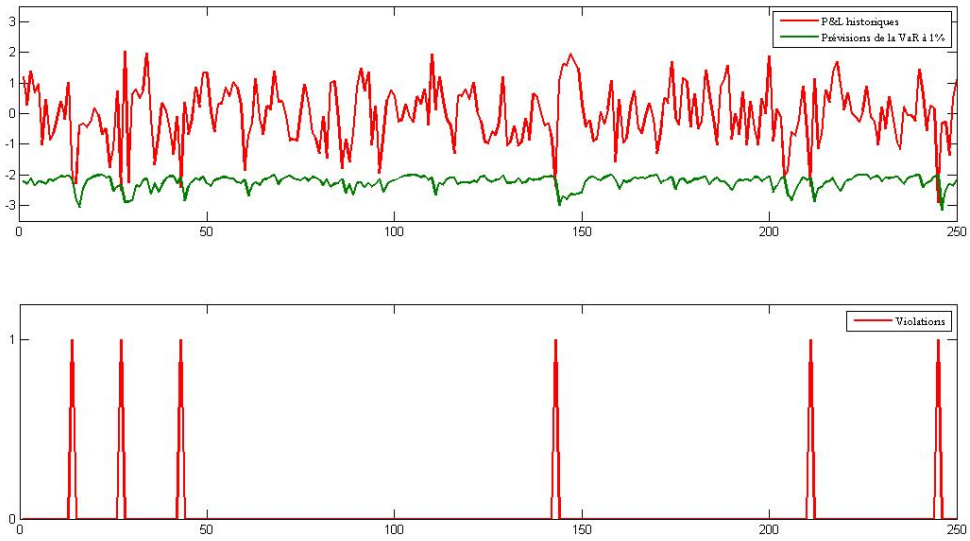


FIG. 4 – Evolution intraday du prix du portefeuille sur 5 jours avec ses moyennes mobiles et évolution du montant journalier de la pollution intraday sur 250 jours.

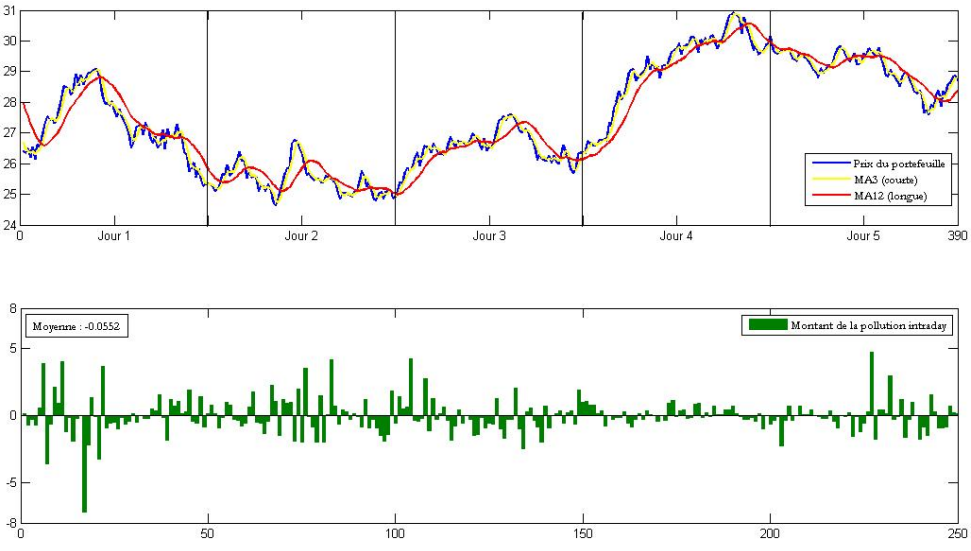


FIG. 5 – P&L non contaminés et prévisions approximées de la VaR avec sa séquence de violations.

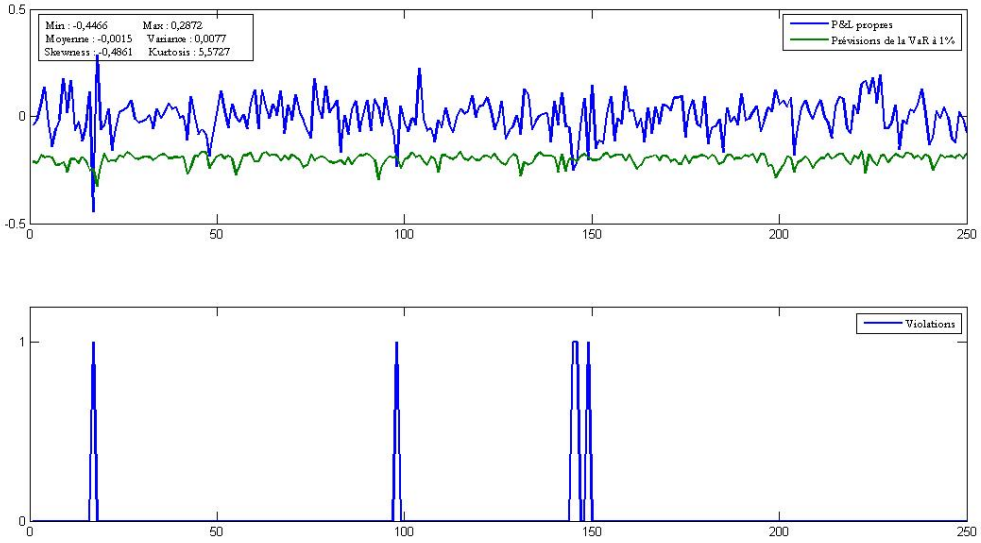


FIG. 6 – P&L contaminés par la pollution intraday et prévisions approximées de la VaR avec sa séquence de violations.

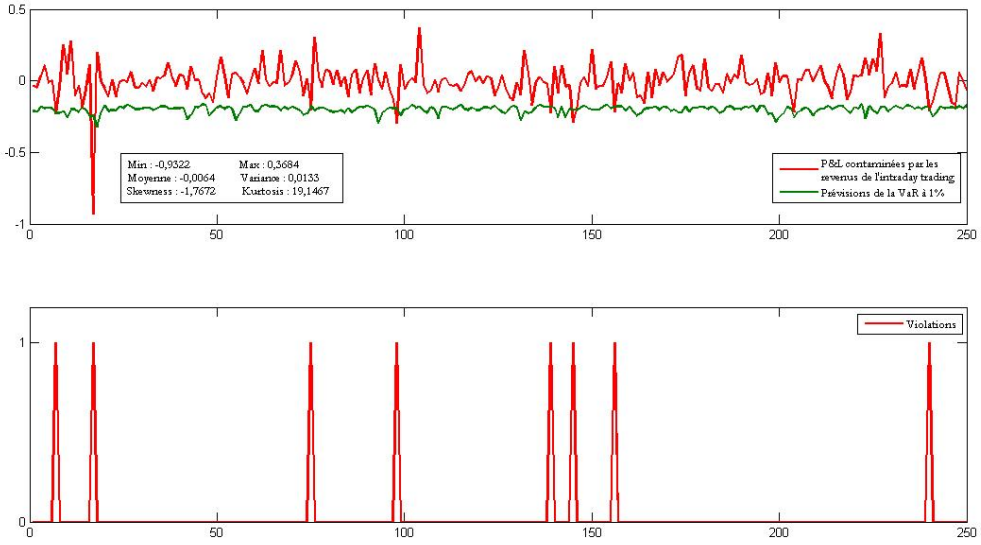


FIG. 7 – P&L non contaminés et contaminés par les frais de gestion et les commissions avec les prévisions de la VaR.

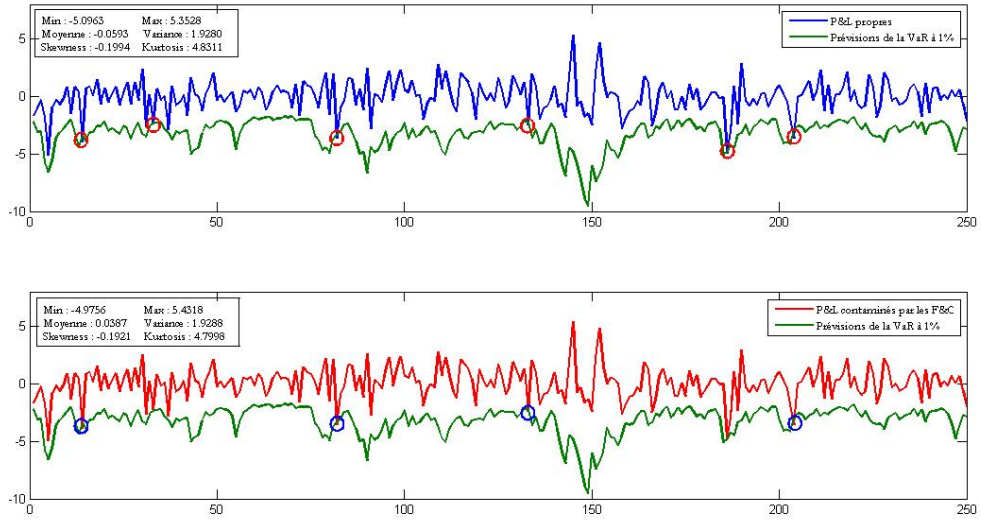


FIG. 8 – Volume quotidien détrendé et montant de pollution associé.