MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE
ET DE LA COHÉSION
DES TERRITOIRES
Liberté
Égalité
Fraternité

NOTE
MÉTHODOLOGIQUE

Estimer une concentration moyenne en présence de données multi-censurées

Juin 2023

Résumé

Les jeux de données relatifs aux concentrations de substances chimiques dans l'eau comportent souvent des observations non quantifiées dont la valeur est inférieure à la limite de quantification du laboratoire d'analyse. Ces données, dites censurées, sont utilisées pour estimer la concentration moyenne d'une substance au même titre que les valeurs quantifiées.

Cette note méthodologique propose un nouvel estimateur de moyenne pour les jeux de données censurées. Nommé ESTIM, cet estimateur s'appuie sur une analyse de la distribution des valeurs observées pour prendre en compte les valeurs non quantifiées. Comparé à l'estimateur LQ2 régulièrement utilisé dans le domaine de l'eau, ESTIM se montre plus performant et mieux adapté aux jeux de données fortement censurés.

Une méthode est également proposée pour redresser les changements de limites de quantification dans le temps ; elle se montre efficace sur les estimateurs LQ2 et ESTIM.

Auteur

Didier EUMONT - SDES

Sommaire

1. Contexte et objectif de l'étude

1.1. Les données environnementales sont fortement censurées

La concentration d'une substance chimique recherchée dans l'environnement est parfois si petite que la méthode d'analyse du laboratoire ne peut pas la déterminer avec exactitude ; le résultat de l'analyse s'exprime alors par une donnée qualitative du type « inférieur à » la limite de quantification de l'appareil de mesure (exemple : « < 0,1 g/l »). La donnée est dite « censurée à gauche » au sens où la concentration exacte de la substance est inconnue et se situe entre zéro et la limite de quantification (LQ) spécifiée.

Quand une substance chimique est recherchée par des laboratoires utilisant des méthodes d'analyse distinctes, les données peuvent comporter plusieurs limites de quantification et devenir « multi-censurées ». L'ensemble des résultats {1, 4, < 3, < 3, 2, 1, < 2, < 2, < 5, 5, 4}, où le signe « < » précède diverses limites de quantification, est un exemple de données multi-censurées.

Selon la substance considérée, les résultats d'analyses peuvent comporter plus de 75 % de données censurées et une grande variabilité des limites de quantification dans le temps (les valeurs de LQ tendent à diminuer dans le temps à mesure que les performances analytiques s'améliorent).

Sachant qu'un estimateur de concentration moyenne peut influencer des décisions importantes (mesures de gestion de substances écotoxiques ou de pollutions environnementales), la qualité d'un tel estimateur se pose sur des données fortement censurées et avec des LQ évolutives dans le temps. Un estimateur de bonne qualité et robuste favorise l'obtention de résultats crédibles (statistiques inférentielles, modèles de régression, tendances temporelles…).

1.2. Les estimateurs de moyenne disponibles ne sont pas performants dans toutes les situations

Dans le domaine de l'eau, les concentrations moyennes sont régulièrement estimées par la moyenne arithmétique des valeurs quantifiées et non quantifiées, après substitution des valeurs censurées par une constante arbitraire (souvent LQ/2 ou LQ/ $\sqrt{2}$). Bien que simple à calculer, la littérature scientifique souligne que cet estimateur produit des résultats peu fiables, notamment lorsque les taux de censure sont élevés [1, 2, 3, 4, 5, 9, 10].

Il est possible d'estimer la moyenne de données censurées par des méthodes alternatives, comme la méthode de Kaplan-Meier, la régression robuste sur statistique ordonnée (rROS) ou la méthode du maximum de vraisemblance sous l'hypothèse d'une distribution sous-jacente connue. Leurs performances dépendent du pourcentage de données censurées, de la taille des échantillons et de la forme de la distribution des données. Par exemple, telle méthode donne de meilleures estimations sur les petits échantillons que sur les grands. Telle autre est fiable jusqu'à 50 % de données censurées et telle autre jusqu'à 75 %. Toutefois, aucun de ces estimateurs ne se montre supérieur à un autre dans toutes les situations, qu'il soit fondé sur une méthode paramétrique ou non paramétrique [2,4,5,6,7].

1.3. Proposition d'un algorithme estimant la moyenne des distributions de données

La présente publication propose un algorithme produisant des estimations de moyenne de bonne qualité sur données multi-censurées à gauche, observables dans le domaine des eaux de surface.

Il est fondé sur l'hypothèse selon laquelle les valeurs non quantifiées (< LQ) suivent la même distribution que les valeurs quantifiées. Il combine plusieurs méthodes d'estimation de moyenne, paramétrique et non paramétrique, lui permettant de s'adapter à la forme de la distribution, à la taille de l'échantillon de données et à son taux de censure. Selon les caractéristiques du jeu de données, l'algorithme produit une estimation de moyenne de distribution par la méthode de Kaplan-Meier, par rROS, ou par maximum de vraisemblance car ces dernières sont reconnues efficaces dans des conditions extrêmes.

La souplesse d'un tel « estimateur composite » peut être intéressante notamment pour l'étude de séries chronologiques et les regroupements de données issues de zones géographiques distinctes. Par exemple, sur de grands espaces, les données acquises à une période P1 peuvent provenir de centaines de points de mesures et comporter autant de limites de quantification distinctes qu'il y a de laboratoires d'analyses. Et compte tenu de la diversité des pressions polluantes exercées sur ces zones de prélèvements, la distribution des données peut présenter une forme différente d'un point de mesure à l'autre (une moyenne identique mais de plus grands écarts-types donnant des formes plus aplaties). Puis, pour une période P2 sur un point de mesure donné, les formes de distribution peuvent changer suite à une modification de la pression polluante. Il en résulte un jeu de données aux caractéristiques évolutives, et un besoin d'estimateur adapté à des conditions changeant rapidement dans le temps et l'espace.

2. Les deux estimateurs de moyenne en comparaison

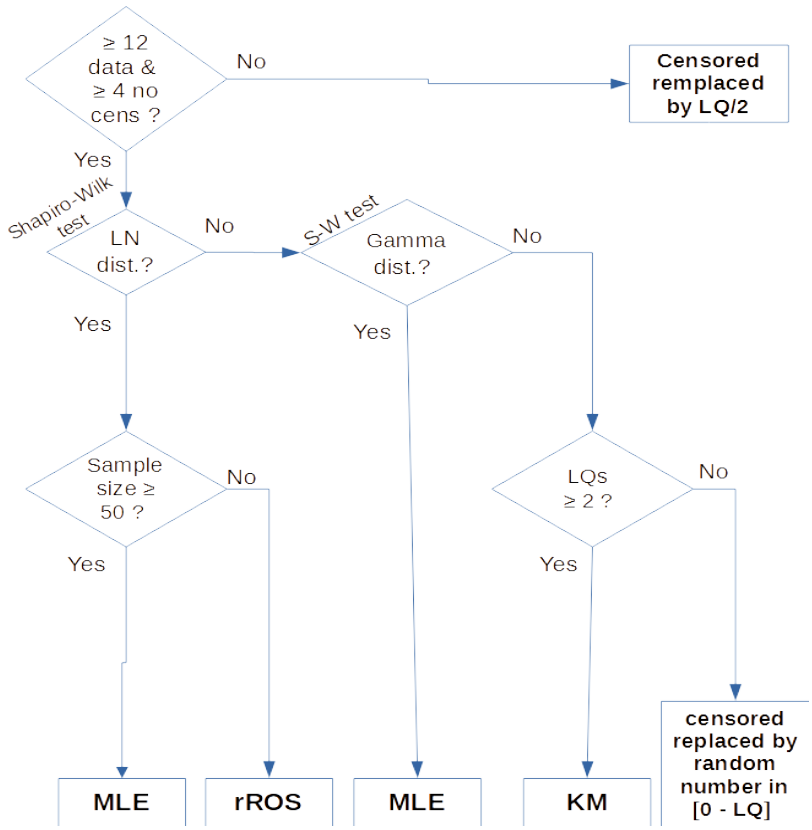
2.1. LQ2 : moyenne arithmétique des valeurs

Cet estimateur constitue la référence compte tenu de son utilisation régulière dans les publications sur la qualité des eaux (par exemple, dans celles relatives aux évaluations de l'état des eaux réalisées au titre de la directive-cadre sur l'eau). Il correspond à la moyenne arithmétique des valeurs quantifiées et non quantifiées, où ces dernières sont remplacées par la demi-valeur de la limite de quantification du laboratoire, $\frac{LQ}{2}$. La distribution réelle des valeurs n'est pas prise en compte.

2.2. ESTIM : moyenne estimée de la distribution des valeurs

L'algorithme ESTIM vise à ajuster une loi de distribution aux échantillons de données comportant au moins 12 valeurs, dont au moins 4 quantifiées, puis à estimer la moyenne de cette distribution ; il estime la moyenne par une méthode non paramétrique lorsqu'aucune distribution pertinente ne peut être ajustée *(figure 1)*.

Figure 1 : algorithme ESTIM



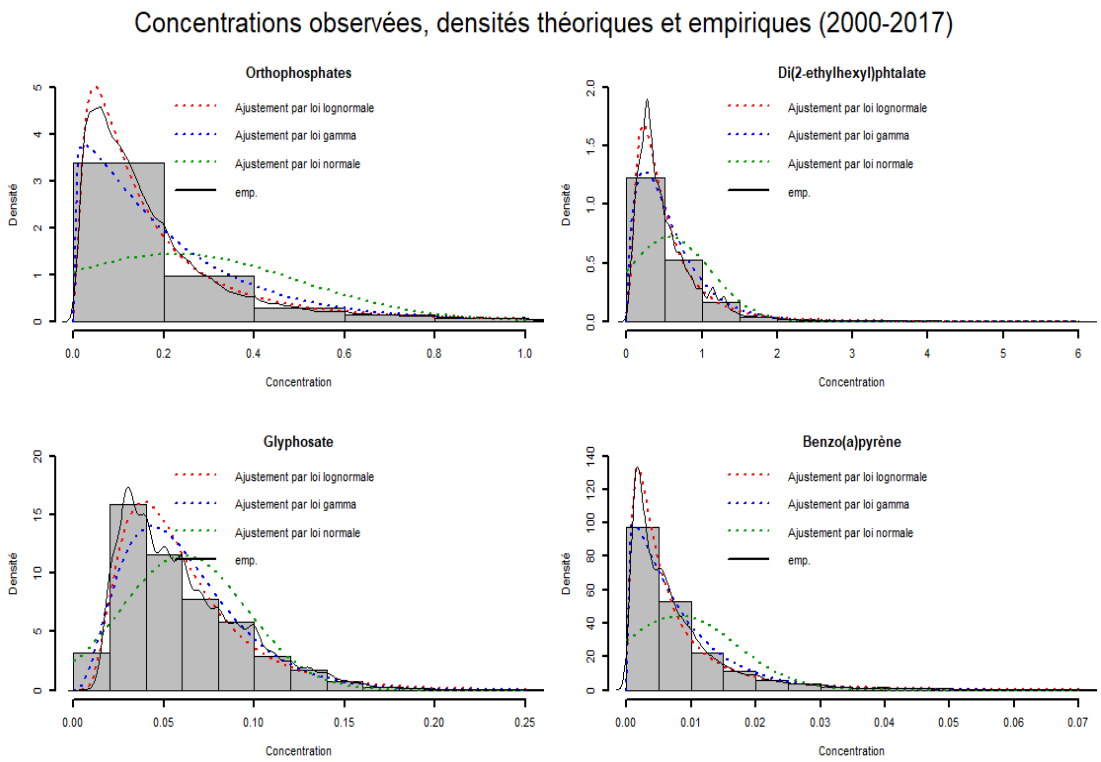
***Source** : SDES*

En présence de moins de 12 valeurs totales ou moins de 4 valeurs quantifiées, ESTIM renvoie la moyenne arithmétique calculée selon la méthode LQ2, l'échantillon étant considéré de taille insuffisante pour un test d'ajustement raisonnablement fiable.

Le test d'ajustement à une distribution candidate est réalisé sur les valeurs quantifiées, sous l'hypothèse que les valeurs non quantifiées (et non les limites de quantifications elles-mêmes) suivent la même distribution. Deux distributions candidates sont retenues parmi celles fréquemment observées sur les données de concentrations dans l'eau : la loi lognormale et la loi gamma *(figure 2)*.

Figure 2 : distributions des concentrations observées dans les cours d'eau et ajustement à des distributions théoriques

Concentration en µg/l



Note : la courbe en noir représente la densité de la distribution empirique.
***Source** : SDES*

L'ajustement à une loi lognormale ou gamma est évalué par le test de Shapiro-Wilk [6,7] et la sélection de l'une d'elles se fait d'après la p-valeur du test :

- Quand la p-valeur du test est favorable à une distribution lognormale et si l'échantillon comporte au moins 50 données, la moyenne est estimée par le maximum de vraisemblance (MLE). Si l'échantillon comporte moins de 50 données, la moyenne est estimée par la méthode des moments après imputation de valeurs issues d'une régression quantiles - quantiles *(voir méthodologie en annexe 1)*.

- Quand la p-valeur du test est favorable à une distribution gamma, la moyenne est estimée par le maximum de vraisemblance.

- Quand la p-valeur du test ne permet pas de sélectionner l'une des distributions candidates et si l'échantillon comporte au moins 2 limites de quantification (LQ) distinctes, la moyenne est estimée par la méthode non paramétrique de Kaplan-Meier *(voir méthodologie en annexe 1)*. Si l'échantillon comporte une LQ unique, la moyenne est calculée en remplaçant les valeurs censurées par un nombre tiré aléatoirement dans l'intervalle]0 - LQ].

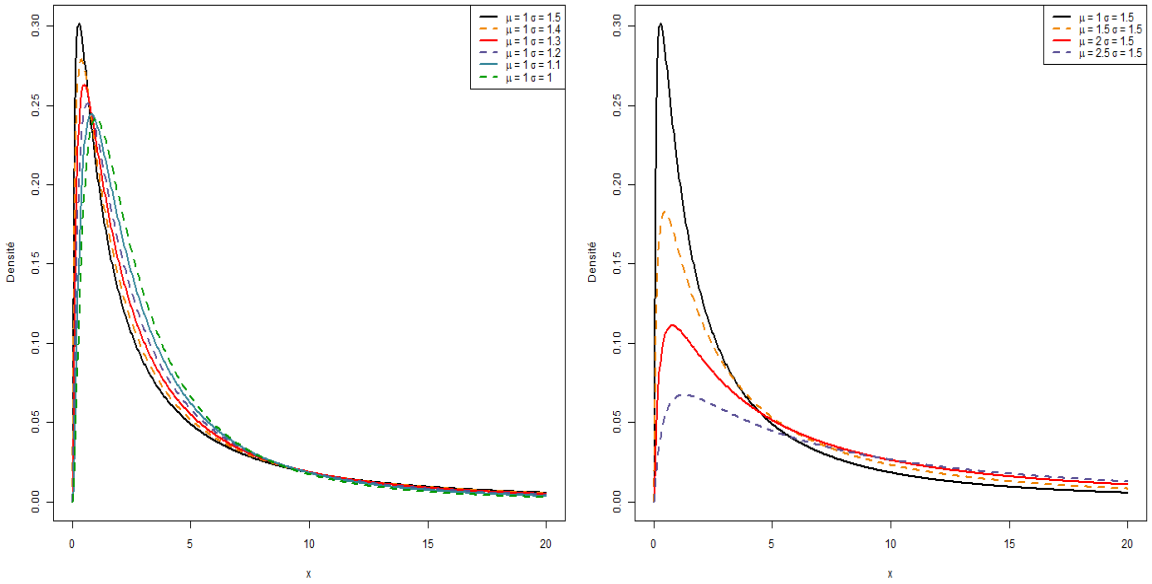
3. Méthode d'évaluation des propriétés des estimateurs LQ2 et ESTIM

3.1. Simulations de Monte-Carlo

La précision des estimateurs LQ2 et ESTIM est évaluée sur des jeux de données couvrant de larges gammes de tailles d'échantillons et de formes de distribution (fonctions de leurs moyennes et écarts-types) et de taux de censure. Chaque scénario fait l'objet de simulations de Monte-Carlo répliquées 2 000 fois.

Les tirages aléatoires d'échantillons composés de 20, 30, 50, 100, 150 ou 200 valeurs sont réalisés à partir d'une loi lognormale de paramètres connus, les données réelles suivant fréquemment cette distribution *(figure 1)*. Les paramètres de la loi lognormale couvrent la gamme $\mu = 1$ à $\mu = 2{,}5$ pour la moyenne et $\sigma = 1$ à $\sigma = 1{,}5$ pour l'écart-type *(figure 3)*.

Figure 3 : gamme des distributions lognormales utilisées pour les simulations



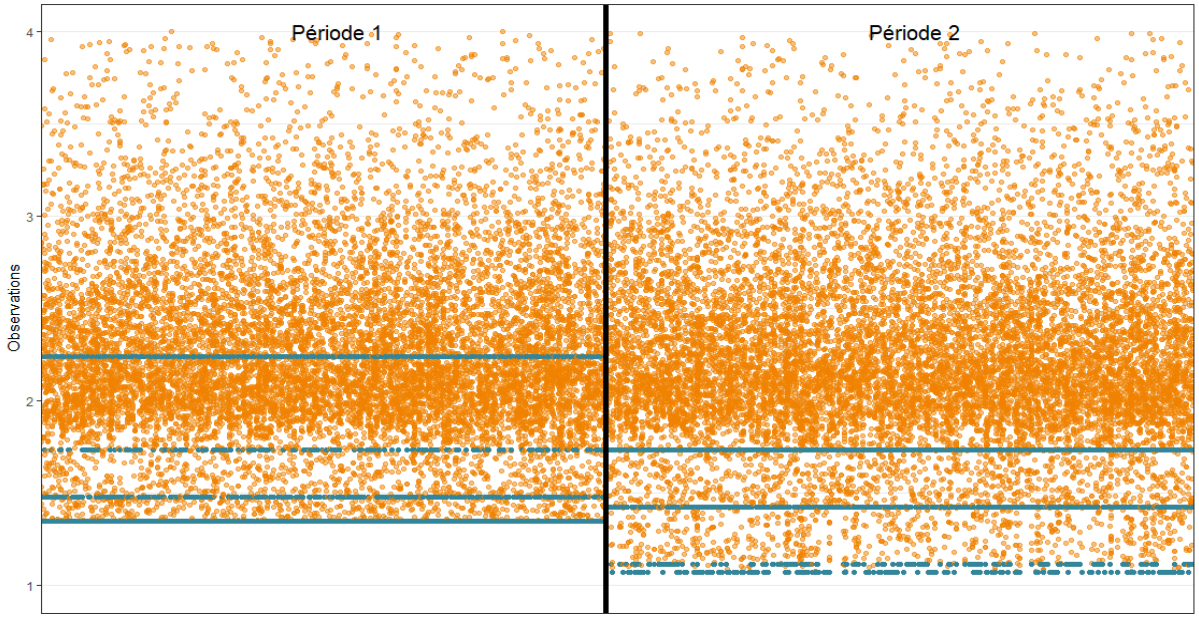
***Source** : SDES*

Les taux de censure cibles sont de 10 %, 25 %, 50 % et 75 % pour chaque tirage. Ils sont obtenus avec cinq valeurs de limites de quantification, chacune d'elles correspondant à un quantile distinct de la distribution générant les données. Les données tirées aléatoirement ayant une valeur inférieure ou égale à ces limites sont qualifiées "NQ" dans le jeu final.

Dans le cas d'une série chronologique, les limites de quantification peuvent changer entre deux périodes successives en raison de progrès techniques ou de changements de laboratoire ; les limites de la première année de la chronique, prise comme année de référence, sont souvent plus élevées que celles des années suivantes. Compte tenu des méthodes de calcul utilisées dans LQ2 et ESTIM, ces changements affecteront les estimations. En conséquence, des scénarios complémentaires sont élaborés afin d'étudier

l'impact de ces changements et tester une méthode visant à corriger les effets d'une variation de LQ dans le temps : une baisse de LQ entre deux périodes est simulée en constituant deux groupes de 1 000 échantillons, l'un comportant des échantillons censurés à des seuils 15 à 20 % plus bas que l'autre. Les échantillons comportent 100 valeurs tirées aléatoirement à partir d'une distribution lognormale de paramètres $\mu = 1{,}5$ et $\sigma = 0{,}6$, et sont censurés à 50 % ou 75 % selon quatre seuils *(figure 4)*.

Figure 4 : série censurée à 75 % avec baisse des seuils de censure



Note : 2 000 échantillons de taille n = 100 sont tirés aléatoirement d'une distribution lognormale de paramètres μ = 1,5 et σ = 0,6 constant sur les périodes 1 et 2. Les échantillons sont censurés à 75 % sur les deux périodes. Les seuils de censure, figurés en gris, baissent de 21,4 %.

***Source** : SDES*

3.2. Correction éventuelle des limites de quantification

Pour les séries chronologiques avec changements de LQ dans le temps, une méthode visant à corriger ou redresser ces variations est testée. La méthode consiste à modifier les valeurs de LQ d'une période (ou année) quelconque d'une chronique qu'il aurait été impossible d'observer si les performances analytiques étaient restées constantes dans le temps.

À titre d'exemple, on pose :

- {<0,1 ; <0,3 ; <0,5 ; <0,6 ; <0,7} : ensemble Y0 des LQ distinctes de l'année de référence ;
- {<0,05 ; 0, 07 ; 0,1 ; <0,1 ; <0,2 ; 0,32 ; <0,45 ; <0,5; 0,5 ; 0,6 ; 1} : ensemble Y1 des résultats observés l'année n (Y1 contient 5 valeurs non quantifiées sur 11, soit un taux de censure de 45 %).

Y1 ne contiendrait pas les valeurs {< 0,05 ; < 0,2 ; < 0,45} si les performances analytiques étaient restées constantes dans le temps, aucun laboratoire ne proposant ces niveaux de LQ l'année de référence. S'appuyant sur une relation d'ordre, ces trois valeurs sont remplacées

par des niveaux de LQ immédiatement supérieurs présents dans Y0 (une valeur inférieure à 0,05 est également inférieure à 0,1). On obtient

- {**<0,1** ; 0,07 ; 0,1 ; <0,1 ; **<0,3** ; 0,32 ; **<0,5** ; <0,5; 0,5 ; 0,6 ; 1} : ensemble Y2 des résultats observables l'année n à méthode d'analyse constante (censure = 45 %).

Y2 comporte désormais une valeur quantifiée à 0,07 alors que la LQ la plus basse est < 0,1. La valeur quantifiée devenue incohérente dans l'ensemble Y2 est écartée ; le taux de censure augmente. On obtient :

- {<0,1 ; ~~**0, 07**~~ ; 0,1 ; <0,1 ; <0,3 ; 0,32 ; <0,5 ; <0,5 ; 0,5 ; 0,6 ; 1} : ensemble Y3 des résultats observables l'année n à méthode d'analyse constante (censure = 50 %).

Pour retrouver le taux de censure initial, une des plus fortes valeurs de LQ est écartée (correspond à un résultat obtenu avec une méthode d'analyse peu performante) et Y3 devient :

- {<0,1 ; 0,1 ; <0,1 ; <0,3 ; 0,32 ; <0,5 ; ~~**<0,5**~~; 0,5 ; 0,6 ; 1} ensemble Y4 des résultats observables l'année n à méthode d'analyse constante (censure = 44 %).

3.3. Critères d'évaluation

La précision des estimateurs est évaluée par leur biais relatif, leur erreur quadratique moyenne relative et leur incertitude relative. La réaction des estimateurs aux changements de limites de quantification est évaluée par un calcul d'élasticité Moyenne-LQ.

3.3.1. Biais relatif (%)

Le biais est la différence moyenne entre la valeur de l'estimateur et la valeur vraie qu'il estime ; le biais relatif est le rapport entre le biais et la valeur vraie estimée.

$$\text{Biais relatif} = \frac{100\times(\bar{x}-\theta)}{\theta}$$

où θ est la valeur vraie à estimer et $\bar{x}$ est la moyenne des 2 000 estimations de cette valeur.

Le biais est positif quand l'estimateur surestime la valeur vraie, et négatif quand il la sous-estime.

3.3.2. Erreur quadratique moyenne relative (%)

L'erreur quadratique moyenne (EQM) combine le biais et la variance de l'estimateur.

$$\text{EQM relative} = \frac{100\times(variance+biais^2)}{\theta}$$

Toujours positive, EQM permet de classer les estimateurs lorsque l'un d'eux est sans biais et de variance modérée, alors que l'autre est biaisé mais de variance plus petite.

Un estimateur de bonne qualité présente à la fois un biais nul ou faible et une erreur quadratique moyenne minimale.

3.3.3. Incertitude relative (%)

L'incertitude relative (IR) se fonde sur la longueur de l'intervalle de confiance à 95 % obtenu par la méthode des quantiles (différence entre les 50e et 1 950e estimations classées par ordre croissant).

$$\text{IR} = \frac{0{,}5 \times IC95}{\bar{x}}$$

où IC95 est la longueur de l'intervalle de confiance et $\bar{x}$ est la moyenne des 2000 estimations.

Cet intervalle empirique ne tient pas compte du biais de $\bar{x}$.

3.3.4. Élasticité moyenne-LQ

Ce critère est utilisé pour mesurer la réaction d'un estimateur au changement de limite de quantification. La réaction de chaque estimateur est évaluée par l'élasticité moyenne-LQ entre deux périodes où le taux d'évolution des LQ est connu.

$$\text{Élasticité M-LQ} = \frac{\Delta\bar{x}(\%)}{\Delta lq(\%)}$$

où $\Delta\bar{x}(\%)$ est le taux de variation de la moyenne M estimée sur 1 000 échantillons et $\Delta lq(\%)$ est le taux de variation des limites de quantification.

L'efficacité de la méthode visant à corriger les variations de LQ est évaluée en comparant les élasticités M-LQ avant et après sa mise en œuvre.

4. Résultats

Les résultats sont présentés par classe de taille d'échantillons, d'abord pour ceux de taille modérée (30 valeurs) puis ceux de grandes taille (100 valeurs). Ils couvrent l'ensemble des combinaisons simulées de moyenne μ, d'écart-type σ et de taux de censure.

Les annexes 3 et 4 présentent la précision des estimations obtenues sur des échantillons de tailles différentes. L'annexe 2 présente les occurrences de méthodes d'estimation utilisées par ESTIM.

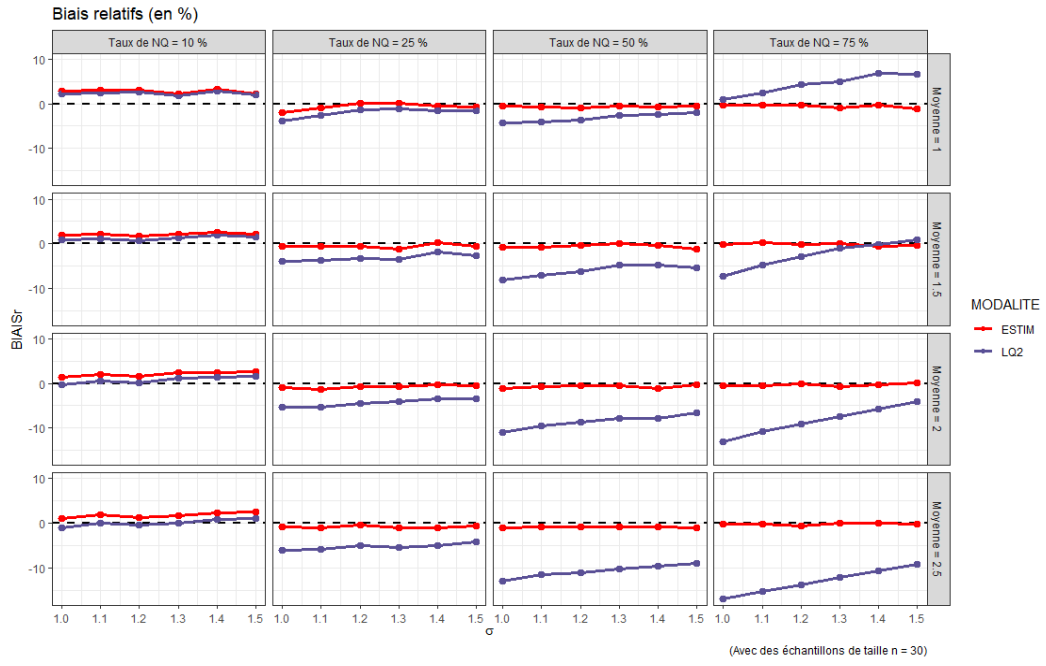
4.1. Qualité d'estimation sur échantillons de taille modérée (30 valeurs)

4.1.1. Biais relatifs

Pour un taux de censure bas (10 %), les biais relatifs de LQ2 et d'ESTIM sont similaires quelle que soit la forme de la distribution sous-jacente (fonction de sa moyenne μ et de son écart-type σ).

À partir de 25 % de censure, les écarts sont plus nets : le biais de LQ2 augmente à mesure que le taux de censure augmente et dépasse parfois 10 % ; celui d'ESTIM demeure inférieur à 1 % *(figure 5)*.

Figure 5 : biais relatifs sur échantillons de taille n = 30



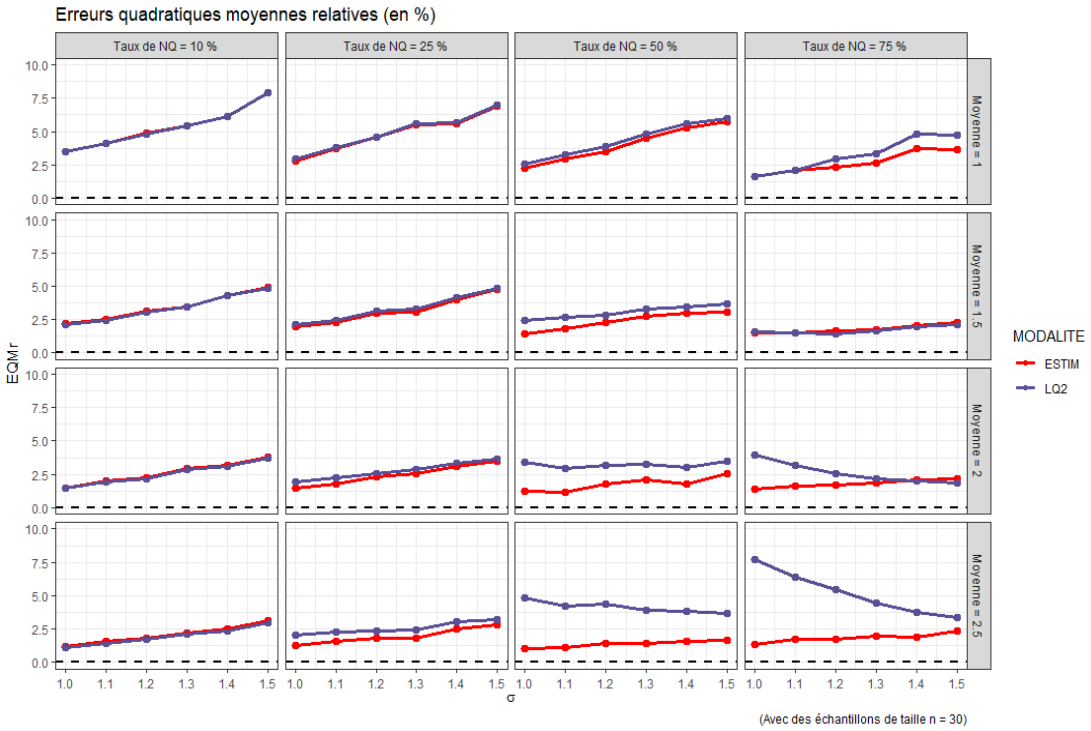
***Source** : SDES*

4.1.2. Erreurs quadratiques moyennes relatives (EQMR)

Les EQMr varient de 1 % à 8 % pour ESTIM et LQ2 sur l'ensemble des combinaisons simulées.

ESTIM présente des erreurs quadratiques quasi identiques à celles de LQ2 jusqu'à 25 % de censure, quelle que soit la forme de la distribution. Les écarts se creusent à partir de 50 % de censure, niveau au-delà duquel ESTIM se comporte bien mieux que LQ2 *(figure 6)*.

Figure 6 : EQM relatives sur des échantillons de taille n = 30



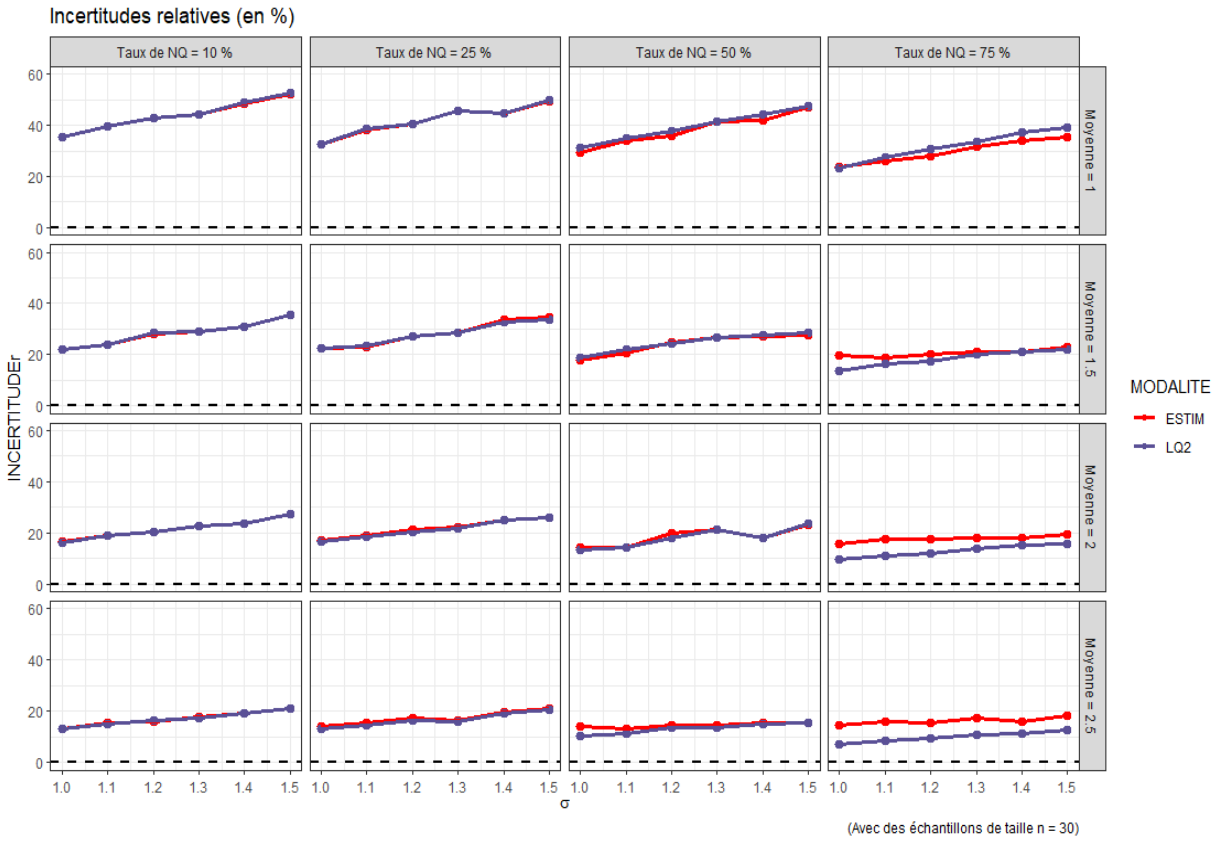
***Source** : SDES*

4.1.3. Incertitudes relatives

Globalement, les incertitudes de LQ2 et ESTIM sont comprises entre 5 % et 50 % et diminuent lorsque la distribution prend des formes plus aplaties (lorsque μ augmente).

Jusqu'à 50 % de censure, les deux estimateurs ne présentent pas de différence, quelle que soit la forme de la distribution. À 75 % de censure, ESTIM donne des intervalles de confiance un peu plus larges que LQ2 sur les distributions aux formes les plus aplaties *(figure 7)*.

Figure 7 : incertitudes relatives sur échantillons de taille n = 30



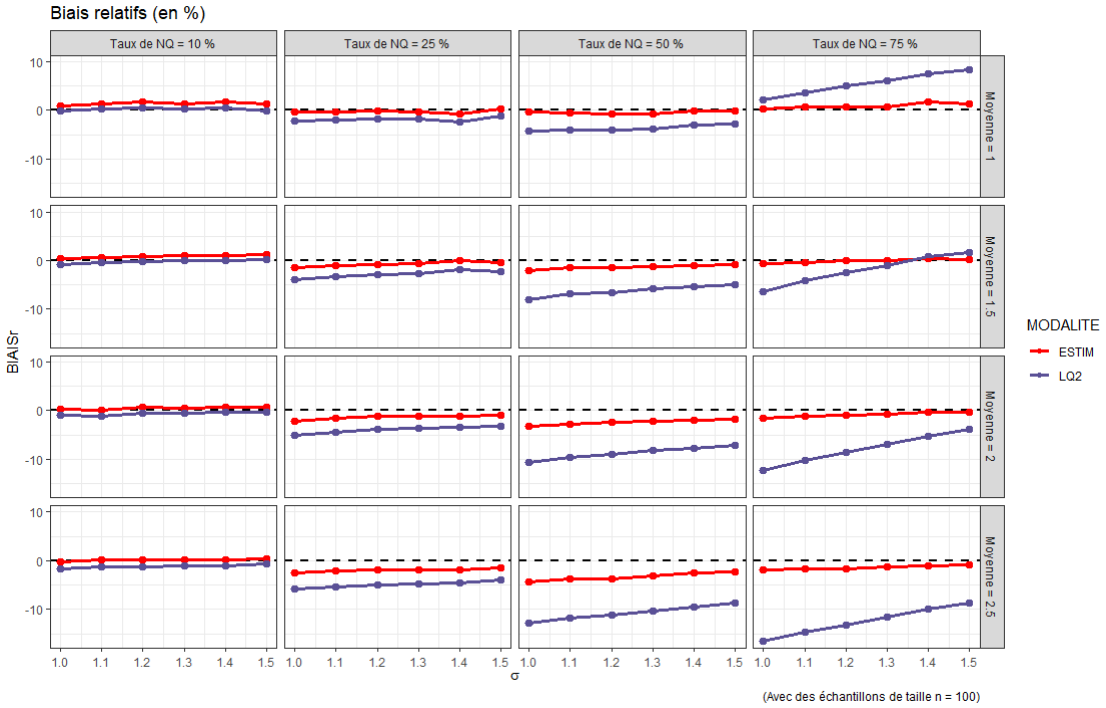
***Source** : SDES*

4.2. Qualité d'estimation sur échantillons de grande taille (100 valeurs)

4.2.1. Biais relatifs

Les comportements sont similaires à ceux observés sur les petits échantillons : au-delà de 10 % de censure, l'estimateur LQ2 est le plus fortement biaisé dans toutes les combinaisons de moyenne $\mu > 1$ et d'écart-type $\sigma > 1$ avec une tendance à sous-estimer les moyennes vraies. ESTIM présente des biais toujours inférieurs à 5 % *(figure 8)*.

Figure 8 : biais relatifs sur des échantillons de taille n = 100

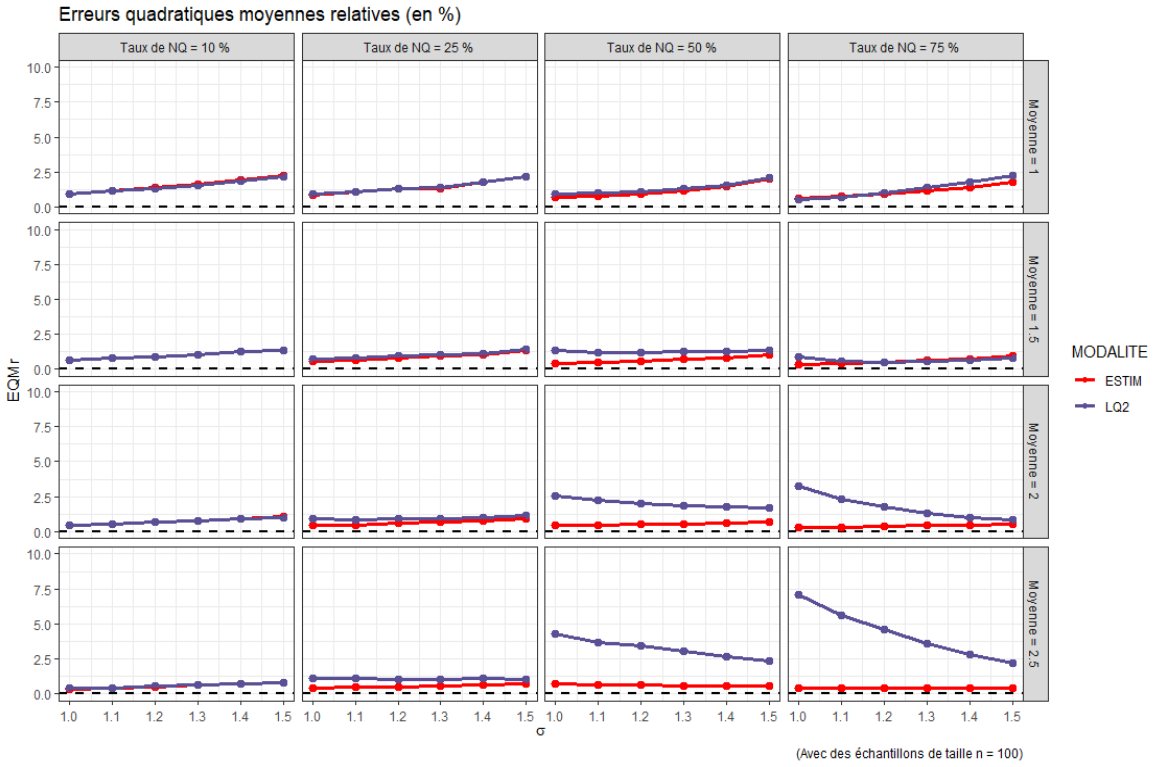


Source *: SDES*

4.2.2. Erreurs quadratiques moyennes relatives (EQMR)

Comme pour les petits échantillons, jusqu'à 25 % de censure, ESTIM présente des erreurs quadratiques quasi identiques à celles de LQ2 et inférieures à 2,5 %. Pour des taux de censure plus élevés, les EQMr d'ESTIM restent à ces faibles niveaux, tandis que celles de LQ2 se dégradent quand les distributions s'aplatissent *(figure 9)*.

Figure 9 : EQM relatives sur des échantillons de taille n = 100



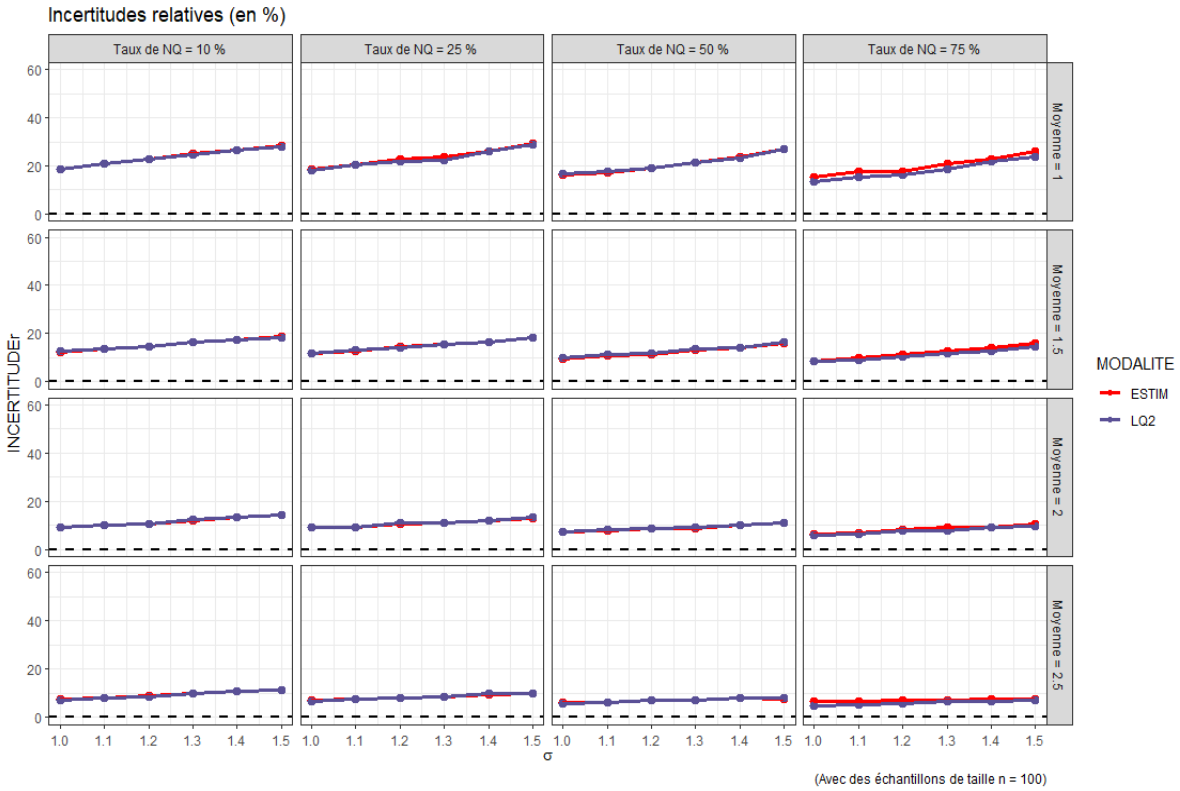
***Source** : SDES*

LQ2 présente des erreurs quadratiques moyennes similaires à celles d'ESTIM dans des situations où il est pourtant plus fortement biaisé (par exemple avec 50 à 75 % de censure et $\mu = 1$ à 1,5). Cela souligne l'impact considérable que la substitution par une valeur fixe comme LQ/2 peut avoir sur la variance de l'estimateur. La faible variance des estimations LQ2 est induite par l'absence de variabilité des valeurs de remplacement et peut être considérée comme douteuse. En conséquence, il semble préférable de ne pas utiliser l'estimateur LQ2 pour faire de l'inférence statistique, les p-valeurs de tests et les intervalles de confiance étant liés de près à la variance des estimations.

4.2.3. Incertitudes relatives

Les incertitudes d'ESTIM et LQ2 sont comprises entre 5 % et 30 % sur l'ensemble des simulations. Elles diminuent lorsque μ augmente. Contrairement au cas des petits échantillons, aucune différence n'est visible entre LQ2 et ESTIM, quels que soient le taux de censure et la forme de la distribution *(figure 10)*.

Figure 10 : incertitudes relatives sur des échantillons de taille n = 100

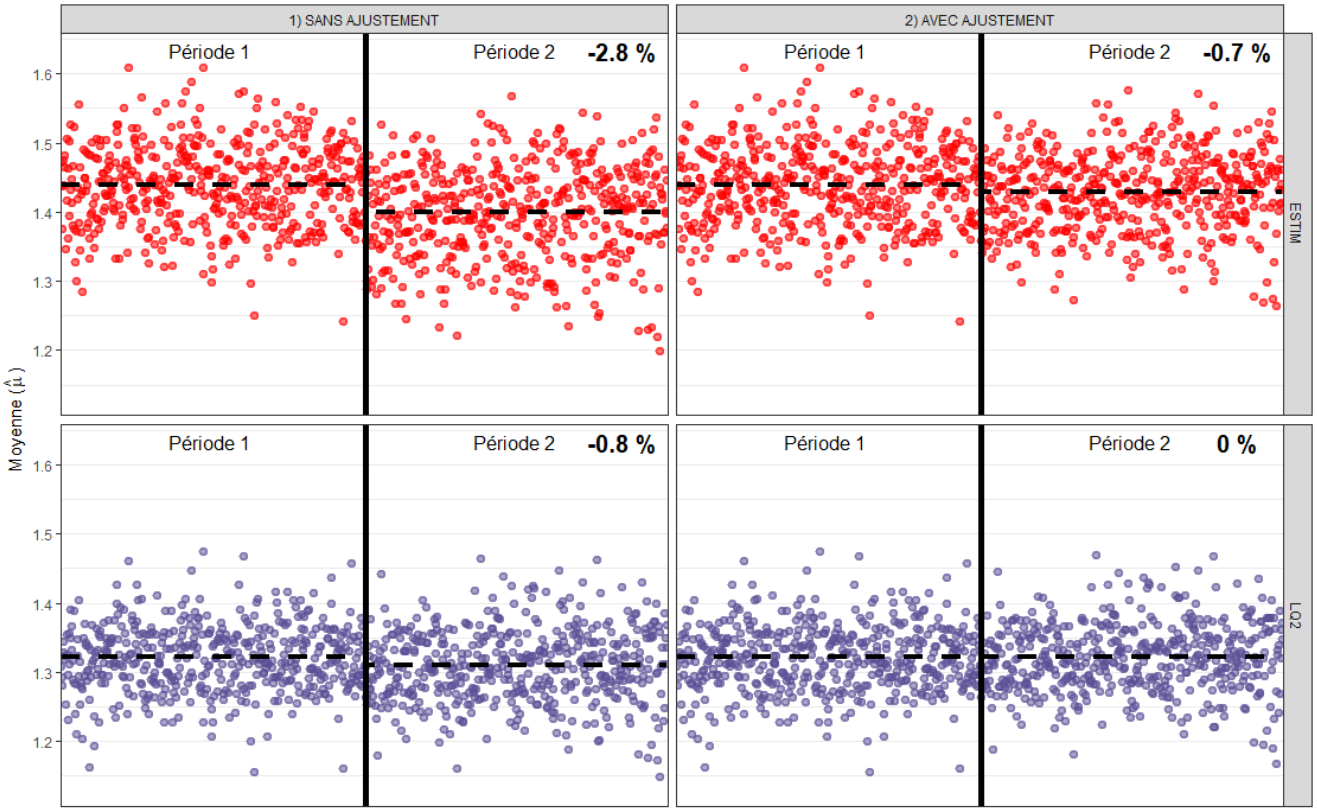


Source : *SDES*

4.3. Réaction au changement de limites de quantification

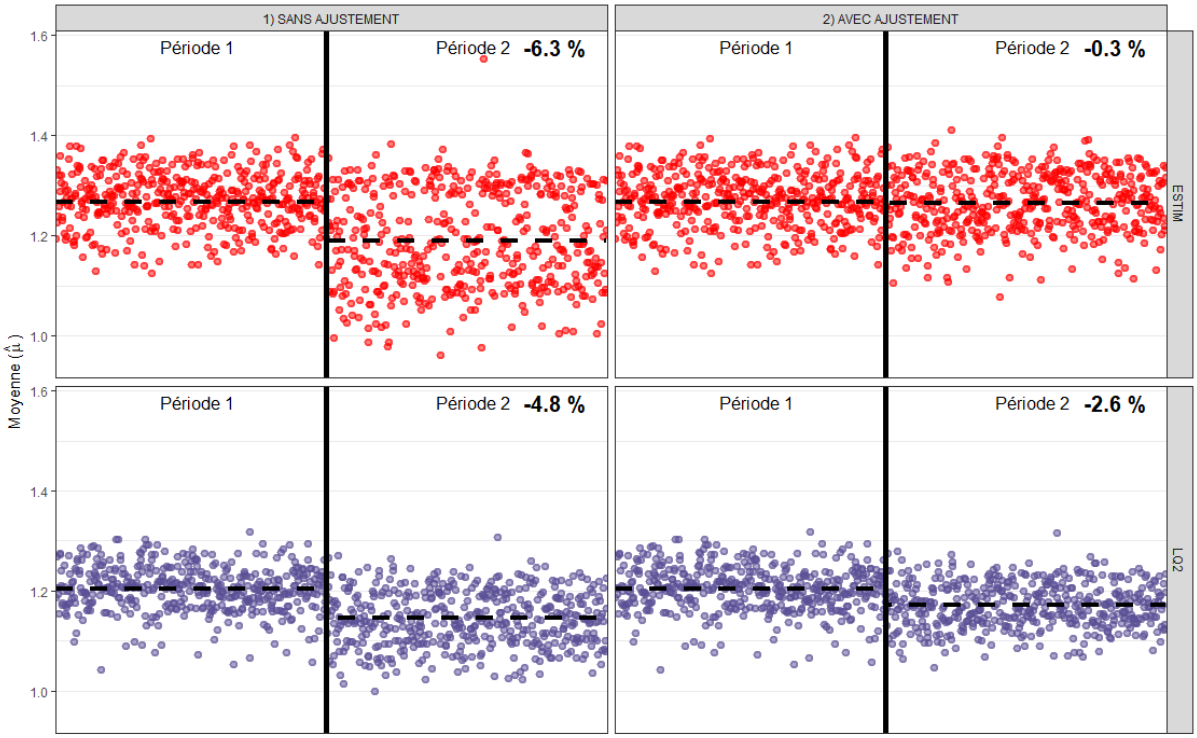
ESTIM et LQ2 sont sensibles au changement de LQ, leurs estimations évoluant dans le même sens que l'évolution des seuils de censure. Sans surprise, les deux estimateurs ont des réactions plus marquées sur les échantillons censurés à 75 % qu'à 50 %. Il semble qu'ESTIM soit plus sensible que LQ2 dans les cas simulés *(figures 11 et 12)*.

Figure 11 : estimations sur échantillons censurés à 50 %, avec et sans correction du changement de LQ



Note : simulation réalisée avec 2 000 échantillons de taille n = 100, censurés à 50 %. De la période 1 à la période 2, la moyenne réelle baisse de 0,1 % et les seuils de censure baissent de 16,6 %. La moyenne des estimations par période est figurée par la ligne pointillée horizontale.
***Source** : SDES*

Figure 12 : estimations sur série d'échantillons censurés à 75 % avec baisse des seuils



Note : simulation réalisée avec 2 000 échantillons de taille n = 100, censurés à 50 %. De la période 1 à la période 2, la moyenne réelle est constante et les seuils de censure baissent de 21,4 %. La moyenne des estimations par période est figurée par la ligne pointillée horizontale.
***Source** : SDES*

La méthode utilisée pour corriger les changements de LQ réduit efficacement l'élasticité M-LQ. Dans les « pires situations » simulées (75 % de censure et 20 % de baisse de LQ), cette méthode neutralise presque complètement l'effet des variations de LQ sur ESTIM et réduit l'élasticité de moitié sur LQ2 *(figure 13)*.

Figure 13 : élasticité M-LQ avant et après correction des changements de seuils de censure

Estimateur	Taux de censure	Avant correction des changements de LQ	Après correction des changements de LQ
ESTIM	50 %	0,16	0,04
ESTIM	75 %	0,29	0,01
LQ2	50 %	0,04	0,00
LQ2	75 %	0,22	0,12

***Source** : SDES*

5. Discussion - conclusions

Les biais relatifs rapportés dans la présente étude peuvent être différents de ceux observables sur des jeux de données réelles. En effet, les paramètres de la distribution sous-jacente des données réelles sont inconnus dans la pratique, et les taux de censure peuvent ne pas être exactement ceux qui ont été simulés.

La présente étude montre que :

- D'une manière générale, les biais et les erreurs quadratiques des estimateurs ESTIM et LQ2 augmentent à mesure que la taille des échantillons diminue, que le taux de censure augmente et que l'asymétrie de la distribution augmente. Toutefois, de nettes différences s'observent selon l'estimateur considéré.

- L'estimateur ESTIM est globalement plus performant que LQ2. Il présente systématiquement des biais faibles (- 5 % à + 5 %) et des erreurs quadratiques moyennes inférieures à 8 %. Il est robuste aux changements de tailles d'échantillons, de taux de censure et de forme de distribution des données (dans les limites des cas simulés).

- L'estimateur LQ2 est peu biaisé jusque 25 % de censure (- 5 % à + 5 %). Pour des taux de 50 % et 75 %, il présente soit de plus grandes erreurs quadratiques moyennes qu'ESTIM (pour des distributions de formes plus aplaties), soit des EQMr similaires à ESTIM mais liées à des variances artificiellement basses (induites par l'absence de variabilité des valeurs de remplacement). Cela confirme ce que d'autres auteurs mettent en évidence sur l'estimateur LQ2 [1, 2, 3, 4, 5, 9, 10].

- ESTIM et LQ2 sont sensibles à un changement de seuils de censure, une baisse de LQ induisant des estimations de moyennes plus basses. Toutefois, la méthode corrective qui a été mise en œuvre diminue efficacement cette sensibilité ; elle peut être utilisée dans le cadre d'une étude de tendance sur une série chronologique où les seuils de censure varient dans le temps.

Le comportement des deux estimateurs sur séries chronologiques réelles est illustré en *annexe 5* pour deux molécules surveillées dans les cours d'eau de métropole.

Les résultats de cette étude incitent à ne pas utiliser l'estimateur LQ2 pour estimer des moyennes sur des séries chronologiques où les formes de distribution, les taux de censure et les seuils de censure changent rapidement et de façon importante d'une période à l'autre. Une analyse de tendance avec l'estimateur LQ2 pourrait conduire à des conclusions erronées. L'estimateur ESTIM, associé à la méthode de redressement des LQ, est plus approprié pour ce genre d'étude.

Bien qu'ESTIM présente des qualités satisfaisantes sur la gamme des situations simulées, il convient de ne pas extrapoler ses performances à des situations « plus dégradées » (par exemple, aux échantillons censurés à plus de 80 %).

6. Annexes

Annexe 1. Méthodes d'estimation de Kaplan-Meier (KM) et par régression robuste sur statistique ordonnée (rROS)

- Notations

Soit X, un vecteur de n observations issu d'une distribution lognormale de moyenne μ et d'écart-type σ.

On suppose que q de ces n observations sont quantifiées ($0 < q < n$) et que c d'entre elles sont censurées à gauche ($c = n - q$), c'est-à-dire inférieures à k limites de quantifications connues $(lq_1, lq_2, \dots, lq_k;\ k \geq 1)$. Ainsi, une observation de vecteur désigne soit une valeur quantifiée, soit une limite de quantification. De plus, si c_j est le nombre d'observations inférieures à la limite de quantification lq_j (pour$j = 1,2, \dots, k$), on a

$$\sum_{i=1}^{k} c_j = c$$

Les n observations sont classées par ordre croissant $(x_1, x_2, \dots, x_n)$. Si une observation censurée à la même valeur qu'une observation quantifiée, l'observation censurée est placée en premier. La quantité x_i ne représente donc pas forcément la $i^{ème}$ plus grande observation de l'échantillon.

Soit Ω, l'ensemble des n indices des observations quantifiées de l'échantillon ordonné.

Finalement, $Y = \log(X)$ désigne un vecteur d'observations issu d'une distribution normale de moyenne μ_* et d'écart-type σ_*.

- Estimateur de Kaplan-Meier (KM)

Il est connu que la moyenne d'une distribution à valeurs positives est égale à l'aire sous la courbe de survie :

$$\mu = \int_0^\infty [1 - F(t)]dt = \int_0^\infty S(t)dt$$

où

$F(t)$ désigne la fonction de répartition évaluée à t et $S(t) = 1 - F(t)$ la fonction de survie évaluée à t.

Quand l'estimateur de Kaplan-Meier est utilisé pour construire la fonction de survie, il est possible d'utiliser l'aire sous cette courbe pour estimer la moyenne de la distribution.

Soit $\hat{F}(t)$, l'estimateur de Kaplan-Meier de la fonction de répartition empirique évaluée à t et $\hat{S}(t) = 1 - \hat{F}(t)$ la fonction de survie évaluée à t. L'estimateur de moyenne est calculé par :

$$\hat{\mu} = \sum_{i=1}^{n} \hat{S}(y_{i-1})(y_i - y_{i-1})$$

où

$y_0 = 0$ et $\hat{S}(y_0) = 1$ par définition.

Ce calcul est équivalent à :

$$\hat{\mu} = \sum_{i=1}^{n} y_i \left[\hat{F}(y_i) - \hat{F}(y_{i-1})\right]$$

où

$\hat{F}(y_0) = \hat{F}(0) = 0$ par définition (USEPA, 2009, p. 15-10).

L'estimateur de Kaplan-Meier est adapté aux échantillons comportant au moins deux limites de quantification distinctes. Il n'est pas recommandé lorsqu'il y a un seul seuil de censure.

- Estimateur rROS

La méthode débute par une régression quantiles-quantiles sur les logarithmes des observations quantifiées afin d'obtenir une estimation initiale de μ_* et de σ_* (échelle logarithmique). Les valeurs des observations censurées sont ensuite imputées avec les valeurs prédites de l'équation de régression, puis transformées dans l'échelle d'origine. Finalement, l'estimateur de moyenne est calculé sur les valeurs quantifiées et imputées par la méthode classique des moments.

Étape 1 : estimation de μ_* et de σ_* par les moindres carrés selon le modèle :

$$y_i = \mu_* + \sigma_* \Phi^{-1}(p_i) + \varepsilon_i,\ i \in \Omega$$

où

p_i est la position de la $i^{ème}$ plus grande valeur sur la courbe des probabilités empiriques cumulées :

Φ^{-1} est la fonction de répartition inverse de la loi normale standard (indique la valeur de la variable associée à une probabilité cumulée donnée)

Étape 2 : calcul des valeurs imputées d'après l'équation précédente :

$$\hat{y}_i = \hat{\mu}_{*qqreg} + \hat{\sigma}_{*qqreg} \Phi^{-1}(p_i),\ i \notin \Omega$$

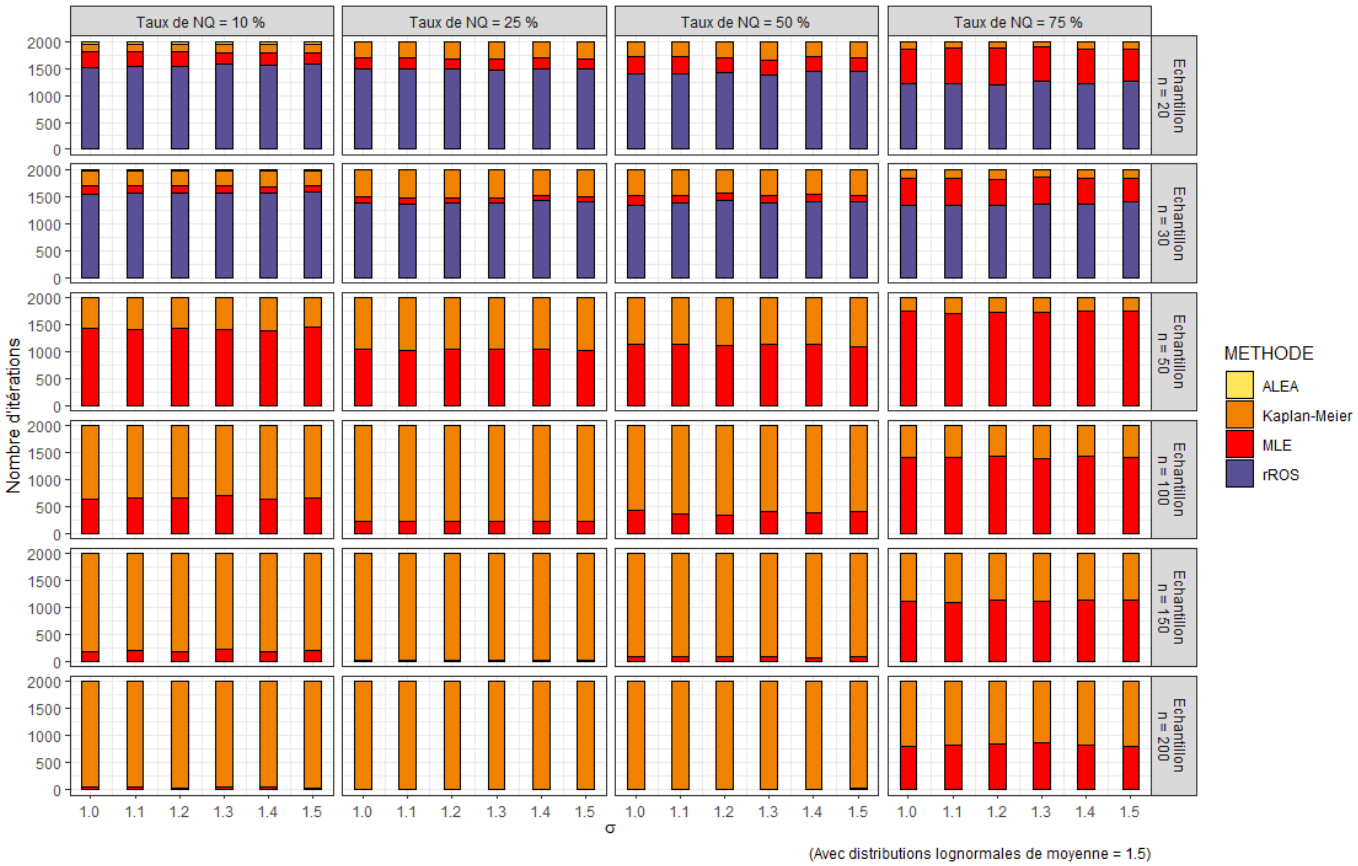
Étape 3 : transformation des valeurs imputées dans l'échelle d'origine :

$$\hat{x}_i = exp(\hat{y}_i),\ i \notin \Omega$$

Étape 4 : calcul de l'estimateur final par la méthode des moments :

$$\hat{\mu} = \frac{1}{N}\left(\sum_{i \notin \Omega} \hat{x}_i + \sum_{i \in \Omega} x_i\right)$$

Annexe 2. Occurrences des méthodes utilisées par ESTIM

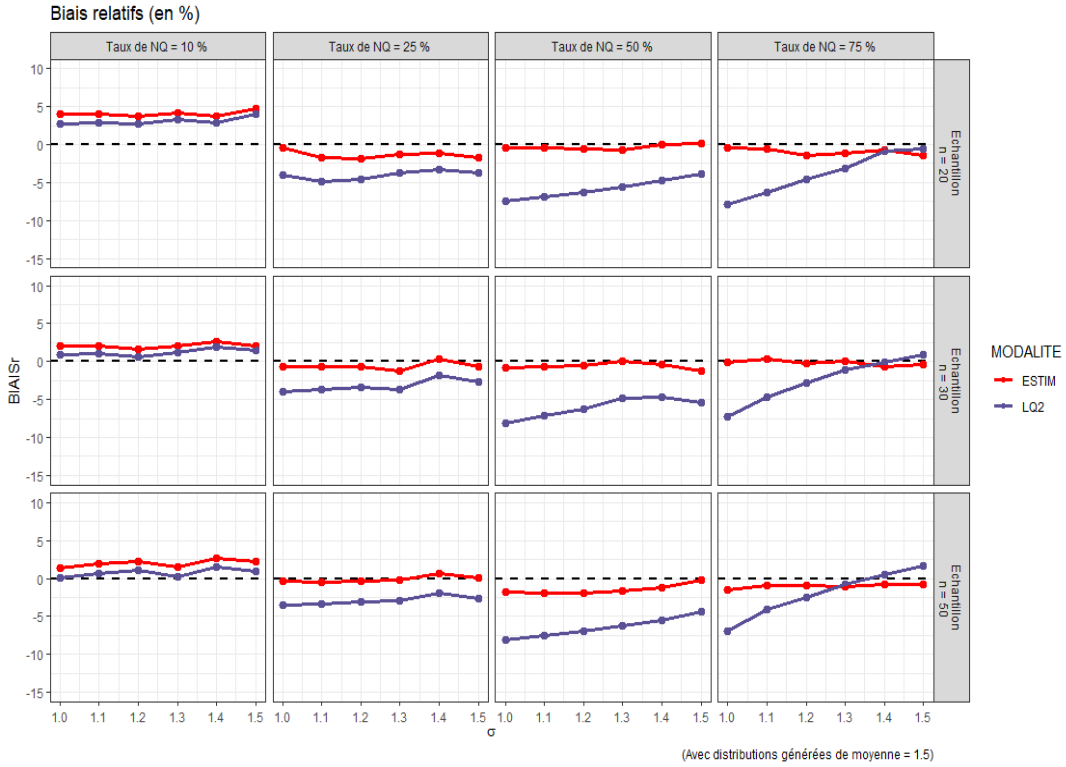


***Source** : SDES*

La répartition des méthodes est similaire sur la gamme allant de $\mu = 1$ à $\mu = 2{,}5$.

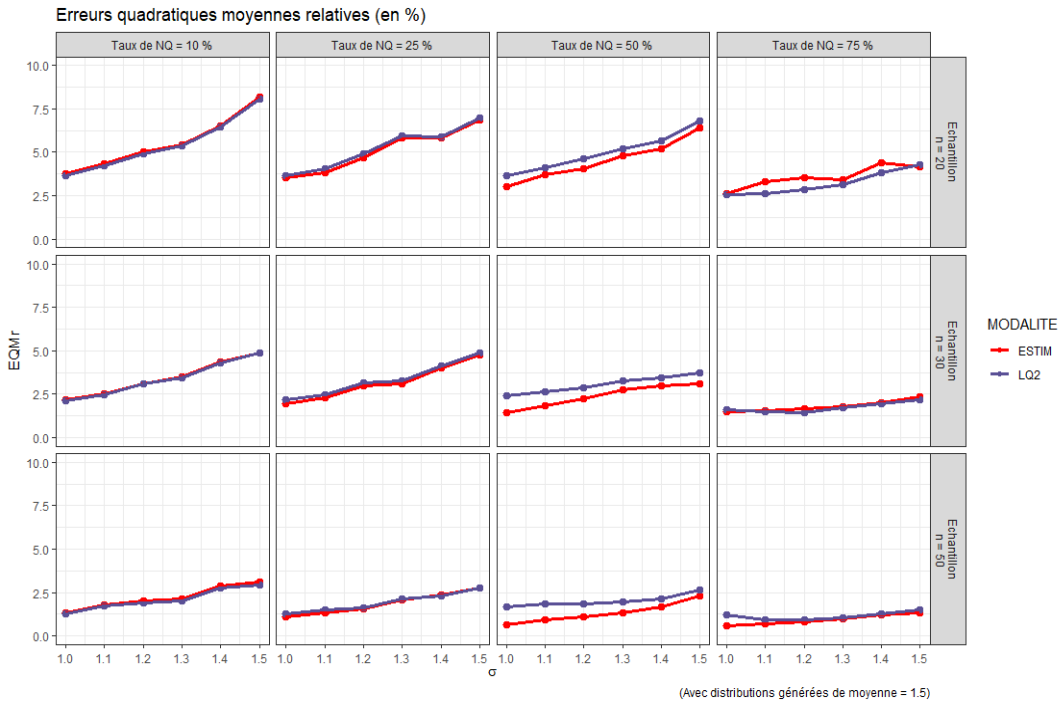
Annexe 3. Résultats complémentaires sur échantillons de taille modérée (20 < n < 50)

Figure 14 : biais relatifs sur échantillons issus de distributions de moyenne $\mu = 1,5$



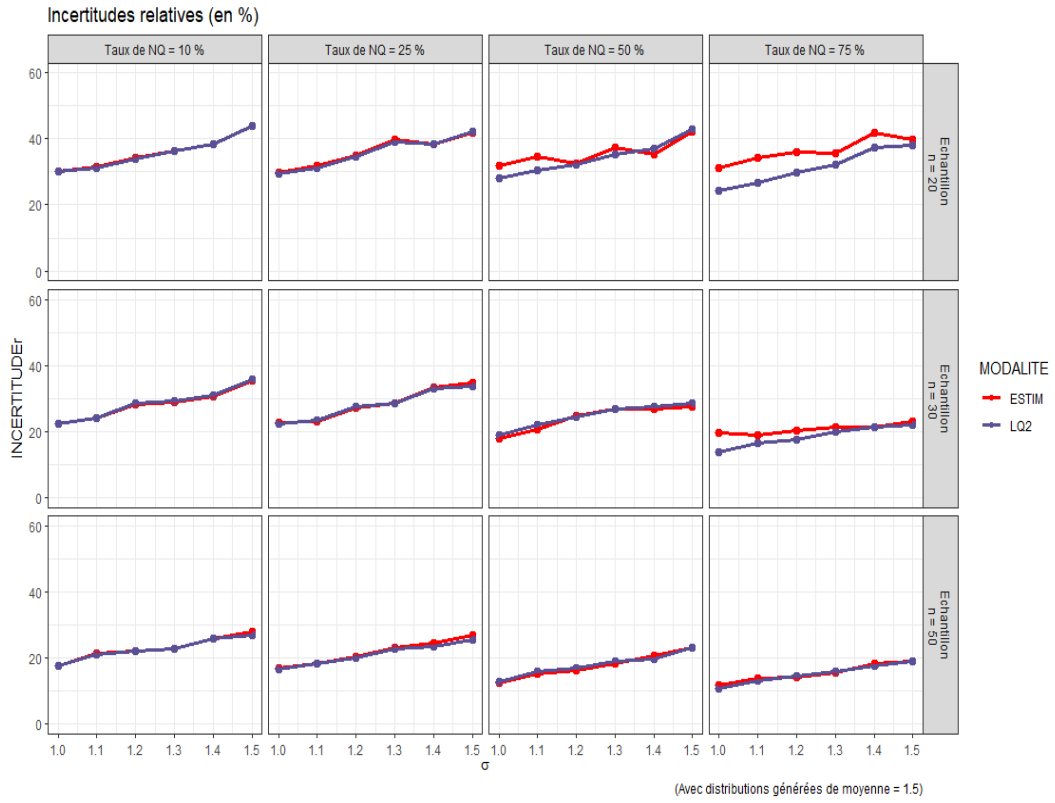
Source *: SDES*

Figure 15 : EQM relatives sur échantillons issus de distributions de moyenne $\mu = 1,5$



Source *: SDES*

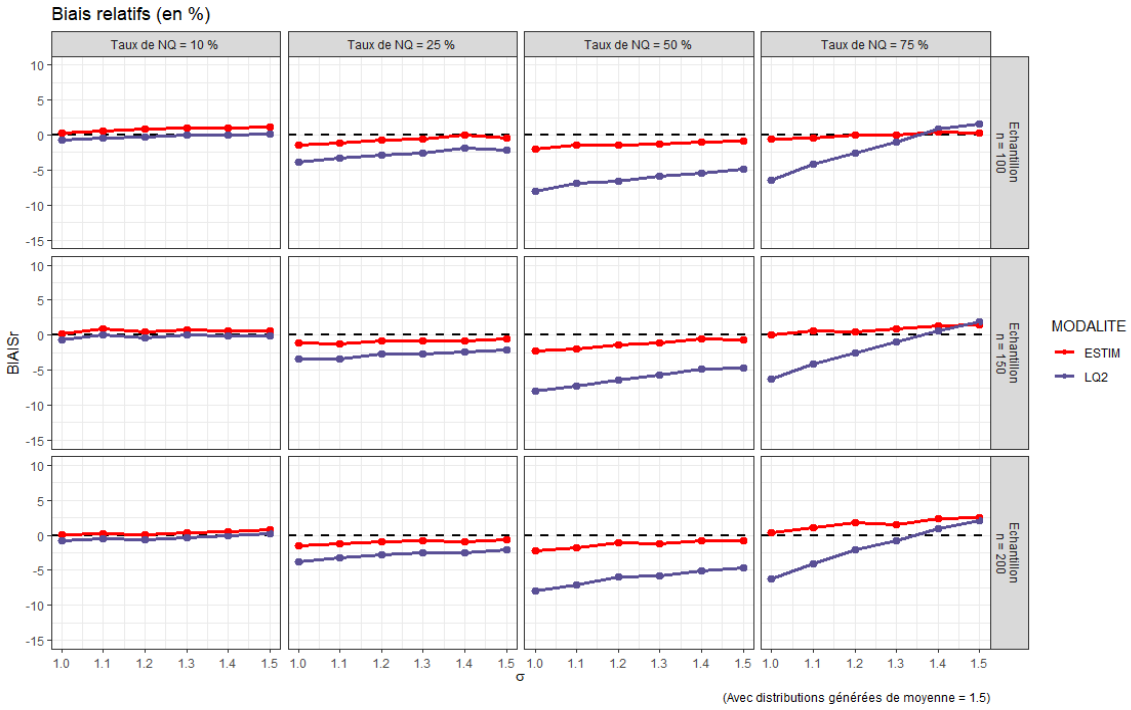
Figure 16 : incertitudes relatives sur échantillons issus de distributions de moyenne $\mu = 1,5$



***Source** : SDES*

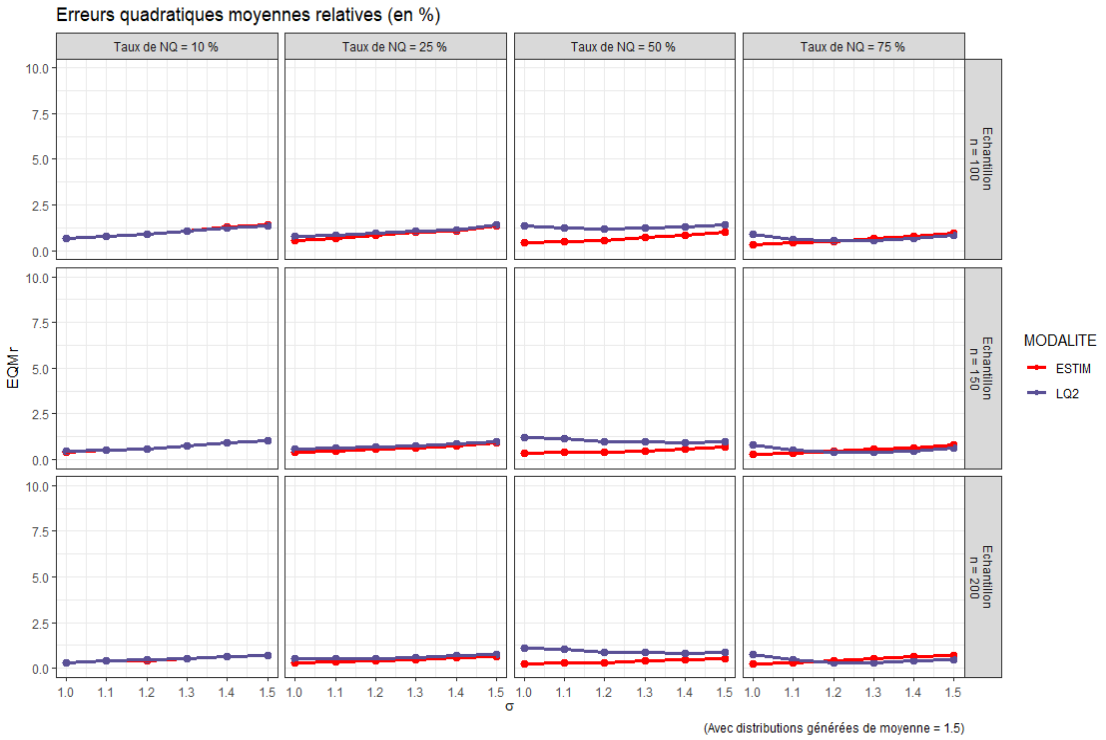
Annexe 4. Résultats complémentaires sur échantillons de grande taille (100 < n < 200)

Figure 17 : biais relatifs sur échantillons issus de distributions de moyenne $\mu = 1,5$



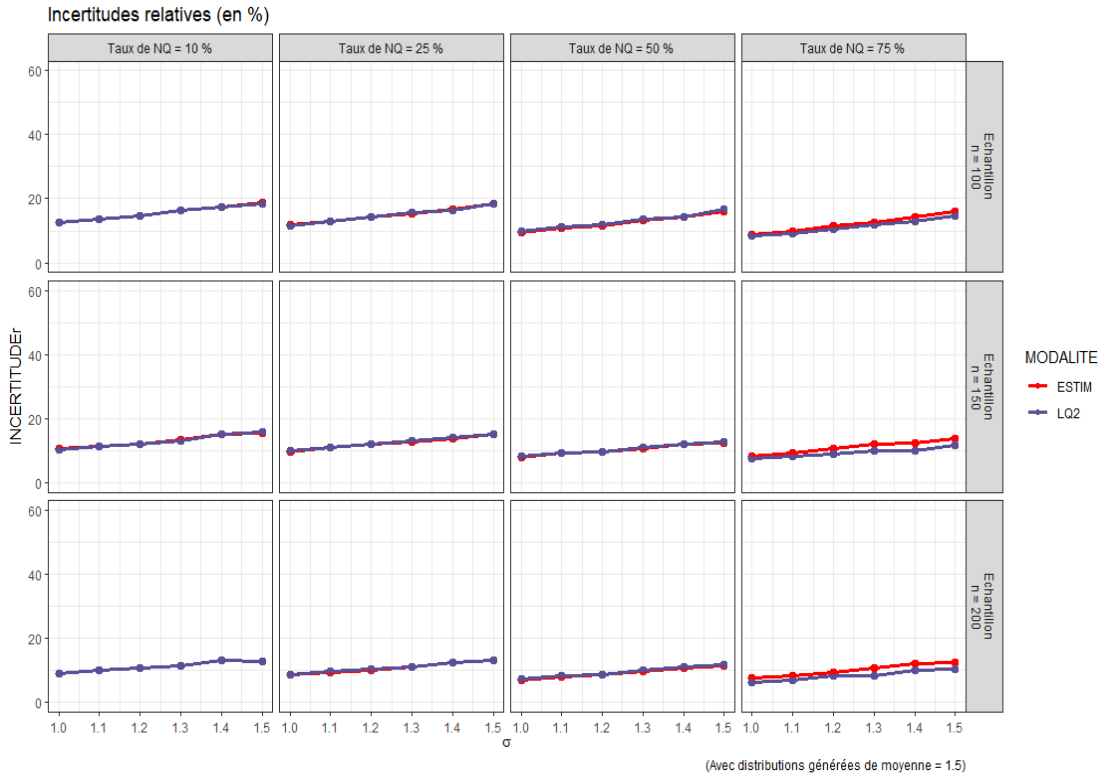
***Source** : SDES*

Figure 18 : EQM relatives sur échantillons issus de distributions de moyenne $\mu = 1,5$



***Source** : SDES*

Figure 19 : incertitudes relatives sur échantillons issus de distributions de moyenne $\mu = 1,5$



***Source** : SDES*

Annexe 5. Comparaison de chroniques obtenues avec ESTIM et LQ2 sur données réelles

Les figures 20 et 21 illustrent les différences de moyennes estimées par ESTIM et LQ2 sur des séries chronologiques de pesticides surveillées dans les cours d'eau de métropole. Les différences peuvent être faibles ou de grande ampleur, tant en valeur absolue qu'en variation entre un instant t et t +1.

Figure 20 : concentrations moyennes du pesticide 1G dans les cours d'eau de métropole

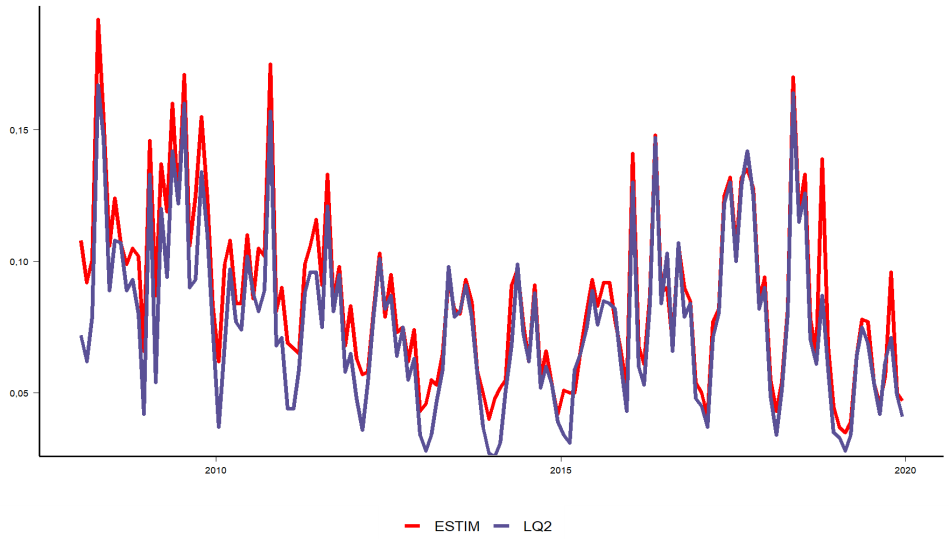
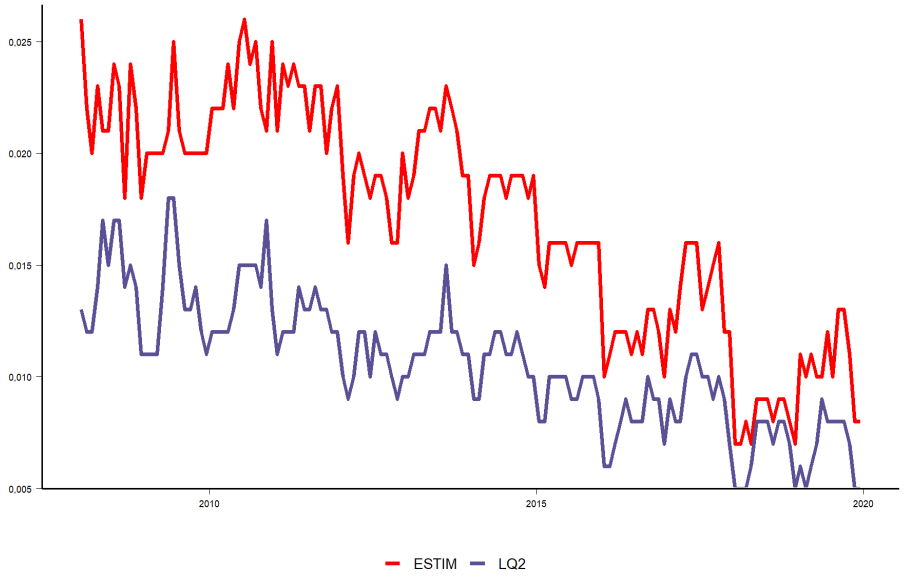


Figure 21 : concentrations moyennes du pesticide 2A dans les cours d'eau de métropole



***Source** : SDES*

Annexe 6. Références

1. Baccarelli, A.*, et al*. (2005). *Handling of dioxin measurement data in the presence of non-detectable values : overview of available methods and their application in the Seveso chloracne study*. Chemosphere. Aug. 60 (7) :898-906.

2. European Food Safety Authority. (2010). *Management of left-censored data in dietary exposure assessment of chemical substances*.

3. Helsel, D. R. (1990). *Less than obvious : statistical treatment of data below detection limit.* Environmental science & technology, 24, 1767-1774.

4. Helsel, D.R. (2012). *Statistics for Censored Environmental Data Using Minitab and R, Second Edition*. John Wiley & Sons, Hoboken, New Jersey.

5. Huynh,T.*, et al*. (2014). *Comparison of Methods for Analyzing Left-Censored Occupational Exposure Data*. The Annals of occupational hygiene. 58.

6. Royston, J.P. (1992). *Approximating the Shapiro-Wilk W-Test for Non-Normality. Statistics and Computing 2*, 117-119.

7. Shapiro, S.S., and M.B. Wilk. (1965). *An Analysis of Variance Test for Normality* (Complete Samples). Biometrika 52, 591-611.

8. Shoari, N., Dubé, J.-S. & Chenouri, S. (2015). *Estimating the mean and standard deviation of environmental data with below detection limit observations : Considering highly skewed data and model misspecification*. Chemosphere, 138, 599-608.

9. Shoari, N., Dubé, J.-S. & Chenouri, S. (2016). *On the use of the substitution method in leftcensored environmental data*. Human & ecological risk assessment, 22 (2), 435-446.

10. Singh, A.; Singh, A.K. (2015). ProUCL Version 5.1.002 Technical Guide - *Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations*. EPA : Washington, WA, USA.

11. USEPA. (2009). *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance*. EPA 530/R-09-007, March 2009. Office of Resource Conservation and Recovery Program Implementation and Information Division. U.S. Environmental Protection Agency, Washington, D.C.

MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE
ET DE LA COHÉSION
DES TERRITOIRES

Liberté
Égalité
Fraternité

Commissariat général
au développement durable

Commissariat général au développement durable
Service des données et études statistiques
Sous-direction de l'information
environnementale
Tour Séquoia - 92055 La Défense cedex

www.statistiques.developpement-durable.gouv.fr

STATISTIQUE
PUBLIQUE