Compléments sur la quantification des résultats

pondération et transformation

Michel Laurentie[1]

Résumé : *La pondération ou la transformation des données sont des outils nécessaires à leur exploitation statistique. Pondérer ou transformer des données repose sur la nécessité de stabiliser la variance. En effet, un principe théorique pour l'utilisation de la régression selon le critère des moindres carrés impose que les variances soient homogènes, c'est à dire quelles soient indépendantes du niveau des concentrations. Cet article décrit ces outils et montre au travers un exemple leur impact sur la validation des méthodes.*

Mots-Clés : pondération, transformation, validation, profil d'exactitude, méthodes d'analyse quantitatives

Problématique

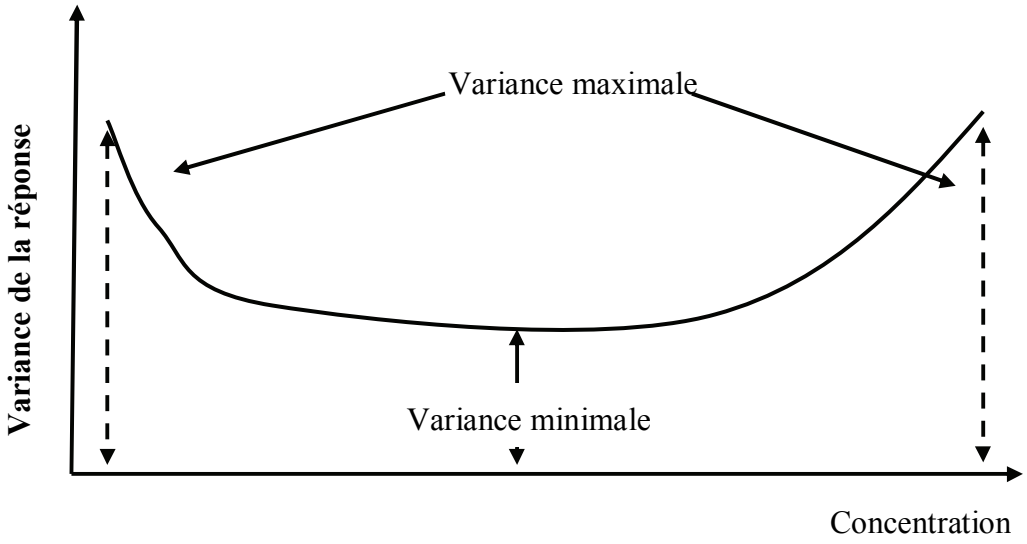
De nombreux logiciels de chromatographie permettent de calculer les critères de performance d'une méthode analytique. En particulier, ils proposent des modèles d'étalonnage (fonction de réponse) permettant de lier la réponse aux différents niveaux de concentration. Bien souvent ces logiciels proposent systématiquement un facteur de pondération.
La pondération est un des moyens de respecter une règle fondamentale dans l'analyse statistique de données : l'homogénéité des variances.
Une autre approche pour respecter cette règle est celle de la transformation des données qui est plus rarement utilisée par les analystes car elle est souvent assimilée à une modification des données qui ne semble pas justifiée. À l'inverse dans d'autres disciplines telles que la microbiologie cette approche est systématique (transformation logarithmique pour analyser les comptages bactériens).
La question est alors la suivante : quand faut-il pondérer ou utiliser des données transformées ?

1. Quand faut-il agir ?

Dans une gamme de concentration, la variance des réponses peut varier en fonction de la concentration, d'autant plus que l'intervalle de gamme est important (Mac Taggart D.L. et Farwell S.O., 1992). On dit alors que la variance n'est pas homogène.

Généralement la variance est plus large aux extrémités du domaine qu'en son milieu. La Figure 1 illustre cette observation.

[1] Unité de pharmacocinétique-pharmacodynamie - AFSSA – F-35302 Fougères cedex- m.laurentie@afssa.fr



Figu**re 1 :** *représentation de l'évolution de la variance des réponses* en fonction *du niveau de concentration*

La représentation graphique de la variance de la réponse en fonction de la concentration n'est pas fournie dans les logiciels de chromatographie, mais cette observation peut être également visualisée sur une droite d'étalonnage (**figure 2**).

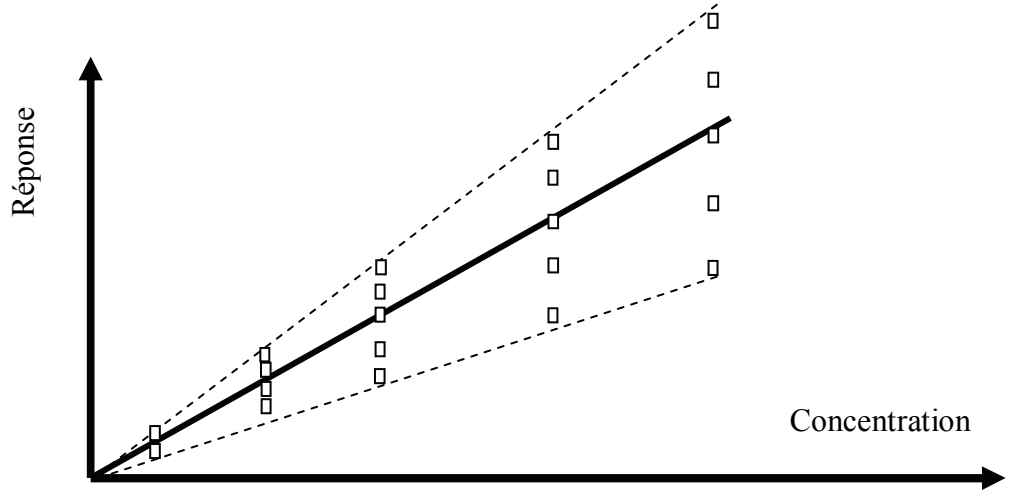


Figure 2 : *évolution des réponses en fonction du niveau des concentrations, représentée par rapport à une courbe d'étalonnage*

Une autre possibilité pour évaluer l'impact de la variance sur le modèle d'étalonnage est l'utilisation du graphique des résidus. Le résidu est la différence entre la valeur observée et la valeur prédite (Équation 1)

$$résidu = Y_{obs} - \hat{Y}_{cal} \qquad \text{Équation 1}$$

Avec Y_{obs} la valeur observée, $\hat{Y}_{cal}$ la valeur calculée par le modèle d'étalonnage *e.g.*un modèle linéaire.

Si la répartition des résidus est homogène dans l'intervalle des concentrations, il est conclu à la validité du modèle et à l'homogénéité des variances par niveau des concentrations. Si au

contraire les résidus paraissent structurés, *i.e.* que distribution suit une courbe particulière (droite par exemple) ou si la distribution des résidus montre une variabilité importante ou une forme particulière, par exemple, en trompette, il y a hétérogénéité des variances et par conséquent il est conclut à la non validité du modèle. Ceci est illustré sur les figures 3A et 3B.

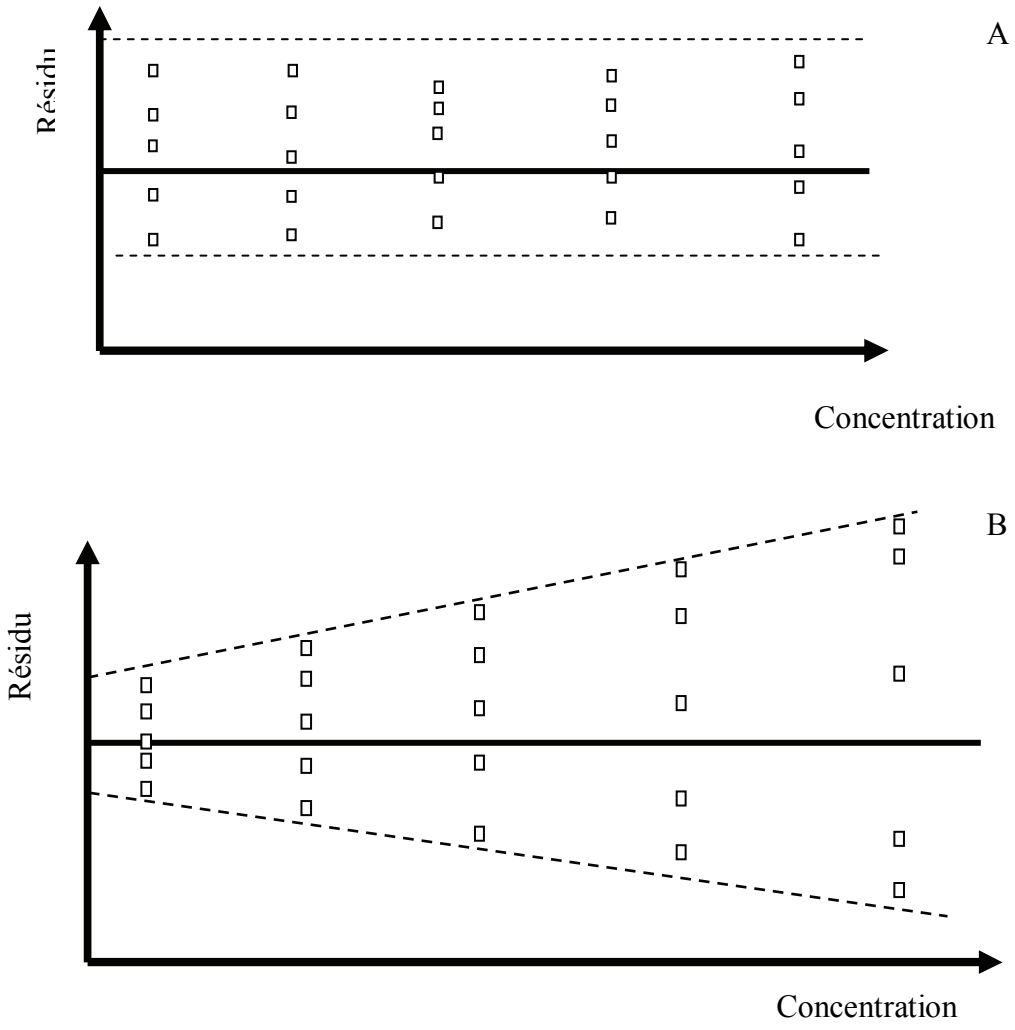


Figure 3 : *(A) répartition des résidus pour un modèle d'étalonnage pour lequel la variance des réponses est homogène.*
(B) répartition des résidus pour un modèle d'étalonnage pour lequel les variances des réponses sont non homogènes (forme dite en trompette).

La connaissance de la variance, à chaque niveau de concentration, est donc primordiale pour établir correctement le modèle d'étalonnage ; ceci implique que le schéma expérimental doit être rigoureux, avec en particulier, un nombre suffisant de répétition pour évaluer la variance aux extrémités du domaine de dosage.

Cette analyse graphique peut-être complétée par une analyse statistique basée sur l'utilisation de tests tel que les tests de Cochran ou de Levène (Hubert Ph. *et al.,* 1999). Ces test sont largement décrits dans la littérature et ne seront pas décrits dans cet article. Toutefois il

convient de préciser que ces tests sont des drapeaux pour indiquer l'hétérogénéité des variances et non des tests pour éliminer des niveaux de concentration qui seraient dits aberrants.

2. La stabilisation des variances

Si une hétérogénéité des variances est observée, il convient de la prendre en compte. Deux approches sont possibles : la pondération ou la transformation des données.

2.1 La transformation des données

La transformation des données permet de stabiliser la variance. Les plus utilisées sont les logarithmes ou les racines carrées.

Dans le cas de l'utilisation de la transformation, logarithmique le modèle utilisé s'écrit :

$$logY = a\,logC + b \qquad \text{Équation 2}$$

Ou le log$Y$ est le logarithme népérien de la réponse Y et log$C$ le logarithme népérien de la concentration.

Les techniques de régression sont alors utilisées comme pour une régression classique, en particulier l'utilisation des moindres carrés. L'observation des résidus doit montrer une stabilisation des variances. Dans ce cas le modèle de fonction de réponse utilisée en routine implique de transformer les réponses avant de calculer les concentrations en retour.

La deuxième approche est l'utilisation de la pondération.

2.2 La pondération

Le principe de la pondération est simple. Pour chaque niveau de concentration d'un intervalle de dosage, la variance n'est pas proportionnelle à la concentration. Ceci implique de donner un poids à chaque niveau de concentration pour corriger le manque de proportionnalité.

Prenons le cas d'une régression linéaire de type (Tomassone R., Lesquoy E., Miller C., 1983) :

$$Y = aC + b \qquad \text{Équation 3}$$

Avec Y la réponse, a et b les coefficients de la droite de régression.

Lors d'une estimation des paramètres a et b nous cherchons à minimiser la différence entre les valeurs calculées et les valeurs estimées ce qui se traduit par l'équation

$$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 \qquad \text{Équation 4}$$

Avec SCE : somme des carrées des écarts entre la énième observation Y_i et l'estimation $\hat{Y}$ par la droite de régression.

La prise en compte du poids (w_i) se fera par l'introduction d'un facteur dans l'équation 4.

$$SCE = \sum_{i=1}^{n} w_i (Y_i - \hat{Y})^2 \qquad \text{Équation 5}$$

La sélection de la valeur du poids peut se faire graphiquement. Sur la **figure 4**, les différentes relations entre la variance des réponses et les niveaux de concentration sont représentées.

Plusieurs types de relation existent entre la variance et la concentration [4]. A chaque type de structure il est possible de définir un poids (w_i pour weighting factor) qui permettra de prendre en compte l'influence de la variance. Quatre cas sont généralement observés :

1 - Il n'y a pas de relation entre la concentration et la variance (le cas idéal) : le poids $w_i = 1$. En d'autres il n'y a pas de nécessité de pondérer

2 - la variance varie avec une relation de type $1/X$ i.e. inversement proportionnelle à la concentration, le poids qui doit être pris en compte est de la forme $w_i = 1/C$.

3 - la variance varie selon une relation de type $1/X^2$ i.e. inversement proportionnelle au carré de la concentration, le poids à utiliser est de la forme $w_i = 1/C^2$

4 - la relation n'est pas une relation simple entre la concentration et la variance, chaque variance variant indépendamment du niveau de concentration, le poids $w_i = 1/S^2$ avec S^2 la variance observée à chaque niveau de concentration.

Les cas 2 et 3 sont les plus fréquemment rencontrés mais le cas 4, cas le plus difficile, peut également s'observer.

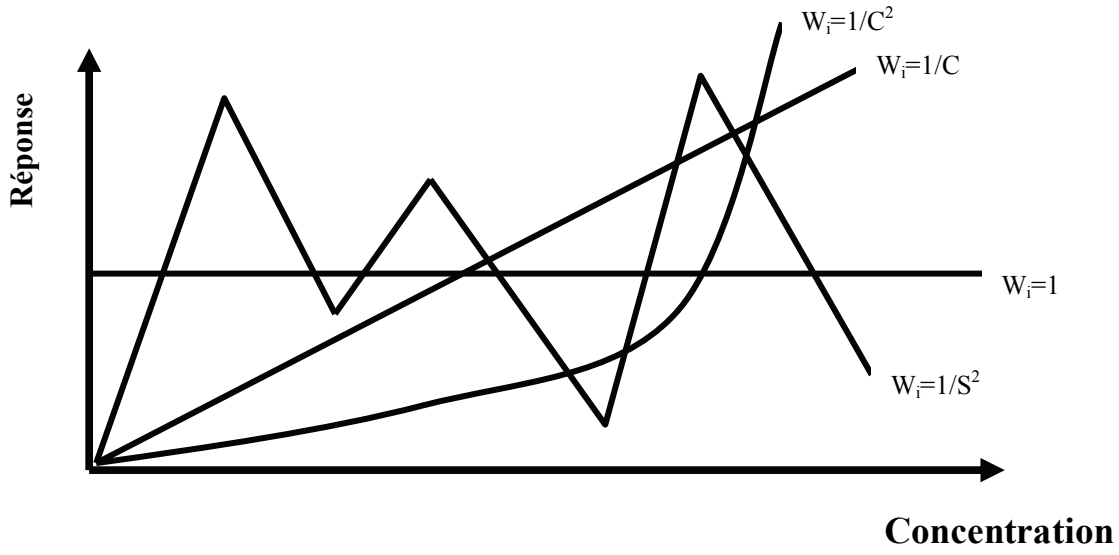


Figure 4 : *types de relation fréquemment observées entre la variance de la réponse et les niveaux de concentration. Les quatre courbes présentées correspondent aux quatre cas décrits dans le texte*

Ce type de graphique peut être facilement réalisé sur une feuille Excel, à partir des données obtenues lors de l'étude de validation ou de pré-validation. Il devrait être un préalable avant toute analyse statistique car ce graphique permet de visualiser s'il faut appliquer une pondération.
La pondération est applicable quel que soit le modèle choisi et l'utilisation d'un facteur permet de minimiser la *SCE* et donc d'avoir des meilleurs estimateurs pour les coefficients de la droite de régression.

3. Exemple

Un exemple d'utilisation de la pondération ou de la transformation des données est montré dans l'exemple du dosage de l'acrylamide.

L'acrylamide est un produit neo-formé lors de la cuisson de certains aliments. Il est issu d'une réaction de Maillard par combinaison de sucre (*i.e.* glucose) et de certains acides aminés comme l'asparagine. Si de nombreuses études documentent la teneur en acrylamide dans les aliments peu d'études ont permis d'en quantifier l'absorption après ingestion. Une étude de pharmacocinétique, réalisée chez le porc doit permettre de déterminer la biodisponibilité de l'acrylamide. Au préalable une méthode analytique est nécessaire pour quantifier dans le plasma de porc les concentrations en acrylamide. La méthode retenue pour effectuer ce dosage est une méthode HPLC associée à une détection par spectrométrie de masse (MS). La gamme étudiée varie de 10 à 5000 ng/ml, compte tenu que nous n'avons pas d'information sur les taux plasmatiques circulants après ingestion d'aliment contenant de l'acrylamide.

Deux gammes sont préparées : une gamme standard d'étalonnage et une gamme dans du plasma de porc (standard de validation). À 200 µl de plasma est ajouté 100 µl de solution saturée de $ZnSO_4$ puis 1000 µl d'acétonitrile et 100 µl de standard interne (acrylamide D5). Après agitation et centrifugation le surnageant est évaporé. L'éluât est repris avec 200 µl d'acétate d'ammonium 0,01M pH 6. Le volume d'injection est de 50 µl. Les conditions analytiques utilisées sont un débit chromatographique de 0,2ml/mn à travers une colonne Hypercarb (5µ 50-2 mm) et une détection par spectrométrie de masse sur l'ion moléculaire 72 de l'acrylamide. La gamme standard d'étalonnage en acrylamide varie de 10 à 5000 ng/ml et est réalisée dans une solution d'acétate d'ammonium 0,01M ramenée à pH 6 avec de l'acide formique.

Le plan d'expérience utilisé consiste en 5 jours, 6 niveaux et 2 répétitions (5 × 6 × 2) soit 60 essais pour les standards d'étalonnage et de validation.

La **figure 5** montre l'évolution de la variance des réponses en fonction des niveaux de concentration.

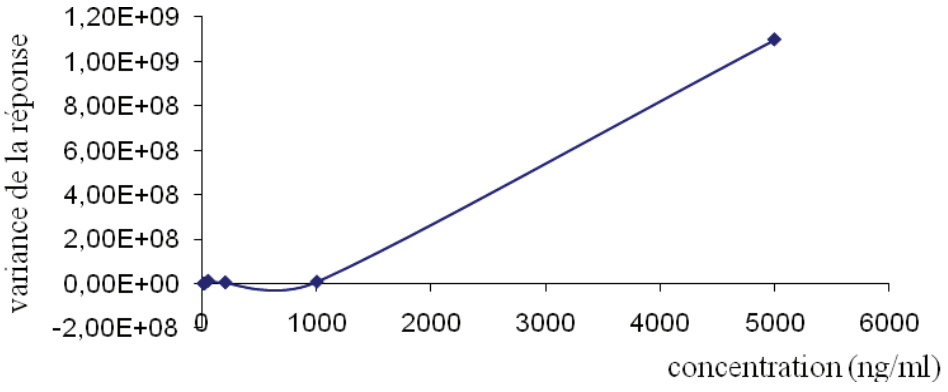


Figure 5 : *évolution de la variance des réponses en fonction des niveaux de concentrations*

La forme de la courbe montre que la variance n'est pas homogène sur l'intervalle de dosage, avec une très forte augmentation pour le niveau de concentration de 5000 ng/ml. Pour

stabiliser cette variance il est possible de pondérer soit par $1/C$ ou par un $1/C^2$ ou tester la transformation logarithmique des données.

La **figure 6** montre la distribution des résidus obtenue par un modèle de régression linéaire pondérée ou non ou appliqué sur des données logarithmiques.

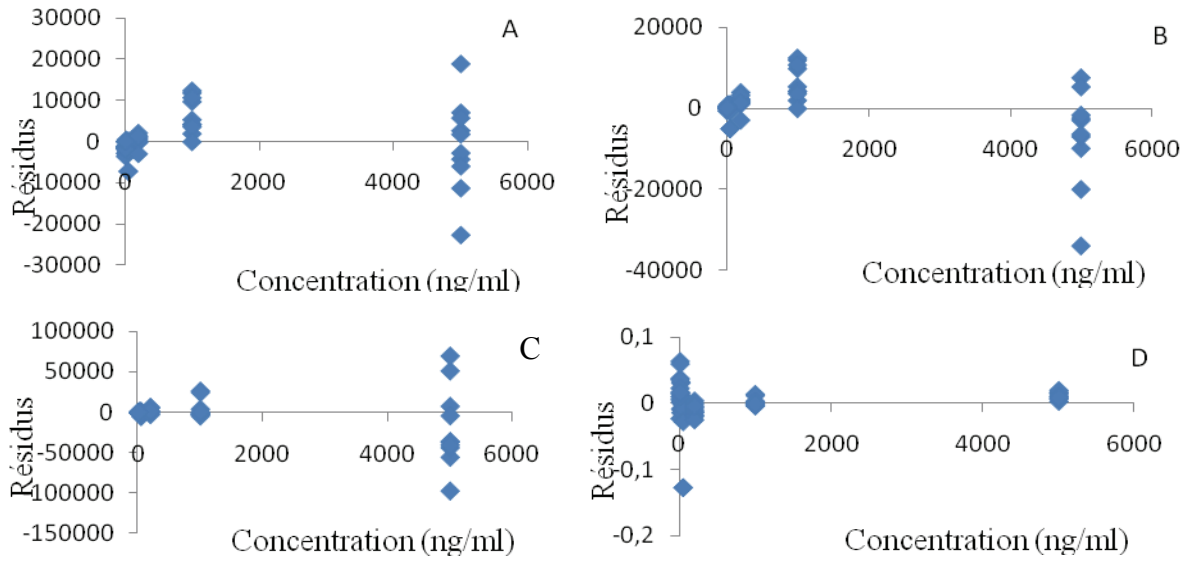
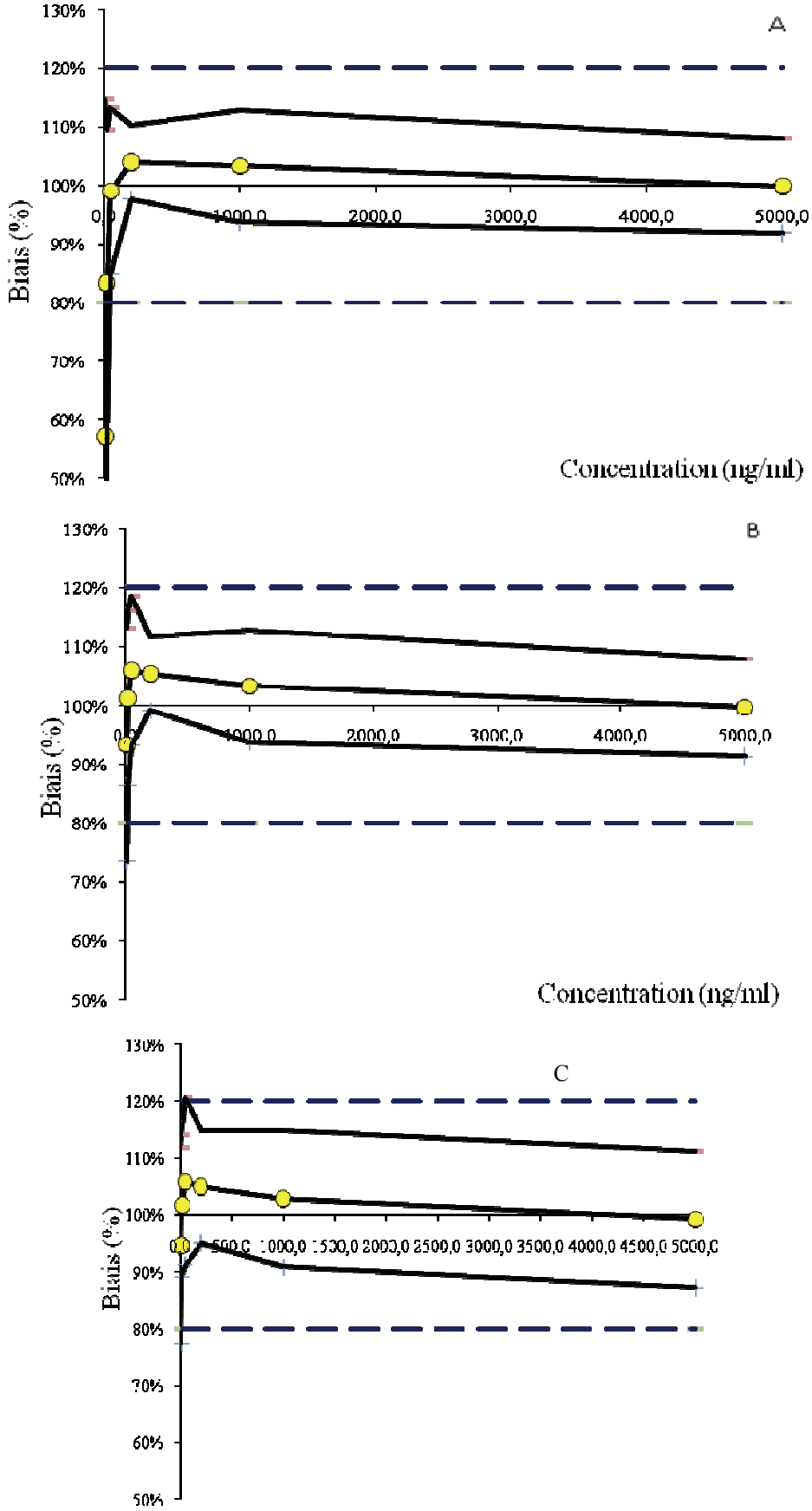


Figure 6 : *évolution des résidus en fonction des concentrations (A) modèle linéaire ; (B) modèle linéaire pondéré (1/C) ; (C) modèle linéaire pondéré ($1/C^2$) ; (D) modèle linéaire appliqué après transformation logarithmique des données*

L'observation des résidus montre que le modèle linéaire simple (**figure 6A**) conduit à une distribution des résidus en forme de trompette. Ceci implique que le modèle ne sera pas très performant. La pondération 1/C (**figure 6B**) diminue cet effet trompette. En revanche la pondération de type $1/C^2$ ne modifie pas la distribution des résidus (**figure 6C**). L'application du modèle linéaire sur les données transformées (**figure 6D**) montre une stabilisation de la variance. Ce modèle pourrait être retenu pour décrire les données.

La **figure 7** montre l'impact de la pondération et de la transformation sur le profil d'exactitude. En utilisant une fonction de réponse linéaire sans pondération ou transformation, les faibles concentrations présentent un biais de plus de 40 % (**figure 7A**). Les profils obtenus après pondération par $1/C$ ou par $1/C^2$ ou après transformation des données sont acceptables. La pondération $1/C$ présente une variabilité moins grande et après transformation logarithmique un moindre biais est observé. Le choix se fera sur l'objectif assigné de la méthode, c'est-à-dire privilégier le biais ou la fidélité.


A
130%
120%
110%
100%
90%
80%
70%
60%
50%
Biais (%)
0,0
1000,0
2000,0
3000,0
4000,0
5000,0
Concentration (ng/ml)
B
130%
120%
110%
100%
90%
80%
70%
60%
50%
Biais (%)
0,0
1000,0
2000,0
3000,0
4000,0
5000,0
Concentration (ng/ml)
C
130%
120%
110%
100%
90%
80%
70%
60%
50%
Biais (%)
0,0
500,0
1000,0
1500,0
2000,0
2500,0
3000,0
3500,0
4000,0
4500,0
5000,0

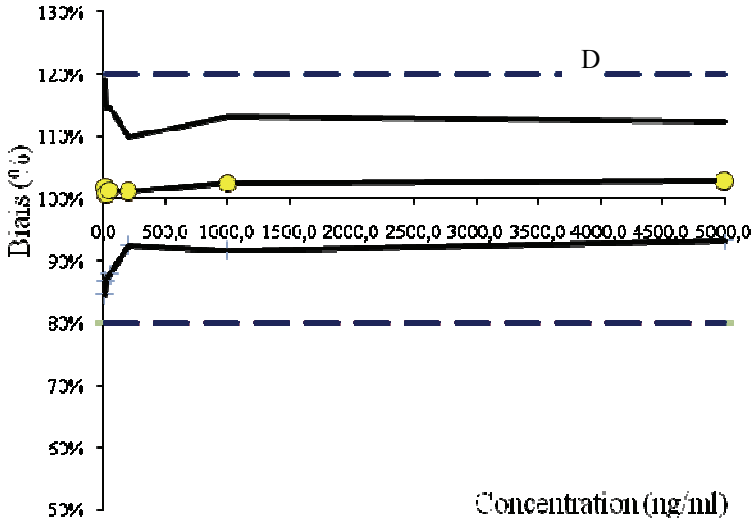


Figure 7 : *profil d'exactitude obtenu avec (A) modèle linéaire, (B) modèle linéaire avec pondération (1/C), (C) modèle linéaire avec pondération ($1C^2$), (D) modèle linéaire après transformation logarithmique des données*

Conclusion

La pondération et / ou la transformation des données sont des outils qui permettent de stabiliser la variance. Elle n'est pas systématiquement nécessaire mais elle est un préalable à l'utilisation de modèle linéaire ou non linéaire, si la variance des réponses n'est pas homogène sur le domaine des concentrations étudiées. Le choix de la pondération ou de la transformation repose sur des aspects graphiques (résidus) ou sur la réalisation de tests statistiques appropriés. Dans certains cas difficiles, il faut avoir recours à la fois à la pondération et à la transformation des données. L'utilisation des ces outils optimise le modèle et ainsi d'obtenir les meilleurs paramètres descriptifs de la relation réponse / concentration.

Bibliographie

Hubert Ph., Chiap P., Crommena J., Boulanger B., Chapuzet E., Mercier N., Bervoas-Martin S., Chevalier P., Grandjean D., Lagorce P., Lallier M., Laparra M.C., Laurentie M., Nivet J.C. (1999) The SFSTP guide on the validation of chromatographic methods for drug bioanalysis: from the Washington Conference to the laboratory. *Analytica Chimica Acta*, 391, 135 -148

Mac Taggart D.L., Farwell S.O. (1992) analysis use of linear regression. Part 1: regression procedures or calibration and quantitation. *J. of the AOAC* 75, 594-614

Tomassone R., Lesquoy E., Miller C. (1983) La régression, nouveaux regards sur une ancienne méthode statistique. Inra, Actualités scientifiques et agronomiques, p. 180