

REVUE DE STATISTIQUE APPLIQUÉE

C. PERRUCHET

G. SADO

Détection de données aberrantes dans le cas d'essais interlaboratoires et fidélité multidimensionnelle

Revue de statistique appliquée, tome 42, n° 1 (1994), p. 81-105

http://www.numdam.org/item?id=RSA_1994__42_1_81_0

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

DÉTECTION DE DONNÉES ABERRANTES DANS LE CAS D'ESSAIS INTERLABORATOIRES ET FIDÉLITÉ MULTIDIMENSIONNELLE

C. Perruchet (1), G. Sado (2)

(1) UTAC, Autodrome, BP 212, F-91311 Monlhery

(2) TOTAL, Centre de Recherches, BP 27, F-76700 Harfleur cedex

RÉSUMÉ

Dans le cas d'essais interlaboratoires où plusieurs laboratoires testent un ou plusieurs produits pour diverses caractéristiques, nous proposons une approche pour la détection de données aberrantes tenant compte de la nature multidimensionnelle des données. En outre une extension multidimensionnelle des concepts de répétabilité et de reproductibilité, quantifiant la fidélité de la méthode d'essai est proposée.

Mots-clés : *Données aberrantes, répétabilité, reproductibilité, fidélité d'une méthode d'essai.*

SUMMARY

For interlaboratory tests, when several laboratories test one or more products for several characteristics, an approach for outliers identification, taking into account the multidimensional nature of the data, is proposed. Moreover a multidimensional generalisation of repeatability and reproducibility, which quantify the precision of a test method is given.

Key-words : *Outliers, repeatability, reproducibility, precision of a test method.*

1. Introduction

Le problème de la détection et du traitement des données dites : aberrantes, douteuses ou étrangères («outliers») est posé par les expérimentateurs depuis de nombreuses années.

En 1755 Boscovich qui tentait d'évaluer l'ellipticité de la terre en moyennant dix valeurs de l'angle entre l'équateur et un des pôles, rejeta deux valeurs extrêmes comme aberrantes.

En revanche, en 1777 Bernouilli qui traitait d'observations astronomiques, concluait que la seule raison valable pour rejeter une donnée, était l'existence d'un incident dans le processus de production de l'information.

Depuis cette date, de nombreuses méthodes de détection des données étrangères ont été proposées. Un ouvrage spécialisé paru en 1984 [1] recensait déjà près de 700 références bibliographiques sur le sujet.

Cette abondance de littérature et l'importance du débat, sont certainement dues au fait que le concept d'«outliers» n'est pas précisément défini, et que les auteurs travaillant sur le sujet ne s'accordent sur aucune définition rigoureuse sur le plan mathématique.

Les données étrangères peuvent être définies comme produites par un phénomène étranger à celui étudié. Toutefois, cette définition ne donne pas les moyens de répondre au problème de leur identification.

S'il est clair qu'il convient de détecter les données répondant à cette définition, pour ne pas déformer les conclusions d'une expérimentation et donc qu'il convient de mettre en œuvre une méthodologie adaptée, il faut aussi souligner qu'une utilisation inadéquate ou abusive des méthodes de détection des données étrangères, conduit aussi à déformer les conclusions d'un essai.

Les travaux menés par l'AFNOR et l'ISO donnent des éléments de réponse au problème de la détection des valeurs aberrantes, par l'usage de tests d'hypothèse, en se limitant toutefois au cas monodimensionnel [5].

Le cas multidimensionnel a cependant déjà fait l'objet de propositions, on en trouvera en particulier un exemple dans [3] où une fonction d'influence est définie à partir du RV estimant la corrélation vectorielle entre variables aléatoires.

L'objet de cet article est de proposer une généralisation au cas multidimensionnel à l'aide de concepts simples à mettre en œuvre.

En outre, les mêmes concepts sont utilisés pour généraliser la fidélité d'une méthode d'essai, quantifiée par sa répétabilité et sa reproductibilité, dans le cas multidimensionnel.

Les données traitées dans cet article, sont des données recueillies lors d'essais interlaboratoires, où K laboratoires testent un ou plusieurs produits en mesurant leur performance pour différentes caractéristiques, généralement en répétant les essais. Les caractéristiques mesurées peuvent être de nature différente (masses, pressions, forces,...) ou de même nature pour des produits différents (le terme d'essai par niveaux est alors généralement employé).

La nature multidimensionnelle des données est donc une conséquence du mesurage de différentes variables ou de mêmes variables à différents niveaux.

Le tableau de données se présente donc de la façon suivante :

		Variables, Niveaux						
Laboratoires	Labo 1							
	LaboK							

Il est à noter que les répétitions des laboratoires sont présentées sur les lignes successives d'une même colonne, afin de pouvoir :

- étudier la dispersion dans chacun des laboratoires (variance intra, répétabilité).
- gérer des nombres différents de répétitions par laboratoire.

Les objectifs de la démarche proposée sont :

- de traiter les données dans leur nature multidimensionnelle afin de ne pas ignorer des phénomènes aberrants en dimension multiple mais non en monodimensionnel [4].

- de permettre d'identifier les zones suspectes du tableau de données, qu'il s'agisse par exemple d'un laboratoire (bloc ligne) ou d'une répétition d'un laboratoire pour une variable ou un niveau (case du tableau), sans avoir à traiter le tableau colonne par colonne et bloc ligne par bloc ligne comme avec les approches monodimensionnelles.

Les avantages de cette démarche sont :

- la simplicité des concepts et des calculs mis en œuvre,
- l'usage de concepts ayant fait leur preuve et largement diffusés grâce aux méthodes d'analyse des données,
- la cohérence avec les concepts classiques de répétabilité et de reproductibilité.

2. Principes généraux

Dans le cas d'un essai interlaboratoires, une situation idéale correspond à une variance intralaboratoire faible et à une variance interlaboratoires faible, ceci entraînant une bonne répétabilité et une bonne reproductibilité.

Une dispersion faible se juge par rapport à l'étendue des niveaux étudiés. Ainsi, une répétabilité de 100 ppm est forte lorsque le niveau est de 200 ppm; une répétabilité de 1000 mPa.s est faible lorsque le niveau est de 300 000 mPa.s.

Des situations suspectes correspondent par exemple :

- à un laboratoire éloigné des autres, même s'il est de faible variance (cf. Figure 1).

- à un laboratoire de forte variance, même s'il est proche des autres (cf. Figure 2).

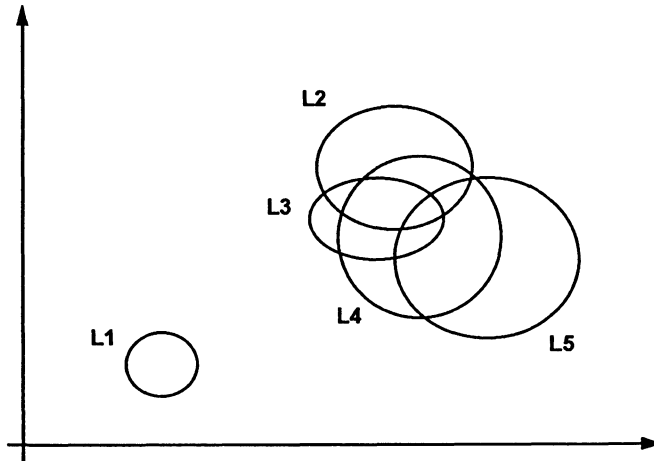


FIGURE 1
*Comparaison de laboratoires
 un laboratoire excentré de faible variance*

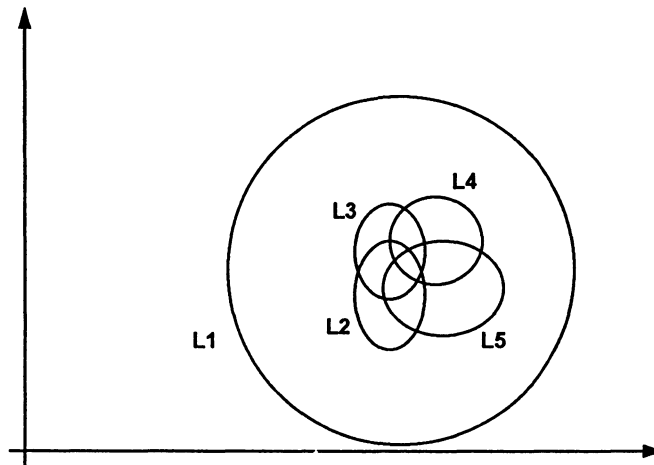


FIGURE 2
*Comparaison de laboratoires
 un laboratoire centré de forte variance*

– ces structures types se généralisent à plus d'un laboratoire, avec la situation extrême où tous les laboratoires sont de fortes variances inégales et sont tous éloignés les uns des autres.

Nous nous proposons donc de définir des indices numériques simples permettant d'identifier ces diverses situations.

Le paragraphe 4 est consacré à l'analyse intralaboratoire et fournit des indices numériques ainsi que des outils graphiques permettant de juger rapidement de la dispersion d'un laboratoire.

Le paragraphe 5 est consacré à l'analyse interlaboratoires et fournit des indices numériques ainsi que des outils graphiques permettant de juger rapidement de l'excentrement d'un laboratoire.

Le paragraphe 6 est consacré à une généralisation des concepts de répétabilité et de reproductibilité au cas multidimensionnel sur la base des éléments introduits précédemment. Il est en particulier montré que la répétabilité et la reproductibilité classiques monodimensionnelles sont des cas particuliers des outils présentés dans cet article.

Le paragraphe 7 est une application numérique basée sur les données utilisées dans la norme NF - ISO 5725.

3. Données

Le tableau de données est noté $X = \{x_i^j; 1 \leq i \leq I, 1 \leq j \leq J\}$

Les J colonnes représentent les caractéristiques mesurées ou les niveaux de l'essai, c'est-à-dire la nature multidimensionnelle du problème.

Les unités statistiques sont les I vecteurs x_i à J composantes. Les descripteurs sont les J vecteurs x^j à I composantes.

Sur l'ensemble des I lignes, une partition est induite par les K laboratoires L_1, \dots, L_K participant à l'essai et présents l_1 à l_K fois par le biais des répétitions, d'où :

$$I = \sum\{l_k; 1 \leq k \leq K\}$$

Afin de se placer dans ce qui sera la pratique dans la grande majorité des applications, l'ensemble des I lignes est muni d'une distance euclidienne, et les masses des lignes sont toutes égales à l'unité.

Toutefois, les développements qui suivent se généralisent sans difficulté à un système de masses quelconques.

Le centre de gravité (ou moyenne) du laboratoire k est noté

$$g_k = \{g_k^j; 1 \leq j \leq J\}$$

avec

$$g_k^j = \sum\{x_i^j; i \in L_K\}/l_k \tag{1}$$

Le centre de gravité (ou moyenne) du nuage de l'ensemble des laboratoires est le vecteur noté g avec :

$$g = \{g^j; 1 \leq j \leq J\} \quad \text{où} \quad g^j = \sum\{x_i^j; 1 \leq i \leq I\}/I \tag{2}$$

$$= \sum\{l_k g_k^j; 1 \leq k \leq K\}/I \tag{3}$$

Dans un espace euclidien avec des masses égales à l'unité, on rappelle que l'inertie totale du nuage par rapport à son centre de gravité g est définie par :

$$M_{T^2} = \Sigma\{d^2(x_i, g); 1 \leq i \leq I\}, \text{ où } d(x_i, g) \text{ désigne la distance entre } x_i \text{ et } g \quad (4)$$

et vérifie : $M_{T^2} = M_{W^2} + M_{B^2}$

où M_{W^2} représente l'inertie intralaboratoire définie par :

$$M_{W^2} = \Sigma\{M_{k^2}; 1 \leq k \leq K\} \quad (5)$$

avec $M_{k^2} = \Sigma\{d^2(x_i, g_k); i \in L_k\}$

et M_{B^2} représente l'inertie interlaboratoires définie par :

$$M_{B^2} = \Sigma\{l_k d^2(g_k, g); 1 \leq k \leq K\} \quad (6)$$

Pour tout ce qui suit, il est fait usage de la distance euclidienne usuelle.

4. Analyse intralaboratoire

L'analyse intralaboratoires est basée sur la décomposition de l'inertie intralaboratoire M_{W^2} en fonction des laboratoires et des variables.

$$\begin{aligned} M_{W^2} &= \Sigma\{M_{k^2}; 1 \leq k \leq K\} \\ &= \Sigma\{M_{W^2}(j); 1 \leq j \leq J\} \quad \text{où} \quad M_{W^2}(j) = \Sigma\{M_{k^2}(j); 1 \leq k \leq K\} \end{aligned}$$

Pour chacun des K laboratoires, l'inertie par rapport à son centre de gravité, notée M_{k^2} est calculée :

$$\begin{aligned} M_{k^2} &= \Sigma\{d^2(x_i, g_k); i \in L_k\} \\ &= \Sigma\{(x_i^j - g_k^j)^2; i \in L_k, 1 \leq j \leq J\} \\ &= \Sigma\{M_{k^2}(j); 1 \leq j \leq J\} \end{aligned} \quad (7)$$

où $M_{k^2}(j) = \Sigma\{(x_i^j - g_k^j)^2; i \in L_k\}$ représente la contribution absolue de la variable j à l'inertie du laboratoire k .

La contribution relative du laboratoire k à l'inertie intralaboratoire est définie par :

$$CTW_k = M_{k^2}/M_{W^2} \quad (8)$$

Cet indice vérifie : $\Sigma\{CTW_k; 1 \leq k \leq K\} = 1$

Les laboratoires dispersés sont ceux dont les inerties intra ou les contributions CTW_k sont les plus élevées.

On construira donc un diagramme en bâton du type représenté en figure 3, pour juger rapidement des dispersions relatives des laboratoires en présentant les laboratoires par contribution décroissante.

Si les dispersions des laboratoires sont du même ordre de grandeur, leur inertie intra commune est égale à l'inertie intra moyenne et vaut M_{W^2}/K . Leur contribution relative commune vaut alors $1/K$.

L'ordonnée du graphique présentera donc les valeurs $1/K$ et $2/K$ au-delà de laquelle les laboratoires fortement dispersés apparaîtront.

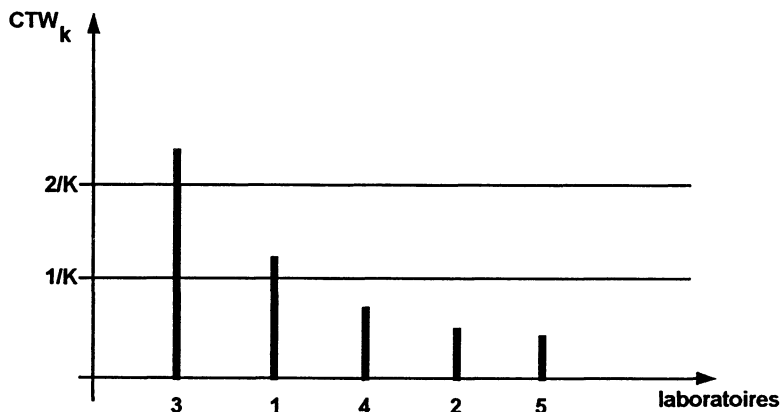


FIGURE 3
Analyse intralaboratoire
identification des laboratoires de forte variance

Pour les laboratoires de forte inertie, l'inertie sera décomposée en calculant les contributions relatives :

$$CTW_k(j) = M_{k^2}(j)/M_{k^2} \tag{9}$$

Cet indice vérifie : $\forall k : \sum\{CTW_k(j); 1 \leq j \leq J\} = 1$

Une représentation graphique du même type permettra pour chacun des laboratoires d'identifier rapidement les variables engendrant une forte inertie du laboratoire.

Il est à noter que si, pour un laboratoire donné, les variables ont la même influence les $M_{k^2}(j)$ ont pour valeur commune, la valeur moyenne égale à M_{k^2}/J .

Leur contribution relative commune (*i.e.* la contribution relative moyenne) vaut alors : $1/J$.

L'ordonnée du graphique associé à chaque laboratoire fera donc apparaître la valeur $1/J$ et la valeur $2/J$ au-delà de laquelle les variables hautement significatives apparaîtront (*cf.* Figure 4).

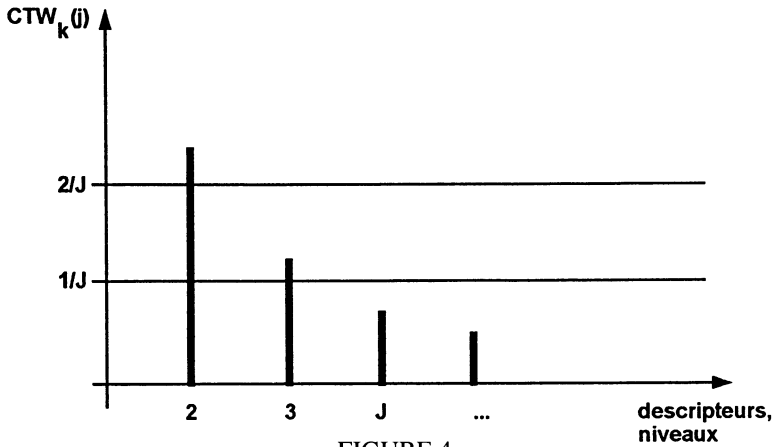


FIGURE 4
Analyse intralaboratoire
identification des variables engendrant une forte variance

5. Analyse interlaboratoires

L'analyse interlaboratoires est basée sur la décomposition de l'inertie interlaboratoires M_{B^2} en fonction des laboratoires et des variables.

$$\begin{aligned}
 M_{B^2} &= \Sigma\{l_k d^2(g_k, g); 1 \leq k \leq K\} \\
 &= \Sigma\{l_k (g_k^j - g^j)^2; 1 \leq k \leq K, 1 \leq j \leq J\} \\
 &= \Sigma\{M_{B^2}(j); 1 \leq j \leq J\}
 \end{aligned} \tag{10}$$

$l_k d^2(g_k, g)$ représente la contribution absolue du laboratoire k à l'inertie interlaboratoires.

Cette quantité se décompose comme la somme des $M_{B_k^2}(j) = l_k (g_k^j - g^j)^2$

$M_{B^2}(j) = \Sigma\{l_k (g_k^j - g^j)^2; 1 \leq k \leq K\}$ représente la contribution absolue de la variable j à l'inertie interlaboratoires.

La contribution relative du laboratoire k à l'inertie interlaboratoires est définie par :

$$CTB_k = l_k d^2(g_k, g) / M_B^2 \tag{11}$$

Cet indice vérifie : $\Sigma\{CTB_k; 1 \leq k \leq K\} = 1$

Les laboratoires excentrés sont ceux dont les contributions sont les plus élevées. On construira donc un diagramme en bâton du type présenté en figure 5 pour juger rapidement des excentrement relatifs des laboratoires.

Si les excentrement des laboratoires sont du même ordre de grandeur, leur contribution absolue commune est égale à la contribution absolue moyenne et vaut M_B^2 / K .

Leur contribution relative commune vaut alors $1/K$.

L'ordonnée du graphique présentera donc les valeurs $1/K$ et $2/K$ au-delà de laquelle les laboratoires fortement excentrés apparaîtront.

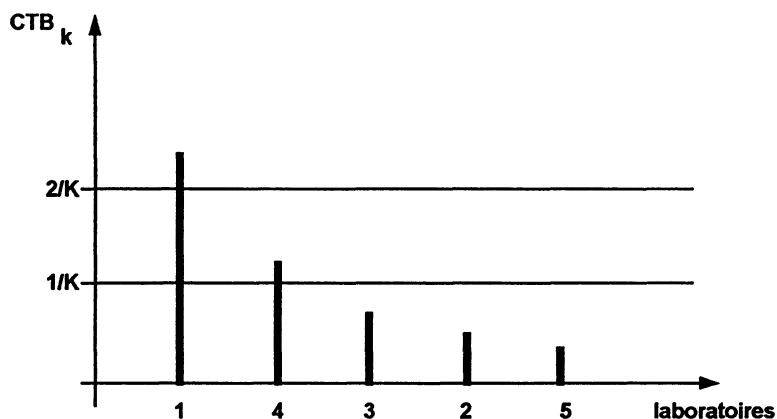


FIGURE 5
Analyse interlaboratoires
identification des laboratoires fortement excentrés

Pour un laboratoire, la contribution relative de la variable j au carré de sa distance à la moyenne des laboratoires est définie par :

$$CTB_k(j) = (g_k^j - g^j)^2 / d^2(g_k, g) \quad (12)$$

Cet indice vérifie : $\forall k : \sum\{CTB_k(j); 1 \leq j \leq J\} = 1$

Un diagramme en bâton permettra de vérifier rapidement si l'excentrement d'un laboratoire est dû à l'influence d'une variable particulière ou bien à plusieurs, voire à l'ensemble des variables.

On répondra par ce moyen à l'identification de laboratoires aberrants ne pouvant être identifiés par des méthodes monodimensionnelles. Il s'agit là de la situation classique et illustrée sur la figure 6, où le laboratoire L clairement aberrant ne l'est ni pour l'axe 1 seul, ni pour l'axe 2 seul.

La figure 7 illustre l'application numérique de ces concepts à une situation simple pour 4 laboratoires notés A, B, C, D.

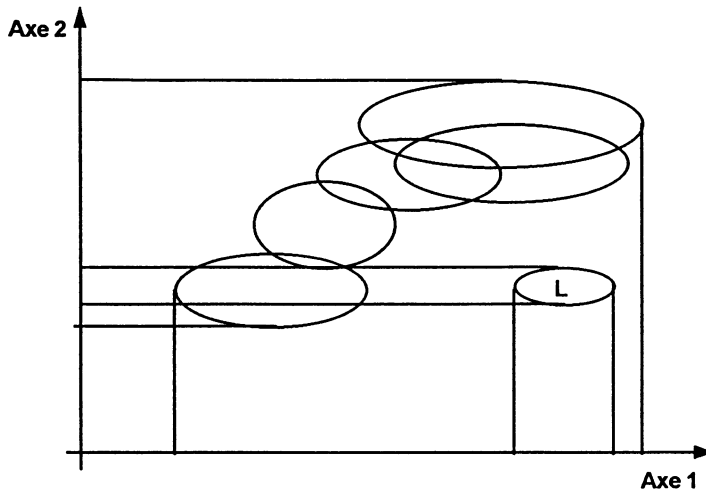
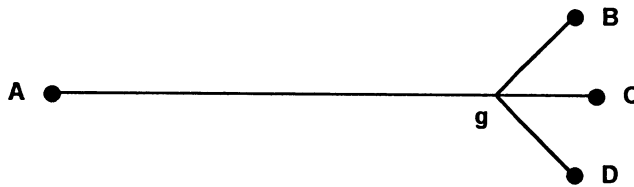


FIGURE 6
Analyse interlaboratoires
exemple de la nécessité de l'approche multidimensionnelle



$$m_A = m_B = m_C = m_D = 1$$

$$d^2(A,g) = 7 \quad d^2(B,g) = d^2(C,g) = d^2(D,g) = 1$$

$$CTB_A = 0,7 \quad CTB_B = CTB_C = CTB_D = 0,1$$

FIGURE 7
Analyse interlaboratoires
exemple de calcul de contributions

6. Fidélité multidimensionnelle

6.1 Répétabilité multidimensionnelle

La répétabilité r , mesure de l'écart probable entre deux mesurages sur des objets de même nature, dans un même laboratoire, sous des conditions opératoires semblables, est définie normativement [5] dans le cas monodimensionnel par :

$$P(|X_1 - X_2| < r) = p$$

où p est une probabilité élevée, généralement égale à 0,95 et X_1, X_2 deux variables indépendamment et identiquement distribuées (*iid*) gaussiennes d'espérance et variance inconnues.

$$\text{Dans ce cas, on obtient : } r = 1,96\sqrt{2}\sigma \approx 2,77\sigma \quad (13)$$

La répétabilité est utilisée pour quantifier la fidélité de la méthode d'essai (*cf.* [5]).

Cependant, dans le cas où plusieurs caractéristiques du produit sont mesurées, ou bien que l'essai est réalisé pour plusieurs niveaux, aucune répétabilité globale n'est définie.

Il s'agit donc, disposant de deux vecteurs aléatoires *iid* gaussiens, à valeur dans R^J , d'espérance M et de matrice de covariance V , de rechercher le scalaire r tel que :

$$P(d(X_1, X_2) < r) = p$$

où d est une distance donnée.

Il pourra s'agir, par exemple, de d_1 et d_2 associés aux deux normes usuelles L_1 et L_2 :

$$\begin{aligned} d_1(X_1, X_2) &= \Sigma\{|X_1^j - X_2^j|; 1 \leq j \leq J\} \\ d_2(X_1, X_2) &= (\Sigma\{(X_1^j - X_2^j)^2; 1 \leq j \leq J\})^{1/2} \end{aligned}$$

d_1 et d_2 sont équivalentes pour $J = 1$ et correspondent à la définition normative de la répétabilité. En revanche, seul le cas de d_2 sera traité par la suite.

D'après les hypothèses sur X_1 et X_2

$$\forall j : (X_1^j - X_2^j) \sim N(0, 2V_{jj})$$

où V_{jj} est le j ème terme diagonal de V , *i.e.* la variance commune de X_1^j et X_2^j .

Alors, $\forall j : (X_1^j - X_2^j)^2 / (2V_{jj}) \sim \chi_1^2$ loi du chi-deux à 1 degré de liberté.

Ainsi, si la matrice V est diagonale :

$$\Sigma\{(X_1^j - X_2^j)^2 / (2V_{jj}); 1 \leq j \leq J\} \sim \chi_J^2$$

et si les termes diagonaux de V sont constants (soit $V = \sigma^2 Id$) :

$$d_2^2(X_1, X_2) \sim 2\sigma^2 \chi_J^2$$

Alors

$$\begin{aligned} p &= P(d_2(X_1, X_2) < r) \\ &= P(d_2^2(X_1, X_2) < r^2) \\ &= P(D < r^2 / (2\sigma^2)) \quad \text{où } D \sim \chi_J^2 \end{aligned}$$

$r^2/(2\sigma^2)$ est donc égal au quantile d'ordre p d'une loi de chi-deux à J degrés de liberté noté $\chi_J^2(p)$,

$$\text{d'où } r = \sqrt{\chi_J^2(p)}\sqrt{2}\sigma \quad (14)$$

Il est alors trivial de constater que pour $J = 1$, par définition de la loi du chi-deux, $\chi_J^2(p) = \chi_1^2(0,95) \approx 3,84 \approx 1,96^2$.

La répétabilité d'un laboratoire dans le cas multidimensionnel est donc définie et est cohérente avec la définition normalisée.

Son estimation l'est aussi et peut être définie par :

$$\forall J \geq 1 : \hat{r} = \sqrt{\chi_J^2(p)}\sqrt{2}\hat{\sigma}$$

où $\hat{\sigma}$ est un estimateur de σ .

Dans le cadre du modèle utilisé, l'estimateur de la variance choisi est celui du maximum de vraisemblance. En effet, l'usage de l'estimateur sans biais de variance minimum conduit à des expressions des fidélités peu satisfaisantes en termes d'homogénéité par rapport à l'équation de décomposition de l'inertie.

Il est bien connu que rien ne permet de privilégier systématiquement l'un de ces deux estimateurs qui sont par ailleurs tous deux presque sûrement convergents.

Toutefois, la norme NF-ISO présente l'usage de l'estimateur sans biais sans le justifier.

En pratique il nous semble simplement nécessaire de préciser quel estimateur de la variance est utilisé, qu'il s'agisse de l'un des estimateurs précédents ou bien d'un estimateur issu d'un modèle linéaire adapté aux données traitées.

L'estimation de σ peut ici être faite dans quatre situations différentes conduisant à quatre estimations de la répétabilité au moyen de la variance de répétabilité.

– la répétabilité monodimensionnelle pour un laboratoire (j et k fixés) à l'aide de :

$$\hat{\sigma}_k^2(j) = \Sigma\{(x_i^j - g_k^j)^2; i \in L_k\}/l_k \quad (15)$$

– la répétabilité multidimensionnelle pour un laboratoire (k fixé) à l'aide de :

$$\begin{aligned} \hat{\sigma}_k^2 &= \Sigma\{\hat{\sigma}_k^2(j); 1 \leq j \leq J\}/J \\ &= M_k^2/(J.l_k) \end{aligned} \quad (16)$$

– la répétabilité monodimensionnelle commune à tous les laboratoires (j fixé) à l'aide de :

$$\begin{aligned} \hat{\sigma}^2(j) &= \Sigma\{l_k \hat{\sigma}_k^2(j); 1 \leq k \leq K\}/\Sigma\{l_k; 1 \leq k \leq K\} \\ &= \Sigma\{M_k^2(j); 1 \leq k \leq K\}/I \end{aligned} \quad (17)$$

– la répétabilité multidimensionnelle commune à tous les laboratoires à l'aide de :

$$\begin{aligned}\hat{\sigma}^2 &= \Sigma\{l_k \hat{\sigma}_k^2; 1 \leq k \leq K\} / \Sigma\{l_k; 1 \leq k \leq K\} \\ &= \Sigma\{\hat{\sigma}^2(j); 1 \leq j \leq J\} / J \\ &= M_W^2 / (IJ)\end{aligned}\tag{18}$$

Dans chaque cas l'estimation de la répétabilité est donc définie par :

$$\begin{aligned}\forall J \geq 1; \forall K \geq 1 : \hat{r} &= \sqrt{\chi_J^2(p)} \sqrt{2} \hat{\sigma} \\ &= (2\chi_J^2(p) M_W^2 / (IJ))^{1/2}\end{aligned}$$

Soit pour $J = 1$, et un laboratoire donné, en notant $N(p)$ le quantile d'ordre p d'une loi gaussienne centrée réduite :

$$\begin{aligned}\hat{r} &= N\left(\frac{1+p}{2}\right) (2\Sigma\{(x_i - g_k)^2; i \in L_K\} / l_k)^{1/2} \\ &\approx 2,77 (\Sigma\{(x_i - g_k)^2; i \in L_K\} / l_k)^{1/2} \quad \text{pour } p = 0,95\end{aligned}$$

Il convient donc de noter, que les concepts utilisés pour l'analyse intralaboratoire au paragraphe 4, sont parfaitement homogènes avec le concept de répétabilité pris aussi bien dans sa définition normative que dans la généralisation multidimensionnelle proposée.

6.2 Reproductibilité multidimensionnelle

La reproductibilité R , mesure de l'écart probable entre deux mesurages sur des objets de même nature, dans des laboratoires différents, est définie normativement [5] dans le cas monodimensionnel par :

$$P(|X_1 - X_2| < R) = p$$

où p est une probabilité élevée, généralement égale à 0,95 et X_1, X_2 deux variables iid gaussiennes d'espérance et variance inconnues.

$$\text{Dans ce cas on obtient : } R = 1,96\sqrt{2}\sigma_R \approx 2,77\sigma_R\tag{19}$$

La reproductibilité quantifie la fidélité de la méthode d'essai appliquée dans des laboratoires différents.

La variance de reproductibilité σ_R^2 se décompose dans le cadre d'un modèle linéaire à effets aléatoires en :

$$\sigma_R^2 = \sigma_L^2 + \sigma_r^2\tag{20}$$

σ_L^2 représentant la composante de la variance de l'effet interlaboratoires.

σ_r^2 représentant la composante de l'effet intralaboratoire (variance résiduelle du modèle), estimée au §6.1.

Dans une situation multidimensionnelle, en reprenant les mêmes hypothèses pour X_1 et X_2 qu'au paragraphe précédent, on obtient :

$$R = \sqrt{\chi_J^2(p)}\sqrt{2}\sigma_R \quad (21)$$

Cette définition est cohérente avec la définition monodimensionnelle de la norme NF-ISO.

L'estimation de σ_R^2 revient à l'estimation de σ_r^2 présentée ci-dessus et à celle de σ_L^2 .

Comme dans le cas de la répétabilité l'estimation de σ_L^2 peut être présentée dans deux situations :

– la reproductibilité monodimensionnelle (j fixé) :

$$\begin{aligned} \hat{\sigma}_L^2(j) &= \Sigma\{l_k(g_k^j - g^j)^2; 1 \leq k \leq K\} / \Sigma\{l_k; 1 \leq k \leq K\} \\ &= M_B^2(j) / I \end{aligned} \quad (22)$$

– la reproductibilité multidimensionnelle :

$$\begin{aligned} \hat{\sigma}_L^2 &= \Sigma\{\hat{\sigma}_L^2(j); 1 \leq j \leq J\} / J \\ &= M_B^2 / (IJ) \end{aligned} \quad (23)$$

L'estimation multidimensionnelle de la variance de reproductibilité s'écrit donc :

$$\begin{aligned} \hat{\sigma}_R^2 &= \hat{\sigma}_L^2 + \hat{\sigma}_r^2 = (M_B^2 + M_W^2) / (IJ) \\ &= M_T^2 / (IJ) \end{aligned} \quad (24)$$

Soit une définition généralisant la définition normative de la reproductibilité.

Il convient donc de noter, que les concepts utilisés pour l'analyse interlaboratoires au paragraphe 5, sont parfaitement homogènes avec le concept de reproductibilité pris aussi bien dans sa définition normative, que dans la généralisation multidimensionnelle proposée.

7. Application

Dans ce paragraphe, nous allons illustrer la méthode exposée en utilisant le jeu de données de la norme NF ISO 5725, données portant sur le dosage de la créosote.

Nous avons fait ce choix pour que le lecteur puisse comparer la méthode multidimensionnelle que nous avons développée aux tests monodimensionnels de Dixon et de Cochran de la norme NF-ISO.

7.1 Présentation des données

Elles sont présentées sur les lignes L1 à L9 du tableau 1, sous forme d'un tableau à 5 colonnes : 5 niveaux de teneur en créosote et 18 lignes représentant les analyses répétées 2 fois par 9 laboratoires.

$$J = 5$$

$$I = 18$$

$$K = 9$$

$$l_k = 2, 1 \leq k \leq 9,$$

7.2 Analyse préliminaire

Elle a pour but d'analyser l'inertie totale M_T^2 en inertie intralaboratoire M_W^2 et inertie interlaboratoire M_B^2 .

7.2.1 Calcul du centre de gravité de chaque laboratoire

Il s'agit ici d'un centre de gravité multidimensionnel, *i.e.* d'un vecteur à J composantes.

On applique la formule (1) pour obtenir les ligne g_1 à g_9 du tableau 1.

Ainsi pour le laboratoire 1 :

$$g_1 = \{4, 415; 9, 34; 17, 15; 19, 23; 24, 14\}$$

7.2.2 Calcul du centre de gravité du nuage de points

Il s'agit là aussi d'un vecteur à J composantes où les composantes se calculent de deux façons possibles à l'aide des formules (2) ou (3).

g se trouve au milieu du tableau 1.

$$g = \{3, 993; 8, 399; 14, 508; 15, 993; 20, 511\}$$

Les composantes de g sont les moyennes par niveau.

7.2.3 Calcul de l'inertie totale du nuage par rapport à son centre de gravité g

On applique la formule (4).

M_T^2 est la somme des carrés des distances des vecteurs x_i au vecteur g , les x_i et g ayant J composantes.

$$M_T^2(j) = \Sigma\{(x_i^j - g^j)^2; 1 \leq i \leq I\}$$
 représente l'inertie de la variable j et

$$M_T^2 = \Sigma\{M_T^2(j); 1 \leq j \leq J\}$$

$$M_T^2 = 0, 818 + 5, 490 + 18, 087 + 28, 379 + 50, 798 = 103, 572$$

TABLEAU I

***** DONNEES *****					
L1	4,44	9,34	17,4	19,23	24,28
L1	4,39	9,34	16,9	19,23	24
L2	4,03	8,42	14,42	16,06	20,4
L2	4,23	8,33	14,5	16,22	19,91
L3	3,7	7,6	13,6	14,5	19,3
L3	3,7	7,4	13,6	15,1	19,7
L4	4,1	8,93	14,6	15,6	20,3
L4	4,1	8,8	14,2	15,5	20,3
L5	3,97	7,89	13,73	15,54	20,53
L5	4,04	8,12	13,92	15,78	20,88
L6	3,75	8,76	13,9	16,42	18,56
L6	4,03	9,24	14,06	16,58	16,58
L7	3,7	8	14,1	14,9	19,7
L7	3,8	8,3	14,2	16	20,5
L8	3,91	8,04	14,84	15,41	21,1
L8	3,9	8,07	14,84	15,22	20,78
L9	4,02	8,44	14,24	15,14	20,71
L9	4,07	8,17	14,1	15,44	21,66

g1 j	4,415	9,34	17,15	19,23	24,14
g2 j	4,13	8,375	14,46	16,14	20,155
g3 j	3,7	7,5	13,6	14,8	19,5
g4 j	4,1	8,865	14,4	15,55	20,3
g5 j	4,005	8,005	13,825	15,66	20,705
g6 j	3,89	9	13,98	16,5	17,57
g7 j	3,75	8,15	14,15	15,45	20,1
g8 j	3,905	8,055	14,84	15,315	20,94
g9 j	4,045	8,305	14,17	15,29	21,185
g j	3,9933	8,3994	14,5083	15,9928	20,5106

MT ² j	0,818	5,4901	18,0868	28,3794	50,7979	MT ² =	103,5722
-------------------	-------	--------	---------	---------	---------	-------------------	----------

Mk ² j	0,00125	0	0,125	0	0,0392	Mk ²	0,16545
	0,02	0,00405	0,0032	0,0128	0,12005		0,1601
	0	0,02	0	0,18	0,08		0,28
	0	0,00845	0,08	0,005	0		0,09345
	0,00245	0,02645	0,01805	0,0288	0,06125		0,137
	0,0392	0,1152	0,0128	0,0128	1,9602		2,1402
	0,005	0,045	0,005	0,605	0,32		0,98
	5E-05	0,00045	0	0,01805	0,0512		0,06975
	0,00125	0,03645	0,0098	0,045	0,45125		0,54375

MV ² j	0,0692	0,2561	0,2539	0,9074	3,0831	MV ² =	4,5697
-------------------	--------	--------	--------	--------	--------	-------------------	--------

MB ² j	0,7488	5,234	17,833	27,4719	47,7147	MB ² =	99,0025
-------------------	--------	-------	--------	---------	---------	-------------------	---------

7.2.4 Calcul de l'inertie intralaboratoire

On applique la formule (7).

En bas du tableau 1, le sous-tableau des $M_k^2(j)$ présente pour chaque niveau j (5 colonnes) et chaque laboratoire k (9 lignes), le carré de la distance des résultats d'un laboratoire à sa moyenne.

L'inertie intralaboratoire M_W^2 est obtenue de deux façons possibles :

– c'est la somme pour les J niveaux et les K laboratoires des $M_k^2(j)$

– on peut aussi sommer d'abord sur les différents j , on obtient les K valeurs M_k^2 puis on applique la formule (5) :

$$M_W^2 = 4,5697$$

7.2.5 Calcul de l'inertie interlaboratoire

On applique les formules (6) et (10) où les g_k et g sont des vecteurs à J composantes.

On calcule d'abord $M_B^2(j)$ puis M_B^2 .

$$M_B^2 = 99,0025$$

7.2.6 Analyse de l'inertie totale

On retrouve bien la décomposition :

$$M_T^2 = M_W^2 + M_B^2$$

$$103,5722 = 4,5697 + 99,0025$$

7.3 Analyse intralaboratoire

On calcule l'inertie M_k^2 de chaque laboratoire par rapport à son centre de gravité par application de la formule (7).

Ces calculs nous ont permis d'obtenir M_W^2 au §7.2.4. et le vecteur M_K^2 à K composantes est présenté au tableau 1.

$$M_K^2 = \{0,16545; 0,1601; 0,2800; 0,09345; 0,1370; 2,1402; 0,9800; 0,06975; 0,54375\}$$

La contribution relative du laboratoire k à l'inertie intralaboratoire CTW_k est donnée par la formule (8).

L'ensemble des contributions relatives $CTW_k(j)$ (cf. formule (9)) est présenté au tableau 2.

La représentation graphique en figure 8 montre les contributions comparées de chacun des laboratoires.

TABLEAU 2
analyse intra et inter-labo

CTW1 j	0,007555	0	0,755515	0	0,23693	CTWk	0,0362
CTW2 j	0,124922	0,025297	0,019988	0,07995	0,749844		0,035
CTW3 j	0	0,071429	0	0,642857	0,285714		0,0613
CTW4 j	0	0,090423	0,856073	0,053505	0		0,0204
CTW5 j	0,017883	0,193066	0,131752	0,210219	0,44708		0,03
CTW6 j	0,018316	0,053827	0,005981	0,005981	0,915896		0,4683
CTW7 j	0,005102	0,045918	0,005102	0,617347	0,326531		0,2145
CTW8 j	0,000717	0,006452	0	0,258781	0,73405		0,0153
CTW9 j	0,002299	0,067034	0,018023	0,082759	0,829885		0,119

MBk ² j	0,35561	1,76929	13,95684	20,95919	26,34567	63,3866
	0,03736	0,0012	0,00467	0,04335	0,25285	0,3394
	0,17209	1,618	1,65013	2,84545	2,04246	8,3281
	0,02276	0,43348	0,02347	0,39211	0,08867	0,9605
	0,00027	0,31117	0,93388	0,22149	0,07561	1,5424
	0,02136	0,72134	0,55827	0,51454	17,29379	19,1093
	0,11842	0,12444	0,2568	0,58922	0,33712	1,426
	0,01561	0,23728	0,22001	0,91877	0,36884	1,7605
	0,00534	0,01784	0,22893	0,9878	0,90974	2,1497
	MB ² j	0,74882	5,23404	17,833	27,47192	47,71475

CTB1 j	0,00561	0,027913	0,220186	0,330656	0,415635	CTBk	0,6403
CTB2 j	0,110077	0,003536	0,01376	0,127725	0,744991		0,0034
CTB3 j	0,020664	0,194282	0,19814	0,341669	0,245249		0,0841
CTB4 j	0,023696	0,451307	0,024435	0,408235	0,092317		0,0097
CTB5 j	0,000175	0,201744	0,605472	0,143601	0,049021		0,0156
CTB6 j	0,001118	0,037748	0,029215	0,026926	0,904993		0,193
CTB7 j	0,083043	0,087265	0,180084	0,413198	0,23641		0,0144
CTB8 j	0,008867	0,13478	0,12497	0,52188	0,209509		0,0178
CTB9 j	0,002484	0,008299	0,106494	0,459506	0,423194		0,0217

Si tous les laboratoires avaient une même contribution relative, celle-ci vaudrait $1/K$, soit dans notre exemple 0,11.

A l'évidence le laboratoire 6 et, dans une moindre mesure, le laboratoire 7 ont une très forte contribution à l'inertie intralaboratoire : ils ont une très médiocre répétabilité.

On peut alors se demander pour quels niveaux ce phénomène est très marqué.

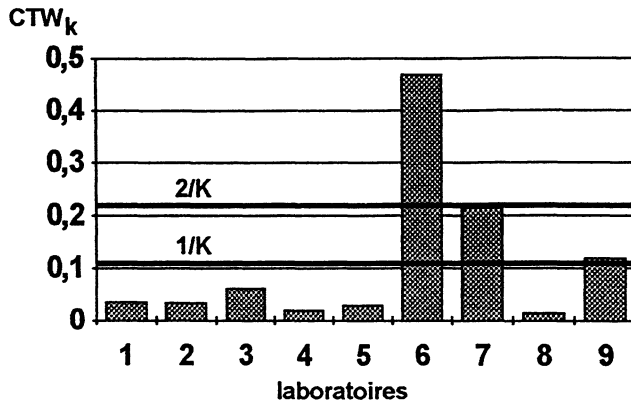


FIGURE 8
Analyse intralaboratoire
données CREOSOTE (NF ISO 5725)

Le graphique de la figure 9 représente les contributions $CTW_k(j)$ à l'inertie intralaboratoire des laboratoires 6 et 7 pour les 5 niveaux.

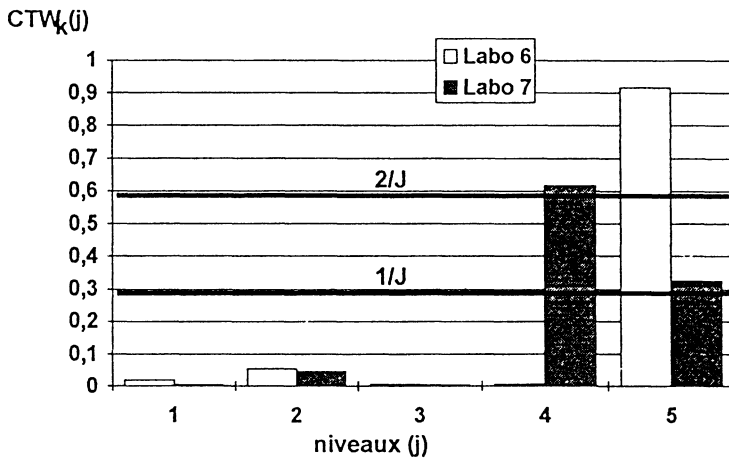


FIGURE 9
Analyse intralaboratoire
données CREOSOTE labos 6 & 7

On y voit clairement que :

- le laboratoire 6 a «raté» la mesure au niveau $j = 5$, cela peut être considéré comme un accident de parcours. Ses résultats sont écartés pour ce seul niveau.
- le laboratoire 7 a une forte contribution pour les niveaux 4 et 5. En conséquence, ce laboratoire devrait peut être revoir sa procédure d'application de la méthode de mesure de la teneur en créosote.

7.4 Analyse interlaboratoire

La décomposition de l'inertie interlaboratoire M_B^2 se fait selon les laboratoires et les niveaux.

On applique la formule (10). Les résultats numériques sont présentés au tableau 2.

On y trouve ensuite les contributions $CTB_K(j)$ à l'inertie interlaboratoire pour chaque laboratoire et chaque niveau.

La figure 10 représente la contribution CTB_K de chacun des laboratoires à l'inertie interlaboratoire.

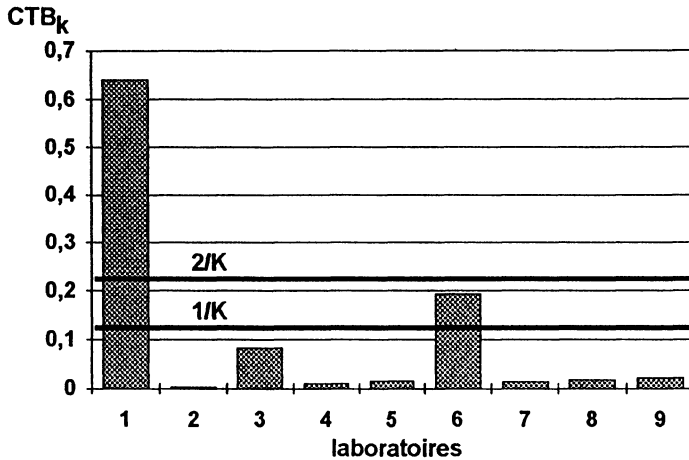


FIGURE 10
Analyse interlaboratoire
données CREOSOTE (NF ISO 5725)

Si les laboratoires avaient des contributions égales elles vaudraient $0,11 (= 1/K)$.

On voit donc que le laboratoire 1 a une contribution considérable à l'inertie interlaboratoire, et, dans une moindre mesure, le laboratoire 6.

Ces laboratoires seront responsables d'une reproductibilité importante.

La figure 11 montre que :

– le laboratoire 1 a une contribution à l'inertie très importante pour les niveaux $j = 3, 4$ et 5 .

Il doit donc être écarté car il fournit des résultats trop différents des autres laboratoires : ses résultats sont beaucoup trop élevés.

– le laboratoire 6 a fourni des résultats trop différents pour le niveau 5. On avait déjà vu au §7.3 que pour ce niveau, ses résultats étaient beaucoup trop dispersés. La valeur 16,58 trop faible peut être responsable de tout cela. Ces résultats ont déjà été écartés.

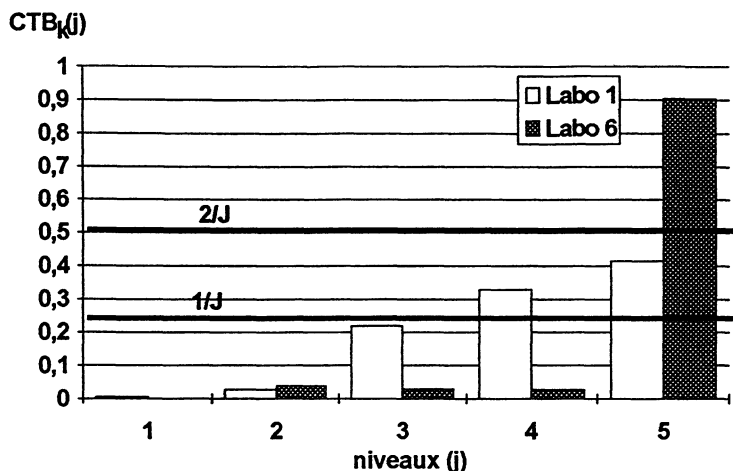


FIGURE 11
Analyse interlaboratoire
données CREOSOTE labos 1 & 6

7.5 Fidélité multidimensionnelle

7.5.1 Répétabilité multidimensionnelle

L'analyse de la variance permet d'atteindre la répétabilité correspondant à quatre situations différentes.

– Répétabilité monodimensionnelle (*j* et *k* fixés). L'application des formules (13) et (15) permet de dresser le début du tableau 3 qui fournit une estimation de la répétabilité pour chaque laboratoire à chaque niveau.

– Répétabilité multidimensionnelle (*k* fixé).

En marge de droite et en haut du tableau 3, on trouve pour chaque laboratoire la répétabilité multidimensionnelle (tous niveaux confondus) par application des formules (14) et (16).

<i>k</i>	\hat{r}_k
1	0,605
2	0,595
3	0,787
4	0,455
5	0,551
6	2,177
7	1,473
8	0,393
9	1,097

TABLEAU 3

Répétabilité monodim. pour un labo. (j et k fixés)						Répét. multidim. (k fixé)	
rk j	0,069	0	0,693	0	0,388	rk	0,605
	0,277	0,125	0,111	0,222	0,679		0,595
	0	0,277	0	0,831	0,554		0,787
	0	0,18	0,554	0,138	0		0,455
	0,097	0,319	0,263	0,332	0,485		0,551
	0,388	0,665	0,222	0,222	2,742		2,177
	0,139	0,416	0,139	1,524	1,108		1,473
	0,014	0,042	0	0,263	0,443		0,393
	0,069	0,374	0,194	0,416	1,316		1,097

Répétabilité monodim. pour tous les labos						Répét. multidim.	
r j	0,172	0,33	0,329	0,622	1,146	r	1,06

calcul des variances de reproductibilité						
V ² L	0,0416	0,290778	0,990722	1,526217	2,650817	
V ² r	0,003856	0,014193	0,014107	0,050422	0,171163	
V ² R	0,045456	0,304971	1,004829	1,576639	2,82198	

Reproductibilité monodimensionnelle						Repro. multidim.	
R j	0,591	1,53	2,777	3,478	4,653	R	5,048

– Répétabilité monodimensionnelle commune à tous les laboratoires.

Elle est obtenue par application des formules (13) et (17) et se trouve en marge inférieure du sous-tableau des répétabilités.

j	\widehat{r}_j
1	0,172
2	0,330
3	0,329
4	0,622
5	1,146

– Répétabilité multidimensionnelle commune à tous les laboratoires

Elle s’obtient par application des formules (14) et (18).

Elle est représentée en bas à droite du sous-tableau des répétabilités.

$$\widehat{r} = 1,060$$

7.5.2 Reproductibilité multidimensionnelle

La variance de reproductibilité est obtenue par application de la formule (20).

$$\sigma_R^2 = \sigma_L^2 + \sigma_r^2$$

où σ_L^2 : composante de la variance de l’effet interlaboratoire

et σ_r^2 : composante de l’effet intralaboratoire.

Comme dans le cas de la répétabilité, l’estimation de σ_R^2 nécessite l’estimation de σ_L^2 qui peut être présentée dans deux situations :

– La reproductibilité monodimensionnelle (j fixé) selon la formule (19).

Les résultats sont rapportés sur le tableau 3 à la ligne V_L^2 , où $V_L^2(j) = M_B^2(j)/I$.

Sur la ligne V_r^2 est rapportée la composante de variance intralaboratoire correspondante, soit : $V^{2r}(j) = M_W^2(j)/I$, où $M_W^2(j)$ est l’inertie intralaboratoire de la variable j .

Sur la ligne V_R^2 est le résultat de l’application de la formule (20) pour chaque niveau j , soit $\widehat{\sigma}_R^2(j)$.

La ligne \widehat{R}_j représente la reproductibilité monodimensionnelle pour chaque niveau J .

– La reproductibilité multidimensionnelle est obtenue en appliquant les formules (21) et (24) :

$$\widehat{R} = \sqrt{(11,07 * 2 * 103,57)/(18 * 5)} = 5,048$$

11,07 étant le quantile d'ordre 0,95 d'une loi du chi-deux à 5 degrés de liberté.

8. Conclusion

Cet article propose, d'une part des méthodes simples de détection de valeurs aberrantes dans des situations multidimensionnelles, d'autre part une généralisation multidimensionnelle du concept normalisé de fidélité.

Ces outils, qui donnent un éclairage multidimensionnel au tableau de données, viennent en complément des méthodes normalisées qui s'appliquent niveau par niveau.

Ces méthodes viennent donc utilement appuyer l'analyste, en particulier lorsqu'il est engagé dans une démarche d'assurance qualité, d'accréditation ou de certification.

A l'heure actuelle, ces outils sont mis en oeuvre à l'UTAC pour l'analyse d'essais interlaboratoires et permettent de synthétiser une information complexe.

Toutefois, des travaux restent à mener sur la signification concrète de la fidélité multidimensionnelle quand les descripteurs statistiques sont de nature ou de niveau différents.

En effet, le recodage abrupt ou l'opération classique de réduction peuvent conduire à éliminer des données les phénomènes d'intérêt (*eg.* rendre le nuage sphérique quand on s'intéresse aux dispersions).

En outre, les hypothèses usuelles de normalité sont ici complétées pour la facilité des calculs d'une hypothèse d'homoscédasticité dont la pertinence peut être contestée.

Nous souhaitons donc que les éléments posés dans cet article soient validés dans différents domaines d'application et complétés de développements confortant cette approche.

Remerciements

Les auteurs remercient la Rédaction de la Revue pour ses commentaires pertinents ainsi que Philippe Poirier (UTAC) pour la programmation en SAS des outils présentés dans cet article.

Références

- [1] BARNETT V., LEWIS T. Outliers in statistical data. - Wiley, 1984
- [2] BECKMAN R.J., COOK R.D. Outlier...s. - Technometrics, 25,2, 1983
- [3] CLEROUX R., HELBLING J.M., RANGER N. Détection d'ensembles de données aberrantes en analyse des données multivariées, Revue de Statistique Appliquée, 38,1, 1990

- [4] GNANADESIKAN R. *Methods for statistical data analysis of multivariate observations.* - Wiley, 1977
- [5] NF-ISO Fidélité des méthodes d'essai. Détermination de la répétabilité et de la reproductibilité d'une méthode d'essai normalisée par essais interlaboratoires. Norme ISO 5725 - 1986, AFNOR NF X 06-041 (1987).