DÉTERMINATION DE LA RÉPÉTABILITÉ ET DE LA REPRODUCTIBILITÉ D'UNE MÉTHODE DE MESURE NORMALISÉE SELON LA NORME ISO 5725-2

R. Palm*

RÉSUMÉ

Cette note présente les aspects statistiques liés à la mise en œuvre de la norme ISO 5725-2 lors de la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée. La démarche est ensuite appliquée à un exemple concret relatif à une méthode de détermination de la superficie de parcelles agricoles à partir de photos aériennes.

SUMMARY

This note describes the statistical aspects related to the norm ISO 5725-2 for the estimation of repeatability and reproducibility of a standard measurement method. The procedure is then applied to a real world example related to the measurement of land parcel areas by remote sensing.

1. INTRODUCTION

La norme ISO 5725-2, intitulée exactitude (justesse et fidélité) des résultats et méthodes de mesure, comprend une partie 2 qui décrit la "méthode de base pour la détermination de la répétabilité et la reproductibilité d'une méthode de mesure normalisée [X, 2000].

L'objectif de cette partie de la norme est de définir la démarche à suivre, en pratique, pour la conduite d'une expérience interlaboratoire, destinée à produire des valeurs numériques de la précision de méthodes de mesures. L'approche est

*Professeur à la Faculté universitaire des Sciences agronomiques de Gembloux (Unité de Statistique, Informatique et Mathématique appliquées).

typiquement définie dans le contexte de mesures chimiques ou physiques mais peut être étendue à d'autres situations.

Bien que la norme envisage de nombreux aspects non directement en relation avec le traitement proprement dit des données, nous nous limitons dans cette note uniquement aux aspects plus statistiques, en insistant particulièrement sur l'examen préliminaire des données en vue de mettre en évidence d'éventuelles inconsistances ou données aberrantes.

Nous présentons d'abord le schéma type d'une expérience interlaboratoire destinée à quantifier la répétabilité et la reproductibilité (paragaphe 2). Ensuite, nous passons en revue différents tests statistiques et d'autres outils préconisés par la norme pour la recherche de données anormales ou inconsistantes (paragraphe 3). Nous précisons alors comment la norme utilise ces outils dans une procédure pas à pas (paragraphe 4). Enfin, nous présentons un exemple numérique concret (paragraphe 5), et nous donnons quelques informations complémentaires (paragraphe 6), avant de conclure (paragraphe 7).

2. SCHÉMA GÉNÉRAL DE L'ANALYSE D'UNE EXPÉRIENCE INTERLABORATOIRE

2.1. Dispositif expérimental et tableaux des données

La procédure de base pour déterminer la répétabilité et la reproductibilité d'une méthode d'analyse donnée consiste à envoyer à p laboratoires q échantillons de matière représentant q niveaux de teneurs différentes du composant à analyser. Chaque laboratoire réalise n répétitions de l'analyse, sous les conditions de répétabilité.

Ces conditions de répétabilité sont des conditions où les résultats de mesures indépendantes sont obtenues par une même méthode sur des matériaux identiques, dans le même laboratoire, par le même opérateur, utilisant le même équipement et pendant un cours intervalle de temps [X, 2000].

Les pqn résultats peuvent alors être rassemblés dans un tableau comportant q colonnes, à raison d'une colonne par niveau de teneur. Chaque colonne comporte p groupes de n valeurs, correspondant aux n répétitions de chacun des laboratoires.

2.2. Examen critique des résultats

L'examen critique des données repose sur le regroupement de celles-ci par laboratoire. Les moyennes et les écarts-types des répétitions sont calculés par laboratoire pour chacun des niveaux de l'essai et sont ensuite comparés à l'aide d'outils statistiques décrits ci-dessous (paragraphe 3). Ces outils permettent de mettre en évidence un éventuel comportement particulier de l'un ou l'autre laboratoire pour un niveau donné de l'essai ou, au contraire, pour l'ensemble des

essais. Ils permettent également de détecter d'éventuelles données aberrantes. La mise en œuvre de ces outils est détaillée au paragraphe 4.

L'examen critique des données peut conduire à la correction de données erronées ou à l'élimination de certaines observations au sein d'un laboratoire ou encore à l'élimination de l'ensemble des observations faites par un laboratoire, pour un ou pour plusieurs niveaux de l'essai.

2.3. Composantes de la variance, répétabilité et reproductibilité

Pour un niveau de l'essai, une analyse de la variance à un critère est réalisée. Cette analyse donne le carré moyen entre laboratoires, CM_{labo}, et le carré moyen résiduel, CM_r. Ce dernier carré moyen est une estimation de la variance entre répétitions au sein des laboraboires :

$$\widehat{\sigma}_r^2 = CM_r \,.$$

Quant à la variance entre laboratoires, elle est estimée par l'équation suivante :

$$\widehat{\sigma}_{\text{labo}}^2 = \frac{CM_{\text{labo}} - CM_r}{n} \,,$$

n étant le nombre de répétitions dans les laboratoires.

Il peut arriver que le nombre de répétitions varie d'un laboratoire à l'autre, du fait de problèmes techniques ou encore du fait de l'élimination d'observations durant la phase de l'examen critique des résultats (paragraphe 3). Dans ce cas, le dispositif expérimental n'est plus équilibré et la valeur de n qui apparaît dans la formule donnant la variance entre laboratoires est remplacée par :

$$n' = \left(n_{\cdot}^2 - \sum_{i=1}^{p} n_i^2 \right) \Big/ \left[n_{\cdot} (p-1) \right] .$$

Dans cette formule, n_i est le nombre de répétitions pour le laboratoire i ($i = 1, \ldots, p$) et $n_{\cdot}$ est le nombre total d'observations :

$$n_{\cdot} = \sum_{i=1}^{p} n_i \,.$$

Lorsque le carré moyen entre laboratoires est plus faible que le carré moyen résiduel, l'application de la relation ci-dessus conduirait à une composante de la variance entre laboratoires qui serait négative. Dans ce cas, on considère que la variance entre laboratoires est nulle.

Les composantes de la variance obtenues ci-dessus permettent de déterminer la variance de répétabilité et la variance de reproductibilité : la variance de répétabilité est la variance résiduelle $\widehat{\sigma}_r^2$, comme signalé ci-dessus, et la variance de reproductibilité $\widehat{\sigma}_R^2$ est la somme de la variance de répétabilité et de la variance entre laboratoires :

$$\widehat{\sigma}_R^2 = \widehat{\sigma}_{\text{labo}}^2 + \widehat{\sigma}_r^2 \,.$$

2.4. Relation entre les écarts-types et les niveaux moyens

Lorsque l'essai porte sur plusieurs niveaux d'une caractéristique donnée, telle que la teneur en un constituant particulier, les valeurs des écarts-types de répétabilité et de reproductibilité sont d'abord déterminées indépendamment pour chacun des niveaux. Dans un second temps, des estimations globales peuvent être recherchées, soit par le calcul d'une relation fonctionnelle entre les écarts-types et les niveaux moyens, soit par le calcul de moyennes pour l'ensemble ou pour un sous-ensemble de niveaux.

Des graphiques reprenant les q écarts-types de répétabilité ou de reproductibilité en fonction des q moyennes générales $\widehat{m}_j$ $(j = 1, \ldots, q)$, calculées pour chaque niveau, sont établis. Ensuite, en fonction de l'allure générale de ces graphiques, on recherche une relation susceptible de modéliser l'évolution des écarts-types en fonction des moyennes. Parmi les modèles les plus simples on a :

$$\widehat{\sigma} = a + b\,\widehat{m}$$

et

$$\log \widehat{\sigma} = c + d \log \widehat{m}\,.$$

Dans ces relations, $\widehat{\sigma}$ désigne soit l'écart-type de répétabilité, soit l'écart-type de reproductibilité, $\widehat{m}$ est la valeur moyenne de la caractéristique pour un niveau donné, a, b, c et d sont les constantes à déterminer.

Le premier modèle, qui est parfois simplifié par la suppression de l'ordonnée à l'origine, est ajusté par la méthode des moindres carrés pondérés car l'erreur standard de l'écart-type est proportionnelle à l'écart-type [Dagnelie, 1998]. La procédure est itérative : à la première itération les écarts-types utilisés pour déterminer les pondérations sont les $\widehat{\sigma}_j$ obtenus lors de l'analyse séparée des données par niveau (paragraphe 2.3). Dans les itérations successives, les pondérations sont déterminées à partir des écarts-types estimés par la relation.

Des informations relatives au calcul de cette relation sont données dans la norme [X, 2000]. D'une manière plus générale une description de la régression pondérée est donnée, notamment, dans Palm [1994].

L'ajustement de la seconde relation est plus simple. En effet, l'écart-type de $\log \widehat{\sigma}$ étant indépendant de σ, une régression ordinaire peut être utilisée.

Lorsqu'un modèle de régression a été ajusté, les écarts-types de répétabilité ou de reproductibilité sont déduits de cette relation. Par contre, si aucune relation ne peut être établie, les écarts-types estimés par niveau sont utilisés tels qu'ils sont obtenus par la procédure décrite au paraphe 2.3 ou sont regroupés sous la forme d'une moyenne pour l'ensemble ou un sous-ensemble de niveaux.

3. OUTILS STATISTIQUES POUR L'EXAMEN CRITIQUE DES DONNÉES

3.1. Considérations préliminaires

La première étape de l'analyse est l'examen critique des données afin d'identifier et de traiter les valeurs aberrantes et autres irrégularités et de tester l'adéquation du modèle statistique utilisé.

Les valeurs aberrantes sont des observations qui s'écartent des données comparables d'une manière telle qu'elles doivent être considérées comme étant incohérentes avec les autres données.

Dans le cadre de la norme ISO 5725-2, la détection des valeurs aberrantes repose sur le regroupement des observations en fonction d'un facteur. Pour une expérience interlaboratoire, ce facteur est typiquement le facteur laboratoire. En relation avec ce facteur de regroupement, le caractère aberrant d'une observation ou d'un ensemble d'observations peut se traduire de plusieurs manières et l'utilisation conjointe de plusieurs outils statistiques vise à mettre en évidence les diverses situations. Ces outils reposent soit sur une approche graphique (paragraphe 3.2), soit sur des tests statistiques (paragraphes 3.3 à 3.5).

3.2. Technique graphique de cohérence

Pour chaque laboratoire et pour chaque niveau de l'essai, la statistique suivante de cohérence entre laboratoires est déterminée :

$$h_i = (\bar{y}_i - \bar{\bar{y}}) \Big/ \sqrt{\sum_{i=1}^{p} (\bar{y}_i - \bar{\bar{y}})^2 / (p-1)}\,,$$

p étant le nombre de laboratoires, $\bar{y}_i$ la moyenne pour le laboratoire i et $\bar{\bar{y}}$ la moyenne générale pour l'ensemble des laboratoires. Il s'agit donc d'une mesure standardisée de l'écart entre la moyenne observée pour le laboratoire i et la moyenne générale.

Une deuxième statistique compare la variabilité au sein du laboratoire i et la variabilité moyenne dans les laboratoires :

$$k_i = \frac{\widehat{\sigma}_i}{\tilde{\sigma}}\,,$$

$\widehat{\sigma}_i^2$ étant la variance entre répétitions dans le laboratoire i et $\tilde{\sigma}^2$ la moyenne arithmétique de toutes les variances dans les p laboratoires :

$$\widehat{\sigma}_i^2 = \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2 / (n_i - 1)$$

et

$$\tilde{\sigma}^2 = \frac{1}{p} \sum_{i=1}^{p} \widehat{\sigma}_i^2\,.$$

Si les nombres de répétitions par laboratoire sont constants ($n_i = n$ pour $i = 1, \ldots, p$), alors $\tilde{\sigma}^2$ est la variance de la répétabilité.

Les statistiques h_i et k_i sont calculées pour chaque niveau j ($j = 1, \ldots, q$). Pour cette raison, elles sont notées de manière plus rigoureuse h_{ij} et k_{ij}.

Les statistiques h_{ij} sont alors portées sur un graphique, sous la forme de bâtons dont la hauteur est proportionnelle à h_{ij}, les valeurs étant groupées par laboratoire, et, au sein d'un laboratoire, ordonnées selon l'indice j. De plus, des droites horizontales situées en $h_{0,05}$, $h_{0,01}$, $-h_{0,05}$ et $-h_{0,01}$ sont portées sur ce graphique. Les valeurs $h_{0,05}$ et $h_{0,01}$ sont les valeurs critiques au seuil 0,05 et 0,01, données dans des tables [X, 2000].

De même, les statistiques k_{ij} sont portées sur un graphique, dans le même ordre que les h_{ij} et deux droites horizontales, situées en $k_{0,05}$ et en $k_{0,01}$ sont ajoutées, les valeurs critiques $k_{0,05}$ et $k_{0,01}$ étant données dans des tables [X, 2000].

L'examen de ces graphiques peut révéler par exemple l'existence de laboratoires présentant une moyenne ou une variabilité systématiquement plus faible ou, au contraire, systématiquement plus grande que les autres laboratoires. Notons que la norme ne donne pas de test statistique formel permettant de juger un laboratoire, la décision finale de conserver ou d'éliminer un laboratoire étant de la responsabilité du statisticien.

3.3. Test de COCHRAN

Le test de Cochran [1941] est un test d'égalité d'une série de variances, l'hypothèse alternative étant qu'au moins une variance est plus grande. Il s'agit donc d'un test unilatéral.

Soit $\widehat{\sigma}_1^2, \ldots \widehat{\sigma}_p^2$, les variances estimées dans les laboratoires et soit $\hat{\sigma}_{\max}^2$ la plus grande de ces variances. On calcule la statistique C de Cochran :

$$C = \widehat{\sigma}_{\max}^2 \Big/ \sum_{i=1}^{p} \widehat{\sigma}_i^2$$

et on rejette l'hypothèse nulle si celle-ci est supérieure à la valeur critique, qui est fonction du nombre n de répétitions par laboratoire, du nombre p de laboratoires ainsi que du seuil de signification. Des tables de ces valeurs critiques sont données dans Pearson et Hartley [1966] et X [2000].

De façon stricte, ce test implique que le nombre de répétitions par laboratoire soit constant. Du fait de l'existence de données manquantes, ce nombre peut varier. Toutefois si l'expérimentation a été bien organisée, on peut s'attendre à ce que la fluctuation dans le nombre de répétitions soit suffisamment faible de manière à pouvoir être ignorée, la valeur n retenue pour le test étant alors la valeur la plus souvent observée dans les laboratoires.

3.4. Test de GRUBBS pour la détection d'une donnée aberrante

Soit une série statistique de n données, ordonnées par ordre croissant :

$$x_{[1]}, x_{[2]}, \ldots, x_{[n]} ,$$

de moyenne $\bar{x}$ et d'écart-type $\widehat{\sigma}$.

Soit :

$$G1_{\min} = \frac{\bar{x} - x_{[1]}}{\widehat{\sigma}} \quad \text{et} \quad G1_{\max} = \frac{x_{[n]} - \bar{x}}{\widehat{\sigma}} ,$$

les écarts standardisés par rapport à la moyenne pour la plus petite et pour la plus grande observation.

Soit $G1$ la plus grande de ces deux valeurs :

$$G1 = \max \left[G1_{\min}, G1_{\max} \right] .$$

La valeurs extrême ($x_{[1]}$ si $G1 = G1_{\min}$, $x_{[n]}$ si $G1 = G1_{\max}$), est appelée valeur isolée[1] et est signalée par un astérisque si la statistique $G1$ est supérieure à la valeur critique pour le niveau de signification de 5 %, mais inférieure à la valeur critique pour le niveau de signification de 1 %. La valeur extrême est appelée valeur aberrante[2], et est signalée par un double astérisque si la statistique $G1$ est supérieure à la valeur critique pour le niveau de signification de 1 %. Des tables de ces valeurs critiques sont données par GRUBBS [1969], GRUBBS et BECK [1971] et sont également reprises dans X [2000].

3.5. Test de GRUBBS pour détection de couples de valeurs aberrantes

Le but est de vérifier si les deux valeurs les plus faibles, $x_{[1]}$ et $x_{[2]}$ ou bien les deux valeurs les plus grandes, $x_{[n-1]}$ et $x_{[n]}$, d'une série d'observations ordonnées, doivent être considérées comme des valeurs aberrantes.

Soit SCE la somme des carrés des écarts pour les n observations :

$$SCE = \sum_{i=1}^{n} (x_{[i]} - \bar{x})^2 \quad \text{avec} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{[i]} .$$

Soient SCE' et SCE'' les sommes des carrés des écarts obtenues après suppression des deux observations les plus petites :

$$SCE' = \sum_{i=3}^{n} (x_{[i]} - \bar{x}')^2 \quad \text{avec} \quad \bar{x}' = \frac{1}{n-2} \sum_{i=3}^{n} x_{[i]}$$

et des deux observations les plus grandes :

$$SCE'' = \sum_{i=1}^{n-2} (x_{[i]} - \bar{x}'')^2 \quad \text{avec} \quad \bar{x}'' = \frac{1}{n-2} \sum_{i=1}^{n-2} x_{[i]} .$$

1. En anglais : *straggler*.
2. En anglais : *outlier*.

Soit :

$$G2_{\min} = SCE'/SCE \quad \text{et} \quad G2_{\max} = SCE''/SCE\,,$$

$G2$ étant la plus petite de ces deux valeurs :

$$G2 = \min[G2_{\min}, G2_{\max}]\,.$$

Les deux observations ayant conduit à $G2$ ($x_{[1]}$ et $x_{[2]}$ si $G2 = G2_{\min}$, $x_{[n-1]}$ et $x_{[n]}$ si $G2 = G2_{\max}$) sont appelées valeurs isolées et signalées par un astérisque si la statistique $G2$ est inférieure à la valeur critique pour le niveau de signification de 5 %, mais supérieure à la valeur critique pour le niveau de signification de 1 %. Elles sont considérées comme valeurs aberrantes et signalées par un double astérisque si la statistique $G2$ est inférieure à la valeur critique pour le niveau de signification de 1 %. Des tables reprenant ces valeurs critiques sont données dans GRUBBS [1969], GRUBBS et BECK [1971] et dans X [2000].

4. PROCÉDURE D'ANALYSE PAS À PAS

4.1. Les grandes étapes

La norme ISO décrit la succession des étapes à suivre pour l'analyse statistique des données. Nous reprenons ci-dessous les étapes principales qui conduisent à la détermination des écarts-types de répétabilité et de reproductibilité pour chacun des q niveaux. Par contre, nous ne détaillons pas les étapes, plus classiques, de la modélisation de ces écarts-types en fonction du niveau moyen.

L'analyse des données commence par un examen visuel destiné à repérer les anomalies évidentes dans les données, telles que, par exemple, des résultats techniquement impossibles, la présence d'un trop grand nombre de données manquantes pour un laboratoire donné, etc. Ces anomalies peuvent éventuellement être corrigées après contact avec les laboratoires concernés ou sont éliminées avant toute analyse ultérieure.

Ensuite, les graphiques des statistiques h_{ij} et k_{ij} de MANDEL décrits au paragraphe 3.2 sont établis et examinés afin de repérer d'éventuelles inconsistances entre laboratoires qui peuvent conduire à identifier l'un ou l'autre laboratoire caractérisé par des résultats globalement en désaccord avec les autres laboratoires, par exemple, du fait de la présence d'un biais (résultats ayant globalement une tendance à être supérieurs ou, au contraire, inférieurs à ceux des autres laboratoires) ou d'une variabilité des résultats systématiquement plus grande ou plus faible que celle des autres laboratoires. Cependant une décision définitive n'est pas prise à cette étape mais est retardée jusqu'après la réalisation des tests statistiques de détection de données isolées ou de données aberrantes.

L'étape suivante consiste à identifier les données aberrantes à l'aide des tests statistiques décrits aux paragraphes 3.3 à 3.5. La mise en œuvre de ces tests est décrite aux paragraphes 4.2 et 4.3.

A l'issue de cette étape, toutes les observations identifiées comme isolées ou aberrantes sont consignées dans le rapport, en même temps que la raison pour

laquelle elles ont été identifiées. Les données aberrantes qui ne peuvent être corrigées sont alors éliminées, mais les valeurs isolées sont, par contre, conservées.

Après cette élimination éventuelle de données aberrantes, les composantes de la variance sont estimées et les écarts-types sont déterminés comme expliqué au paragraphe 2.3.

Enfin, si plusieurs niveaux ont été considérés au cours de l'expérimentation, les écarts-types de répétabilité et de reproductibilité sont mis en relation avec le niveau, pour autant qu'une relation puisse être déterminée. Sinon des valeurs moyennes pour l'ensemble des niveaux ou pour des groupes de niveaux sont déterminées (paragraphe 2.4).

4.2. Identification des données isolées et des données aberrantes

Cette identification se fait à travers les étapes décrites ci-dessous et est réalisée indépendamment pour chaque niveau.

1. La statistique C de COCHRAN est calculée pour le laboratoire présentant la plus grande variance entre répétitions. Si cette statistique est inférieure à la valeur critique au niveau de signification de 5 %, on passe au point 5.

2. Si la statistique C est plus grande que la valeur critique à 5 %, les observations provenant de ce laboratoire sont soigneusement examinées afin d'identifier d'éventuelles observations aberrantes à l'origine de cette trop grande variance. Cette identification est basée sur les tests de GRUBBS, selon la procédure décrite au paragraphe 4.3 ci-dessous. Si des observations sont déclarées aberrantes, elles sont éliminées et le test de COCHRAN est appliqué à nouveau (retour au point 1).

3. Si aucune donnée aberrante n'est détectée par les tests de GRUBBS au point 2 ci-dessus et si la statistique C de COCHRAN déterminée au point 1 est supérieure à la valeur critique à 5 %, mais inférieure à la valeur critique à 1 %, on passe au point 5.

4. Si aucune observation n'est considérée comme aberrante au point 2 et si la statistique C de COCHRAN est supérieure à la valeur critique à 1 %, toutes les observations du laboratoire qui, pour le niveau considéré, présente la plus grande variance sont éliminées et le test de COCHRAN est appliqué à nouveau (retour au point 1).

5. Si, après élimination éventuelle de données aberrantes, la statistique de COCHRAN est inférieure ou égale à la valeur critique à 1 %, les moyennes des laboratoires sont examinées afin d'identifier d'éventuelles moyennes aberrantes. Cette identification repose sur les tests de GRUBBS selon la procédure décrite au paragraphe 4.3. Si des moyennes sont identifiées comme aberrantes, toutes les observations des laboratoires présentant des moyennes aberrantes, pour un niveau donné, sont éliminées.

4.3. Réalisation des tests de GRUBBS

Les deux tests de GRUBBS ont été présentés aux paragraphes 3.4 et 3.5. Le premier test, désigné ci-dessous par le symbole GRUBBS/1 vérifie si le minimum ou le maximum d'une série d'observations doit être considéré ou non comme une valeur aberrante. Ce test est basé sur la statistique $G1$ définie précédemment. Le second test, désigné ci-après par le symbole GRUUBS/2 vérifie si les deux plus petites ou les deux plus grandes observations doivent être considérées comme aberrantes. La statistique associée à ce test est $G2$.

La recherche de données aberrantes est réalisée sur les observations provenant du laboratoire qui présente une variabilité considérée comme exagérément grande (point 2, paragraphe 4.2). Si une ou deux observations sont déclarées aberrantes, au niveau de signification de 1 %, elles sont éliminées.

L'identification des données aberrantes est également réalisée sur la série des valeurs moyennes par laboratoire (point 5, paragraphe 4.2). Si une ou deux moyennes sont déclarées aberrantes, au niveau de signification de 1 %, toutes les observations des laboratoires concernés sont éliminées pour le niveau considéré.

Qu'il s'agisse de la recherche d'observations aberrantes au sein d'un laboratoire ou de moyennes par laboratoire aberrantes, la procédure suivie est la suivante.

1. La statistique $G1$ du test GRUBBS/1 est déterminée.
2. Si la statistique $G1$ est inférieure ou égale à la valeur critique au niveau 1 %, aller au point 6.
3. Si la statistique $G1$ est supérieure à la valeur critique au niveau 1 %, la valeur extrême est éliminée et le test GRUBBS/1 est appliqué à nouveau à la valeur extrême située à l'autre extrémité (valeur maximum si l'extrême éliminé est le minimum, valeur maximum si l'extrême éliminé est le maximum).
4. Si la statistique $G1$ pour ce second test de GRUBBS/1 est inférieure ou égale à la valeur critique au niveau 1 %, la détection de données aberrantes est arrêtée.
5. Si, par contre, la statistique $G1$ pour ce second test de GRUBBS/1 est supérieure à la valeur critique au niveau 1 %, la valeur concernée est éliminée et la détection des données aberrantes est arrêtée.
6. Si aucune valeur aberrante n'est identifiée lors de la première exécution du test GRUBBS/1, à l'étape 1, le test de GRUBBS/2 est appliqué.
7. Si la statistique $G2$ associée à ce test est supérieure ou égale à la valeur critique au niveau de 1 %, la détection des données aberrantes est arrêtée.
8. Si la statistique $G2$ est inférieure à la valeur critique au seuil de 1 %, les deux valeurs extrêmes supérieures (si $G2 = G2_{\max}$) ou les deux valeurs extrêmes inférieures (si $G2 = G2_{\min}$) sont éliminées. Le test GRUBBS/2

est appliqué à nouveau à l'autre extrémité de la distribution (aux deux valeurs extrêmes inférieures, si les valeurs qui viennent d'être supprimées étaient les extrêmes supérieurs; aux deux valeurs extrêmes supérieures si les valeurs qui viennent d'être supprimées étaient les extrêmes inférieurs).

9. Si la statistique $G2$ pour ce second test GRUBBS/2 est supérieure ou égale à la valeur critique au niveau de signification de 1 %, la détection des données aberrantes est arrêtée.

10. Si par contre cette statistique est inférieure à la valeur critique au niveau de signification de 1 %, les deux observations sont éliminées et la détection des valeurs aberrantes est arrêtée.

5. EXEMPLE : MESURES DE SUPERFICIES DE PARCELLES AGRICOLES

5.1. Présentation de l'étude

Nous avons signalé dans l'introduction que la méthode décrite dans la norme ISO 5725 est typiquement définie dans le contexte de mesures chimiques ou physiques. Des exemples numériques relevant de ce domaine peuvent d'ailleurs être trouvés dans cette norme [X, 2000].

Le champ d'application peut cependant être étendu à d'autres situations, comme l'illustre l'exemple ci-dessous, où il s'agit de quantifier l'écart-type de répétabilité et de reproductibilité de diverses méthodes de mesure de surfaces de parcelles agricoles. Les données utilisées proviennent d'une étude réalisée à la demande de l'Union Européenne [HEJMANOWSKA *et al.*, 2005]. Cette étude visait à définir une méthode de validation de techniques de mesure de superficies de parcelles agricoles pour leur utilisation lors de contrôle des déclarations d'agriculteurs en vue des paiements de subsides.

Les méthodes de mesure consistent, d'une part, en l'utilisation sur le terrain de différents équipements de mesure par GPS et, d'autre part, des mesure de surfaces sur différents types de photos aériennes. Nous nous limiterons toutefois à l'étude d'une seule méthode, basée sur l'utilisation d'un type de photos aériennes. Trente-six parcelles de surface variant de 0,3 à 4 ha environ ont été sélectionnées. Elles ont ensuite été mesurées par douze opérateurs à trois occasions différentes. On dispose ainsi de 1.296 observations.

L'objectif est d'abord de déterminer la répétabilité et la reproductibilité pour chacune des parcelles. Ensuite, on examinera la possibilité de synthétiser les résultats. Par rapport à une expérience relative à une méthode d'analyse chimique, les parcelles jouent, dans cette application, le rôle des niveaux de teneur et les opérateurs, le rôle des laboratoires.

5.2. Résultats pour une parcelle donnée

Afin d'illustrer les détails de l'analyse, nous avons retenu une parcelle particulière, la parcelle 5, choisie en raison de la présence de données aberrantes dont la mise en évidence nécessite l'utilisation de différents tests.

Le tableau 1 donne, pour les différents opérateurs, les moyennes, les écarts-types, en m^2, et les statistiques h et k de MANDEL. Nous allons tout d'abord illustrer le calcul des valeurs h et k.

Tableau 1. Moyennes, écarts-types et valeurs h et k par opérateur (parcelle 5).

Opérateurs	Moyennes	h	Ecarts-types	k
1	12.412,4	0,96	138,2	1,28
2	12.026,2	–2,30	203,2**	1,88*
3	12.310,6	0,10	167,0	1,55
4	12.365,3	0,56	46,2	0,43
5	12.401,9	0,87	69,5	0,64
6	12.257,0	–0,35	89,7	0,83
7	12.320,2	0,18	12,6	0,12
8	12.390,9	0,78	34,3	0,32
9	12.343,2	0,38	61,9	0,57
10	12.266,5	–0,27	76,7	0,71
11	12.370,7	0,61	39,0	0,36
12	12.117,8	–1,53	153,9	1,43

La moyenne des douze moyennes est égale à 12.298,5 et l'écart-type des moyennes vaut 118,3. Ces valeurs permettent le calcul des statistiques h. Ainsi, pour l'opérateur 1, pris à titre d'exemple, on a :

$$h = (12.412,4 - 12.298,5)/118,3 = 0,96\,.$$

La moyenne quadratique des douze écarts-types est égale à 107,8 et les valeurs de k sont obtenues en divisant les écarts-types par cette moyenne. Pour l'opérateur 1, on a donc :

$$h = 138,2/107,8 = 1,28\,.$$

Pour un nombre de répétitions égal à 3 et un nombre d'opérateurs égal à 12, les valeurs seuils données dans X [2000] sont :

$$h_{0,05} = 1,83 \quad \text{et} \quad h_{0,01} = 2,25\,,$$

$$k_{0,05} = 1,69 \quad \text{et} \quad k_{0,01} = 2,02\,.$$

L'opérateur 1 ne présente donc pas, pour cette parcelle, une moyenne et un écart-type des observations qui doivent être considérés comme extrêmes.

Par contre, pour l'opérateur 2, la moyenne est relativement faible par rapport à celles des autres opérateurs et la variabilité est plus importante. Les valeurs de h et k dépassent les valeurs critiques au niveau de probabilité de 1 % ou de 5 % et sont donc identifiées par deux astérisques ou un astérisque.

Les calculs de h et de k sont répétés pour chacune des 36 parcelles et toutes les valeurs devraient, en principe, être portées sur un graphique (paragraphe 3.2). Compte tenu du nombre élevé de valeurs h_{ij} et k_{ij} calculées dans le cas de l'expérience (12 opérateurs × 36 parcelles), il n'est pratiquement pas possible de représenter toutes les valeurs dans un même graphique. Une solution alternative consiste à construire un graphique par opérateur. A titre d'illustration, nous donnons, dans les figures 1 et 2, les graphiques pour l'opérateur 2, identifié ci-dessus comme présentant des résultats relativement extrêmes pour la parcelle 5.

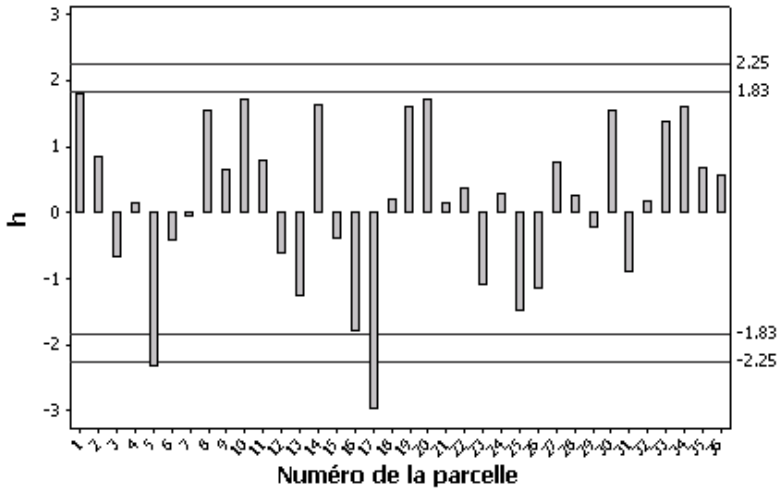

Figure 1. Valeurs de h en fonction du numéro de la parcelle (opérateur 2).

L'examen de ces graphiques ne semble pas indiquer, pour l'ensemble des parcelles, un comportement particulier de l'opérateur 2 par rapport aux autres opérateurs, qui mériterait *a priori* son exclusion de l'étude.

La variabilité plus importante observée pour l'opérateur 2 et la parcelle 5 conduit à la statistique C de COCHRAN suivante :

$$C = (203,2^2/(12)(107,8^2) = 0,296\,.$$

Cette valeur est inférieure à la valeur critique pour le niveau de signification de 5 %, qui est égale à 0,392 [X, 2000]. Aucune recherche de données aberrantes au sein de l'échantillon relatif à l'opérateur 2 n'est donc réalisée.

La série des 12 moyennes fait ensuite l'objet de la recherche de données exceptionnelles par les tests de GRUBBS.

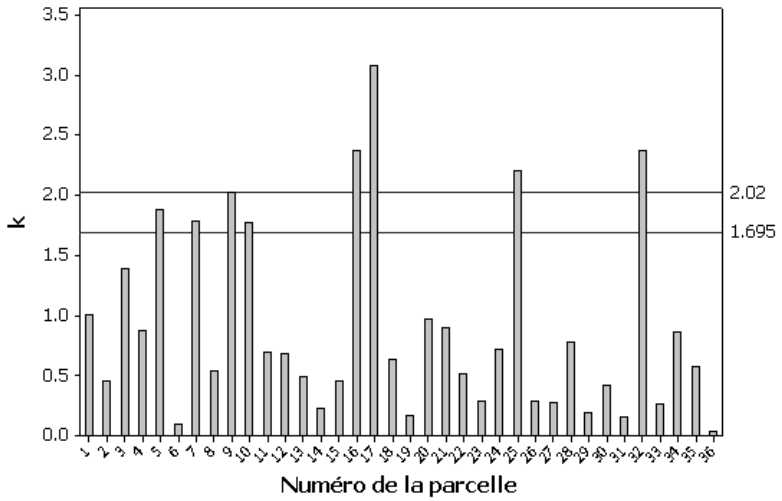

Figure 2. Valeurs de k en fonction du numéro de la parcelle (opérateur 2).

Pour le test GRUBBS/1, on a :

$$G1_{\min} = (12.298,5 - 12.026,2)/118,3 = 2,30\,,$$

$$G1_{\max} = (12.412,4 - 12.298,5)/118,3 = 0,96\,.$$

La valeur de $G1$ est donc égale à 2,30 et doit être comparée à la valeur critique pour le niveau de signification de 5 %, égale à 2,41 [X, 2000]. La valeur $G1$ étant inférieure à 2,41, la moyenne la plus extrême ne doit pas être considérée comme aberrante et le test GRUBBS/2 doit être réalisé. Les deux valeurs les plus petites correspondent aux opérateurs 2 et 12, les deux valeurs les plus grandes correspondent aux opérateurs 1 et 5. Les sommes des carrés des écarts de l'ensemble des observations, SCE, après suppression des deux valeurs les plus faibles, SCE' et après suppression des deux valeurs les plus grandes SCE'', sont respectivement :

$$SCE = 154.022\,, \quad SCE' = 26.657 \quad \text{et} \quad SCE'' = 125.669\,.$$

Il en résulte que :

$$G2_{\min} = 26.658/154.022 = 0,1731 \quad \text{et} \quad G2_{\max} = 125.669/154.022 = 0,8159\,.$$

La valeur de $G2$ est donc égale à 0,1731. Elle est inférieure à la valeur critique au seuil de 1 %, égale à 0,1738 [X, 2000]. Les deux moyennes les plus faibles sont donc considérées comme aberrantes et les observations provenant des opérateurs 2 et 12 sont éliminées.

Après élimination de ces deux opérateurs, le test est recommencé sur les deux valeurs extrêmes supérieures. On a :

$$SCE = 26.657 \quad \text{et} \quad SCE'' = 16.592\,,$$

la première somme des carrés des écarts étant calculée sur 10 données (suppression des opérateurs 2 et 12), la deuxième somme des carrés des écarts étant calculée sur 8 données (suppression des opérateurs 2, 12, 1 et 5). La statistique $G2$ est égale à :

$$G2 = 16.592/26.657 = 0,6224\,.$$

Cette valeur est supérieure à la valeur critique au niveau de signification de 1 %, qui est égale à 0,1150 [X, 2000]. Les deux valeurs extrêmes supérieures ne sont donc pas aberrantes.

En conclusion, la procédure de recherche de données aberrantes pour la parcelle 5 conduit à l'identification de 6 observations réalisées par deux opérateurs (opérateur 2 et opérateur 12). Après vérification que les données ne résultent pas d'une erreur de transcription ou d'encodage, elles sont supprimées.

L'analyse de la variance est ensuite réalisée sur les 30 observations restantes et provenant de 10 opérateurs. Les résultats sont repris dans le tableau 2.

Tableau 2. Tableau d'analyse de la variance (parcelle 5, après élimination de 2 opérateurs).

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Opérateurs	9	79.971	8.886
Répétitions	20	149.177	7.459
Total	29	229.148	

A partir des carrés moyens du tableau 2, on définit les variances entre répétitions (par opérateur) et entre opérateurs :

$$\widehat{\sigma}_r^2 = 7.459 \quad \text{et} \quad \widehat{\sigma}_{\text{oper}}^2 = (8.886 - 7.459)/3 = 476\,.$$

L'écart-type de répétabilité est donc égal à :

$$\widehat{\sigma}_r = \sqrt{7.459} = 86,4$$

et l'écart-type de reproductibilité est égal à :

$$\widehat{\sigma}_R = \sqrt{7.459 + 476} = 89,1\,.$$

5.3. Résultats pour l'ensemble des parcelles

La procédure détaillée au paragraphe précédent a été répétée pour les 36 parcelles. L'identification des observations aberrantes a conduit à éliminer 30 observations sur un total de 1.296, soit 2,3 %. Ces éliminations ont toujours porté sur

des groupes de 3 observations réalisées par un opérateur, essentiellement sur la base du résultat du test de COCHRAN ou alors sur la base des tests de GRUBBS appliqués aux moyennes par opérateur.

Pour les 36 parcelles, on dispose ainsi de l'écart-type de la répétabilité et de l'écart-type de la reproductibilité. Nous ne considérerons par la suite que la reproductibilité, la démarche étant identique pour la répétabilité.

La figure 3 donne les valeurs de l'écart-type de reproductibilité en fonction des surfaces de référence des parcelles, obtenues à partir des relevés cadastraux. L'écart-type augmente avec la surface, mais cette relation semble non linéaire et la variabilité des écarts-types augmente avec la surface.

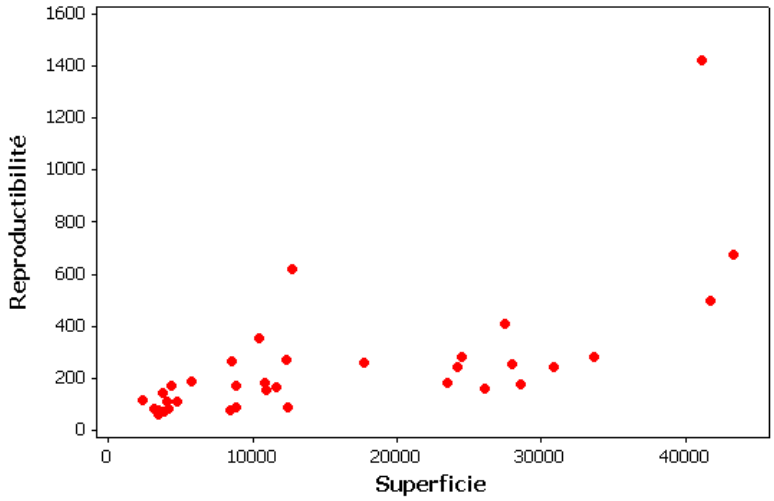

Figure 3. Diagramme de dispersion de l'écart-type de reproductibilité en fonction de la superficie.

Une double transformation logarithmique stabilise la variabilité conditionnelle et linéarise la relation (figure 4).

La droite de régression obtenue s'écrit :

$$\log\left(\widehat{\sigma}_R\right) = -0,3645 + 0,6007 \log\left(\text{surface}\right).$$

Différents tests de validation (test d'adéquation de la relation, test de normalité des résidus, test d'égalité des variances conditionnelles) décrits notamment dans PALM [2002], confirment la pertinence de ce modèle de régression.

En conclusion, au terme de l'expérience, on conclut que l'écart-type de reproductibilité peut s'exprimer en fonction de la surface de la parcelle par le modèle :

$$\widehat{\sigma}_R = 0,6945 \ (\text{surface})^{0,6007}.$$

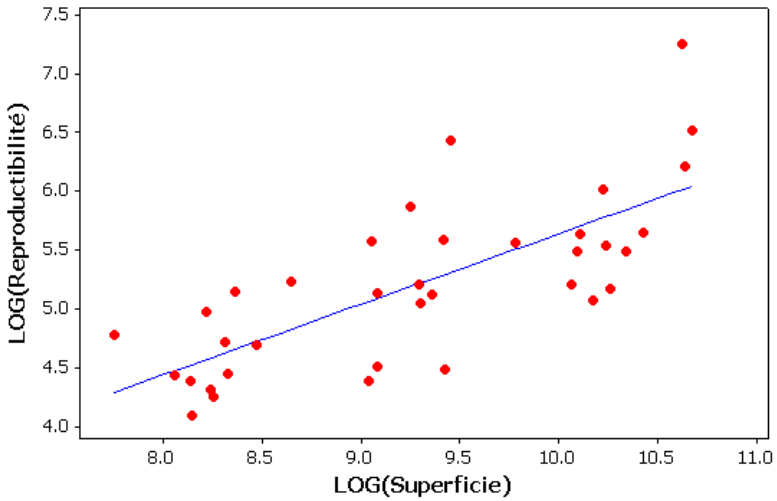

Figure 4. Diagramme de dispersion et droite de régression du logarithme de l'écart-type de la reproductibilité en fonction du logarithme de la superficie.

On notera cependant que la variabilité résiduelle du modèle de régression est importante, le logarithme de la superficie des parcelles ne permettant d'expliquer que 54 % de la variabilité des logarithmes des écarts-types de reproductibilité.

Une autre approche pour synthétiser les résultats obtenus pour les différentes parcelles consiste à rechercher une transformation des écarts-types dont les résultats ne seraient pas liés à la taille de la parcelle. Une telle solution aurait l'avantage de la simplicité. La première idée qui vient à l'esprit est de calculer le coefficient de variation, en divisant l'écart-type par la surface. Dans le cas présent, cette solution n'est cependant pas acceptable, car le coefficient de variation diminue avec la surface, l'exposant dans le modèle défini ci-dessus étant nettement inférieur à l'unité.

Une autre transformation couramment utilisée dans les études relatives à la précision des mesures de surfaces consiste à diviser l'écart-type par le périmètre de la parcelle. Ce rapport définit la largeur d'une zone tampon [3] entourant la parcelle. La figure 5 montre effectivement que la largeur de cette zone tampon ne semble pas liée à la superficie de la parcelle, bien qu'elle soit fort variable d'une parcelle à l'autre. La largeur moyenne est de 0,37 m et l'écart-type vaut 0,26 m.

3. En anglais : *buffer, buffer width.*

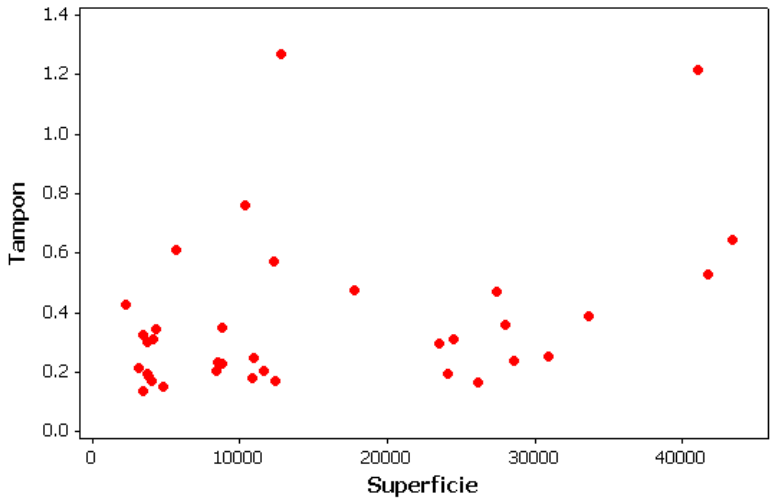

Figure 5. Largeur de la zone tampon en fonction de la superficie.

6. QUELQUES INFORMATIONS COMPLÉMENTAIRES

6.1. Application de la norme hors laboratoire

Nous avons signalé à plusieurs reprises que la norme ISO 5725-2 a été établie dans le contexte d'analyses en laboratoire. Dans ce cas, les notions de répétabilité et reproductibilité sont bien définies. Les conditions de répétabilité sont les conditions où les résultats d'essais indépendants sont obtenus par la même méthode sur des individus d'essai identiques dans le même laboratoire, par le même opérateur, utilisant le même équipement et pendant un cours intervalle de temps et les conditions de reproductibilité sont les conditions où les résultats d'essais sont obtenus par la même méthode sur des individus d'essais identiques dans différents laboratoires, avec des opérateurs et utilisant des équipements différents [X, 2000]. Il en résulte que, dans une expérience classique de validation d'une méthode, le facteur laboratoire est clairement le facteur de regroupement des observations lors de la recherche des données anormales et est aussi le facteur pris en compte lors de l'analyse de la variance.

Les contraintes propres à une étude peuvent justifier qu'on s'écarte du dispositif standard. Ainsi, dans le cas des méthodes de mesure de superficies (paragraphe 5), il est trop complexe pour des raisons pratiques, surtout pour les mesures avec instruments GPS, d'envisager la participation de différentes équipes géographiquement dispersées (équivalentes aux laboratoires d'analyse) et de proposer à un seul opérateur par équipe de réaliser les mesures. Les responsables de l'étude ont donc choisi de confier les mesures à une seule équipe et le dispositif

utilisé correspondait en réalité à un dispositif factoriel, les différents opérateurs réalisant les mesures les mêmes jours.

On se trouve donc en présence d'un modèle d'analyse de la variance à deux critères croisés, que l'on ramène à un seul critère, en considérant le deuxième critère comme un facteur répétition. Dans le cas de la mesure des superficies à partir de photos aériennes on s'attend, et on le vérifie, à ce que la variabilité entre opérateurs soit plus importante que la variabilité entre jours. En conséquence, les données ont été regroupées par opérateur.

Dans le cas de mesures à partir d'appareils GPS, la situation est différente : la variabilité entre jours s'est avérée être supérieure à la variabilité entre opérateurs et le facteur de regroupement a été le facteur jour, les opérateurs jouant le rôle de répétitions. Techniquement, l'effet jour peut se justifier du fait que la qualité des signaux des satellites utilisés par les instruments de mesure varie dans le temps, ce qui n'est évidemment pas le cas pour l'examen des photos aériennes sur ordinateur.

6.2. Les différentes parties de la norme ISO 5725

Nous avons signalé, dans l'introduction, que nous nous intéressions à la deuxième partie d'une norme qui en comporte six. Afin de resituer les aspects examinés dans cette note dans leur contexte général, nous donnons ci-dessous un bref résumé de l'objet des différentes parties de la norme, numérotées ISO 5725-1 à ISO 5725-6.

La première partie concerne les principes généraux et les définitions. Y sont également abordés, les problèmes de détermination du nombre de laboratoires et du nombre de répétitions par laboratoire nécessaires pour l'estimation des variances avec une précision fixée. La norme donne en effet des formules permettant de déterminer la marge d'erreur relative pour un degré de confiance de 95 %, en fonction de n, p et du rapport σ_R/σ_r. D'après cette même norme, il est courant de choisir, en pratique, une valeur de p comprise entre 8 et 15 et cinq niveaux sont généralement suffisants pour couvrir la gamme des valeurs pour lesquelles la méthode de mesure est utilisée.

La seconde partie décrit, comme nous l'avons vu, la méthode de base pour la détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée.

La troisième partie concerne les mesures intermédiaires de la fidélité d'une méthode de mesure normalisée. Les expériences réalisées dans ce contexte ont pour but de quantifier la variabilité des résultats en fonction de diverses sources de variation : facteur temps, étalonnage, opérateur, équipement. Ces facteurs sont fixés dans les conditions de répétabilité (mesures faites à un moment donné, par un même opérateur, avec un même équipement, sans réétalonnage) mais varient dans les conditions de reproductibilité. Les mesures intermédiaires de variabilité quantifient l'aptitude de la méthode de mesure à reproduire des résultats d'essais dans des conditions définies. Les résultats d'expériences conduites dans

le but de quantifier cette variabilité intermédiaire sont soumis à des analyses de la variance à plusieurs critères hiérarchisés.

La quatrième partie traite des méthodes de base pour la détermination de la justesse d'une méthode de mesure normalisée. Elle précise comment estimer le biais d'une méthode de mesure et le biais d'un laboratoire et fait usage des outils statistiques suivants : test de conformité d'une variance, test de conformité d'une moyenne, détermination du nombre de mesures en fonction de la puissance pour le test de conformité d'une moyenne.

La cinquième partie est consacrée aux méthodes alternatives pour la détermination de la fidélité d'une méthode de mesure normalisée. Un premier dispositif est proposé pour éviter une éventuelle influence entre les résultats des répétitions dans un même laboratoire. La solution préconisée est de considérer, pour chaque niveau de teneur, deux matériaux possédant à peu près la même teneur, chacun de ces deux matériaux étant envoyés aux différents laboratoires. L'analyse de l'expérience se fait alors à l'aide d'un modèle croisé à deux critères (critère laboratoire et critère matériau). Un autre dispositif concerne les matériaux très hétérogènes, pour lesquels il n'est pas possible d'envoyer des échantillons parfaitement identiques à chaque laboratoire. Il faut alors prendre en compte un facteur "échantillon dans laboratoire" et réaliser une analyse de la variance à deux critères hiérarchisés (laboratoire, et échantillon dans laboratoire) avec répétitions de mesure sur chaque échantillon. Cette cinquième partie de la norme propose encore l'utilisation de méthodes robustes pour l'estimation des variances de répétabilité et de reproductibilité qui ne nécessitent pas l'élimination de données aberrantes.

Enfin, la sixième partie traite de l'utilisation dans la pratique des valeurs d'exactitude. Y sont présentées :

- des méthodes de contrôle de l'acceptabilité des résultats d'essai;
- des méthodes de contrôle de la stabilité des résultats d'essai dans un laboratoire;
- l'utilisation des écarts-types de répétabilité et de reproductibilité dans l'évaluation des laboratoires;
- la comparaison de méthodes de mesures alternatives.

7. CONCLUSIONS

Dans cette note, nous avons mis l'accent sur les aspects statistiques liés aux expériences de validation de méthodes de mesure. De nombreux autres aspects, plus pratiques, sont encore développés dans la norme. Le lecteur impliqué dans une telle expérience y trouvera notamment de nombreuses informations liées à l'organisation de celle-ci.

De plus, nous avons principalement développé l'examen préliminaire des données et la quantification de la variabilité, directement liée à la notion de fidélité.

Pour la variabilité, nous nous sommes limités à la notion de répétabilité et de reproductibilité, qui sont les deux extrêmes de la fidélité, la première donnant le minimum et la deuxième le maximum de la variabilité des résultats. L'appréciation de la variabilité pour des situations intermédiaires est cependant possible et fait l'objet de la troisième partie de la norme. Quant au problème du biais d'une méthode, qui mesure sa justesse, nous ne l'avons pas examiné. Ce point aussi est développé dans la norme, notamment dans la première et dans la quatrième partie.

La démarche utilisée dans la norme ISO 5725-2 repose sur un certain nombre de conditions d'application. La méthode de mesure doit être normalisée et conduire à des observations se situant sur une échelle continue. Les modèles statistiques supposent l'égalité des variances au sein des laboratoires ainsi que la normalité des erreurs aléatoires. Des écarts modérés par rapport à ces conditions, notamment pour la normalité, n'entraînent pas de conséquences importantes sur les résultats. Par contre, dans le cas de distributions très dissymétriques, des transformations de variables peuvent être envisagées pour se rapprocher davantage des conditions d'application.

Enfin, rappelons que l'objectif de cette note était de présenter les outils statistiques utlisés dans le contexte de la norme ISO 5725-2 et non de donner un aperçu général concernant l'analyse de l'aptitude de méthodes de mesures. Pour ce problème, le lecteur se reportera à la synthèse proposée par BURDICK *et al.* [2003].

BIBLIOGRAPHIE

BURDICK R.K., BORROR C.M., MONTGOMERY D.C. [2003]. A review of methods for measurement systems capability analysis. *J. Qual. Technol.* 35(4), 342-354.

COCHRAN W.G. [1941]. The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugen.* 11, 47-52.

DAGNELIE P. [1998]. *Statistique théorique et appliquée. Tome 1 : Statistique descriptive et bases de l'inférence statistique*. Bruxelles, De Boeck et Larcier, 508 p.

GRUBBS F. [1969]. Procedures for detecting outlying observations in samples. *Technometrics* 11(1), 1-21.

GRUBBS F., BECK G. [1971]. Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* 14(4), 847-854.

HEJMANOWSKA B., PALM R., OSZCZAK S., CIECKO A. [2005]. *Validation of methods for measurements of land parcels areas. Final report.* AGH University of Science and Technology. Cracow, Faculty of Mining Surveying and Environmental Engineering, Department of Photogrammetry and remote Sensing, 166 p.

PALM R. [1994]. La régression linéaire pondérée : principes et application. *Notes Stat. Inform.* (Gembloux) 94/4, 20 p.

PALM R., IEMMA A.F. [2002]. Conditions d'application et transformations de variables en régression linéaire. *Notes Stat. Inform.* (Gembloux) 2002/1, 34 p.

PEARSON E.S., HARTLEY H.O. (ed.) [1966]. *Biometrika tables for statisticians* (Vol. 1). Cambridge, University Press, 264 p.

X [2000]. *Statistical methods for quality control* (Vol. 2) Measurement methods and results, Interpretation of statistical data, Process control. ISO Standards Handbook, Genève. International Organization of Standardization, 747 p.