G. Matheron
1963

COMPARAISON DES ECHANTILLONNAGES MINE ET USINE A MURVILLE

1.- Données

La comparaison porte sur trois listes de résultats, marquées A, B et C. La liste C donne les résultats de l'échantillonnage mine analysé à la mine. La liste A donne le résultat de l'analyse faite au laboratoire de l'usine sur la même prise de 100 g analysée par ailleurs à la mine (liste C) : ainsi la comparaison de A et C ne concerne que les opérations de laboratoire, et non pas l'échantillonnage proprement dit.

La liste B, enfin, donne les résultats des analyses faites à l'usine sur les produits obtenus par l'échantillonnage continu effectué à l'usine. Le contenu de chaqueconvoi de minerai est, en principe, caractérisé par trois chiffres A,B,C. Dans la suite, nous désignerons par A_i, B_i et C_i les résultats relatifs au convoi N°i. Les séries chronologiques A_i, B_i et C_i étaient disponibles pour les mois de Janvier à Septembre 1963, (avec un à trois convois quotidiens). En fait, nous avons dû renoncer à utiliser les données relatives aux trois premiers mois, pour lesquels les séries A_i et C_i présentent des interversions manifestes (correspondant à des confusions dans l'attribution des deux chiffres quotidiens aux deux convois correspondants).

2.- Méthode, et résumé des résultats.

La méthode de travail utilisée consiste dans l'analyse des auto corrélations des séries chronologiques A,B, et C et des corrélations croisées. Les résultats sont les suivants :

a/. Les deux séries A et C (analyses d'une même prise par l'usine et la mine respectivement) sont en très bon accord. Elles ne comportent pas de divergence systématique. La précision du laboratoire de l'usine (± 0,64%) est légèrement supérieure à celle du laboratoire de la mine (± 0,86 %).

... / ...

b/.- Les deux séries A et B présentant une différence systématique. La valeur moyenne de $A_i - B_i$ est de l'ordre de + 0,4, valeur significative statistiquement, correspondant à une divergence systématique.

c/.- L'examen des variances montre une première anomalie. La variance de A(1,26) est notablement plus faible que celle de B(2,03). Si les deux modes de prélèvement étaient corrects, la variance du dispositif continu B devrait être plus faible que celle du dispositif A.

d/.- L'examen des auto-corrélations met en évidence une deuxième anomalie. La série A_i ne présente que des autocorrélations nulles ou négligeables : Entre un convoi i et celui qui le précède ou le suit, les analyses A_i et A_{i+1} sont pratiquement sans corrélation : tout se passe ici comme si chaque convoi était à peu près indépendant de ceux qui le précèdent ou le suivent.

Au contraire, la série des B_i a de très fortes autocorrélations. L'analyse B_i du convoi i a tendance à être supérieure à la moyenne lorsque l'analyse B_{i-1} du convoi précédent a donné un résultat élevé, et réciproquement. Plus généralement, B_i et B_{i+n} présentent des corrélations sensibles jusque vers n = 10. A raison de deux convois par jour en moyenne, on voit que ces influences s'étendent sur une période de l'ordre de 5 jours.

Ainsi les images A_i et B_i d'un même phénomène (le convoi i) sont contradictoires entre elles et incompatibles. L'image fournie par B est celle d'un phénomène structuré dans le temps (évoluant dans le temps avec un certain degré de continuité), tandis que A n'évoque aucune structure, mais plutôt une succession d'erreurs d'analyse et de prélèvements aléatoires et indépendantes les unes des autres.

On peut penser (mais sans avoir le droit de l'affirmer) que la structure révélée par la série B n'est pas une propriété des convois eux-mêmes, mais un phénomène parasite ajouté par le dispositif d'échantillonnage continu : elle pourrait correspondre à une dérive systématique dans le temps du réglage des appareils, dérive qui serait rectifiée à intervalles plus ou moins réguliers par une remise au point de ce réglage : telle pourrait être l'origine de la période de 5 jours qui régit ces autocorrélations.

e/.- L'examen des corrélations croisées permet de préciser les conclusions précédentes. En premier lieu, la corrélation entre les chiffres A_i et B_i relatifs à un même convoi est extrêmement faible, mais positive (coefficient de corrélation de + 0,15). Ce chiffre confirme qu'il n'y a qu'un rapport lointain entre A_i et B_i mais que ce rapport existe.

En deuxième lieu, nous avons examiné les corrélations croisées, c'est-à-dire les corrélations entre B_i et A_{i-k} et entre B_i et A_{i+k}. Une dissymétrie brutale est apparue, qui doit traduire un lien de causalité :

- Entre B_i et A_{i+1}, A_{i+2} ..... (c'est-à-dire entre le chiffre usine relatif au convoi i et les chiffres mines relatifs aux convois suivants $i + 1$, $i+2$..) les corrélations sont négligeables (- 0,04, - 0,08), de sorte que B_i est indépendant de A_{i+1}, A_{i+2} ...

- Au contraire, entre B_i et A_{i-1}, A_{i-2} ... A_{i-10} (c'est-à-dire entre le chiffre usine relatif au convoi i et les chiffres mines relatifs à chacun des 10 convois précédents) apparaissent des corrélations toutes positives. Chacune d'elles est relativement faible (coefficients variant de + 0,10 à + 0,20), mais, prises dans leur ensemble, elles constituent un phénomène très significatif.

Cette dissymétrie dans le temps doit s'interpréter par un lien de causalité. Tout se passe comme si B_i reflétait non pas la teneur t_i du convoi i seul, mais une moyenne pondérée du convoi i et des 10 convois antérieurs. Ceci suggère une interprétation simple : le minerai séjournerait dans l'appareil, de sorte que le prélèvement attribué au convoi i serait en réalité un mélange complexe de minerais provenant des convois antérieurs.

f/.- Ces résultats suggèrent l'interprétation suivante. Désignant par t_i la teneur moyenne réelle (inconnue) du convoi i, les chiffres mines et usine seraient donnés par des relations de la forme :

$$\begin{cases} A_i = t_i + e_i \\ B_i = \frac{1}{11}\left[t_i + t_{i-1} + \dots + t_{i-10}\right] + \varepsilon_i \end{cases}$$

En ce qui concerne A_i, chiffre mine, les e_i seraient des erreurs d'analyse et de prélèvement, aléatoires, indépendantes les unes des autres et des t_i. La variance des convois σ_t^2 serait de l'ordre de 0.4 (teneurs des convois variant dans une fourchette de $\pm$ 1,5 points), celle des e_i de l'ordre de 0.8 (erreur d'analyse et de prélèvement de $\pm$ 2 points). Les e_i se comportent comme des erreurs usuelles, mais on ne peut cependant pas affirmer que leur valeur moyenne soit nulle, c'est-à-dire qu'ils n'introduisent pas d'erreur systématique.

En ce qui concerne B_i, le terme correspondant à une moyenne (en principe, une moyenne pondérée, mais, pour simplifier, nous avons pris une simple moyenne arithmétique) des 11 convois t_i à t_{i-10} introduit une variance de l'ordre de 0.2 (teneurs des 11 convois consécutifs variant dans une fourchette de $\pm$ 1 point). L'erreur ε_i a une variance très forte, de l'ordre de 1.8 (fourchette de $\pm$ 2,7 points). Alors que la corrélation entre B_i et t_i est faible (0,15 environ), la corrélation entre B_i et la moyenne des 11 convois est notable (0.30 environ). Cette même corrélation règne entre B_i et la moyenne des 11 A_i correspondants, et ce point établit que les A_i et les B_i se rapportent bien à un même phénomène (les teneurs t_i des convois), l'image fournie par B_i ayant subi une certaine déformation.

De plus, B_i dépendant des 11 convois t_i, t_{i-1}.. t_{i-10} et B_{i+1}, par exemple, des convois t_{i+1}, t_i ... t_{i-9}, se rapportent, pour 10 onzièmes , à un même lot de minerai (t_i.. t_{i-9}), et pour 1/11 seulement à un lot différent (t_{i+1} ou t_{i-10}). Ce point rend compte, en partie, de la corrélation régnant entre B_i et B_{i+1} : mais en partie seulement.

En fait, la covariance de B_i et B_{i+1} est de 0.8, alors que l'effet de mélange des t_i, t_{i-1} ... ne justifie qu'une covariance de 0.2. Pour l'essentiel, (0.6 sur 0.8) cette covariance ne provient pas des t_i, mais des erreurs de ε_i. Il en résulte que les erreurs ε_i de l'usine ne sont pas des erreurs de type ordinaire : elles ne sont pas indépendantes les unes des autres, mais l'erreur ε_i est fortement liée à l'erreur ε_{i-1} commise sur le convoi précédent.

Cette corrélation entre erreurs successives peut sans doute s'interpréter par une dérive dans le temps du réglage des appareils (avec des remises au point périodiques). Il est plausible (mais on ne peut l'affirmer) qu'une telle dérive soit génératrice d'une cause d'erreur systématique.

Ayant résumé l'essentiel, nous allons maintenant préciser certains détails de méthode ou de calcul.

3.- Comparaison des deux laboratoires.

Pour une même prise de teneur inconnue y, le laboratoire de l'usine donne le résultat A et celui de la mine C. On a

$$\begin{cases} A = y + \alpha \\ C = y + \gamma \end{cases}$$

les erreurs α et γ étant aléatoires, indépendantes l'une de l'autre et de y, et de valeurs probables nulles (en l'absence de différence systématique des deux laboratoires).

En prenant les variances de A et C et la covariance AC, il vient :

$$\begin{cases} \sigma_A^2 = \sigma_y^2 + \sigma_\alpha^2 \\ \sigma_C^2 = \sigma_y^2 + \sigma_\gamma^2 \\ \sigma_{AC} = \sigma_y^2 \end{cases} \Longrightarrow \begin{cases} \sigma_\alpha^2 = \sigma_A^2 - \sigma_{AC} = 0.102 \\ \sigma_\gamma^2 = \sigma_C^2 - \sigma_{AC} = 0.184 \end{cases}$$

D'où les fourchettes d'erreur des labo usine ($\pm 2\ \sigma_x = \pm\ 0{,}64$) et mine ($2\ \sigma_y = \pm\ 0.86$).

4/.- Test sur la différence A - B

La différence $A_i - B_i$ a une valeur moyenne de 0.4 et une variance de 2,9215, c'est-à-dire un écart type de 1.71, pour 241 échantillons. On a :

$$\frac{0.4}{1.71} \times \sqrt{241} = 3.64$$

Cette différence de 3,6 écarts type est hautement significative (Il n'y a que 1,5 chances sur 10.000 d'en observer une aussi forte si la valeur vraie est égale à 0).

5/.- Autocorrélations.

Le fait que les A_i ne présentent pas d'autocorrélation sensible ne signifie pas nécessairement que les teneurs t_i des convois successifs soient indépendantes. Il indique simplement que, si l'on écrit $A_i = t_i + e_i$, les erreurs aléatoires e_i sont indépendantes entre elles et ont une variance nettement plus forte que les teneurs réelles t_i des convois, de sorte que l'autocorrélation des convois, si elle existe, est masquée à peu près entièrement par les erreurs aléatoires.

En fait, la covariance $\sigma(A_i A_{i+1})$ a une valeur positive faible, mais à peu près significative (0.162). Comme les e_i sont indépendants des t_i, on en déduit donc :

$$\sigma\,(t_i t_{i+1}) = \sigma(A_i\ A_{i+1}) = 0.162$$

La variance des convois est nécessairement supérieure à ce chiffre. Elle pourrait être de l'ordre de 0.3 ou 0.4.

Un autre indication dans le même sens est fournie par les valeurs moyennes $\frac{1}{10}\,(A_i + a_{i+1} + \ldots + A_{i+10})$ des analyses mines de dix convois consécutifs.

Ces moyennes mobiles ont une variance de 0.26, valeur nettement supérieure à celle que l'on observerait si les convois étaient indépendants ($\frac{1}{10}\sigma_A^2 = 0,125$).

Si l'on admet que les convois ont une variance σ_t^2 de l'ordre de 0.4, il reste pour la variance de l'erreur e_i

$$\sigma_{e_i}^2 = \sigma_A^2 - \sigma_t^2 = 1,25 - 0,4 \# 0.8$$

Dans une moyenne mobile par 10, la composante de la variance due aux e_i est $\frac{1}{10}\,0,8 = 0,08$, et il reste $0.26 - 0.08 = 0.18$ pour la variance de la moyenne mobile des teneurs vraies. Cet ordre de grandeur est compatible avec ce que l'on peut déduire des corrélations croisées (parag. suivant). Il implique, au niveau des teneurs vraies, des corrélations assez importantes entre convois consécutifs : mais celles-ci sont complètement masquées par les erreurs aléatoires, et il n'est pas possible de les calculer à partir des données disponibles.

Au contraire, la série des B_i (échantillonnage usine) est fortement autocorellée. La première covariance est de l'ordre de 0,8, donc notablement supérieure à la valeur attribuée ci-dessus à la variance des convois. Il en résulte que cette autocorrélation trouve sa source non pas dans le phénomène primaire (la série des teneurs vraies t_i des convois) mais dans un phénomène secondaire surajouté par le dispositif d'échantillonnage continu : interprétation qui s'accorde bien avec l'hypothèse d'une dérive dans le temps du réglage de ce dispositif.

4/.- Corrélations croisées.

Le contraste entre les corrélations inexistantes dans le sens $B_i A_{i+k}$ et, au contraire, faibles mais constamment positives dans le sens $B_i\,A_{i-k}$ est un fait majeur. Sans faire de test précis, il suffit d'observer qu'en l'absence de corrélation vraie la probabilité d'observer une séquence de 10 coefficients positifs serait $2^{10} \# 10^{-3}$, soit 1/1000 environ. Ce fait implique une relation causale. Naturellement, les A_{i-k} ne peuvent pas être la cause directe de B_i, mais n'influent sur lui que par l'intermédiaire des teneurs vraies t_{i-k} des convois. Comme $A_i = t_i + e_i$, les ε_i étant indépendants des t_i, et par suite aussi des B_i, la covariance des A_{i-k} avec les B_i ne diffère pas de celle des t_{i-k} avec les mêmes B_i :

$$\sigma(A_{i-k}, B_i) = \sigma(t_{i-k}, B_i)$$

La théorie de la régression linéaire montre que B_i doit être de la forme :

$$Bi = \varepsilon_i + \alpha_0 t_i + \alpha_1 t_{i-1} + \dots$$

c'est-à-dire la somme d'une erreur ε_i indépendante des t_i et d'une combinaison linéaire des t_{i-k} : d'où l'idée que l'usine n'échantillonne pas le convoi t_i mais un mélange de tous les convois antérieurs.

Un calcul précis des coefficients α_0 α_1 ... n'aurait pas grand sens, du fait que les covariances expérimentales, faibles vis-à-vis des variances, sont entachées d'importantes fluctuations aléatoires. (les corrélations entre convois sont masquées par les erreurs aléatoires). On peut tout au plus chercher à dégager des ordres de grandeur. Observant que les corrélations $\sigma(B_i\ A_{i-k})$ sont notables de $k = 0$ à $k = 10$, et changent ensuite d'allure et d'ordre de grandeur, on peut admettre, en première approximation, que les B_i ne dépendent que des 11 convois t_i, t_{i-1} .. t_{i-10} antérieurs, et, puisque l'on renonce au calcul précis des α_i, qu'ils ne dépendent que de la moyenne arithmétique de ces convois

$$y_i = \frac{1}{10}\left[t_i + t_{i-1} + \dots + t_{i-10}\right]$$

étant entendu que cette hypothèse simpliste est simplement destinée à mettre en évidence un ordre de grandeur. Dans ces conditions, B_i serait de la forme :

$$B_i = \lambda\, y_i + \varepsilon_i$$

Nous prendrons en fait $\lambda = 1$: en premier lieu, parce qu'il est plausible que B_i ne diffère de la teneur y_i échantillonnée que par une erreur ε_i indépendante de y_i. En deuxième lieu, parce que la moyenne

$$\frac{1}{11}\sum_{k=0}^{10} \sigma(B_i\ A_{i-k}) = 0{,}22$$

des covariances croisées doit être égale à $\lambda^2 \sigma_y^2$ et que la variance σ_y^2 des y_i est - nous l'avons vu - de l'ordre de 0,2, d'où $\lambda \# 1$.

5.- Interprétation (conjecturale)

On arrive ainsi à l'idée que A_i et B_i seraient de la forme :

$$\begin{cases} A_i = e_i + t_i \\ B_i = \varepsilon_i + \frac{1}{11} \sum_{k=0}^{10} t_{i-k} \end{cases}$$

Les e_i - erreurs de prélèvement et d'analyse de la mine - sont indépendants les uns des autres, des t_i et des ε_i. Les t_i présentent des autocorrélations, mais sont indépendants des erreurs e_i et ε_i.

Par contre les erreurs ε_i de l'usine ne peuvent pas être indépendantes les unes des autres. En effet, la covariance $\sigma(B_i B_{i+1}) = 0,84$ est de la forme :

$$\sigma_{B_i B_{i+1}} = \sigma_{\varepsilon_i \varepsilon_{i+1}} + \sigma_{y_i y_{i+1}}$$

Or la covariance $\sigma_{y_i y_{i+1}}$ est majorée par la variance σ_y^2, qui est de l'ordre de 0,2. Il reste donc :

$$\sigma_{\varepsilon_i \varepsilon_{i+1}} > 0,84 - 0,2 \# 0.6$$

L'autocorrélation des erreurs de prélèvement à l'usine s'interprète commodément par l'hypothèse d'une dérive dans le temps du réglage des appareils (mais cela ne constitue pas une démonstration formelle).

La variance $\sigma_B^2 = 2,03$ des B_i est la somme de la variance $\sigma_y \# 0.2$ des convois mélangés et de la variance de l'erreur ε_i :

$$\sigma_{\varepsilon_i}^2 \# 1,8$$

Cette valeur de 1,8 est forte, vis-à-vis de la première covariance $\sigma_{e_i\, e_{i+1}}$ qui est de l'ordre de 0.6. On peut en déduire l'idée que l'erreur e_i de l'usine est la somme d'une erreur aléatoire de type usuel η_i et d'un terme d_i reflétant la dérive (responsable de la totalité de l'autocorrélation des e_i), soit :

$$e_i = \eta_i + d_i$$

Comme $\sigma_{d_i d_{i+1}} = \sigma_{e_i\, e_{i+1}} \# 0.6$, on peut penser que la variance du terme de dérive est de l'ordre de 0,8 à 0,9. D'où :

$$\sigma^2_{di} \# \sigma^2_{\eta_i} \# 0.9$$

Dérive et composante aléatoire seraient d'un ordre de grandeur comparable.

G. MATHERON

25 Novembre 1963.

TABLEAU DES VALEURS NUMERIQUES EXPERIMENTALES

A/ - Variances et covariances (calculées sur 241 échantillons).

$$\begin{cases} \sigma_A^2 = 1.2585 \\ \sigma_B^2 = 2.0256 \\ \sigma_C^2 = 1.3409 \end{cases} \qquad \begin{cases} \sigma_{AB} = 0.1813 \\ \sigma_{BC} = 0.1397 \\ \sigma_{CA} = 1.1566 \end{cases} \qquad \begin{cases} \rho_{AB} = +\ 0.115 \\ \rho_{BC} = +\ 0.108 \\ \rho_{CA} = +\ 0.89 \end{cases}$$

B/ - Covariances et autocorrélations des A_i et des B_i

$\sigma(B_i\ B_{i+1}) = 0.8381 \longrightarrow \rho_1 = 0.413$

$\sigma(B_i\ B_{i+2}) = 0.5951 \longrightarrow \rho_2 = 0.296$

$\sigma(B_i\ B_{i+3}) = 0.4533 \longrightarrow \rho_3 = 0.224$

$\sigma(B_i\ B_{i+4}) = 0.3762 \longrightarrow \rho_4 = 0.185$

$\sigma(B_i\ B_{i+6}) = 0.3493 \longrightarrow \rho_6 = 0.173$

$\sigma(B_i\ B_{i+8}) = 0.0483 \longrightarrow \rho_8 = 0.024$

$\sigma(B_i\ B_{i+10}) = 0.1050 \longrightarrow \rho_{10} = 0.052$

$\sigma(A_i\ A_{i+1}) = +\ 0.1624 \longrightarrow \rho_1 = +\ 0.129$

$\sigma(A_i\ A_{i+2}) = -\ 0.0594 \qquad \rho_2 = -\ 0.040$

$\sigma(A_i\ A_{i+3}) = -\ 0.0401 \qquad \rho_3 = -\ 0.032$

$\sigma(A_i\ A_{i+4}) = +\ 0.1333 \qquad \rho_4 = +\ 0.105$

$\sigma(A_i\ A_{i+5}) = +\ 0.1086 \qquad \rho_5 = +\ 0.087$

$\sigma(A_i\ A_{i+6}) = +\ 0.1163 \qquad \rho_6 = +\ 0.093$

C/ Covariances et corrélations croisées.

$$\left\{\begin{array}{lll} \sigma(B_i\,A_{i+2}) = -\,0.096 & \longrightarrow & \rho_{-2} = -\,0.060 \\ \sigma(B_i\,A_{i+1}) = -\,0.051 & \longrightarrow & \rho_{-1} = -\,0.032 \end{array}\right.$$

$$\left\{\begin{array}{lll} \sigma(B_i\,A_i) = +\,0.1813 & \longrightarrow & \rho_0 = +\,0.113 \end{array}\right.$$

$$\left\{\begin{array}{lll} \sigma(B_i\,A_{i-1}) = +\,0.2350 & \longrightarrow & \rho_1 = +\,0.146 \\ \sigma(B_i\,A_{i-2}) = +\,0.3289 & \longrightarrow & \rho_2 = +\,0.205 \\ \sigma(B_i\,A_{i-3}) = +\,0.3296 & \longrightarrow & \rho_3 = +\,0.205 \\ \sigma(B_i\,A_{i-4}) = +\,0.1500 & \longrightarrow & \rho_4 = +\,0.093 \\ \sigma(B_i\,A_{i-5}) = +\,0.1484 & \longrightarrow & \rho_5 = +\,0.092 \\ \sigma(B_i\,A_{i-6}) = +\,0.1999 & \longrightarrow & \rho_6 = +\,0.125 \\ \sigma(B_i\,A_{i-7}) = +\,0.1530 & \longrightarrow & \rho_7 = +\,0.095 \\ \sigma(B_i\,A_{i-8}) = +\,0.3439 & \longrightarrow & \rho_8 = +\,0.214 \\ \sigma(B_i\,A_{i-9}) = +\,0.0991 & \longrightarrow & \rho_9 = +\,0.062 \\ \sigma(B_i\,A_{i-10}) = +\,0.2665 & \longrightarrow & \rho_{10} = +\,0.166 \end{array}\right.$$

$$\left\{\begin{array}{lll} \sigma(B_i\,A_{i-11}) = +\,0.0192 & \longrightarrow & \rho_{11} = +\,0.012 \\ \sigma(B_i\,A_{i-13}) = -\,0.1260 & \longrightarrow & \rho_{13} = -\,0.079 \\ \sigma(B_i\,A_{i-15}) = -\,0.0089 & & \rho_{15} = -\,0.006 \end{array}\right.$$