



# REMARQUES SUR LES CALCULS STATISTIQUES POUR L'ÉVALUATION DE LA QUALITÉ DE L'EAU

Chantal de FOUQUET

---

Rapport N-04/05/G

Mars 2005

**CENTRE DE GÉOSTATISTIQUE**  
35 rue Saint-Honoré  
77305 FONTAINEBLEAU (France)  
Tél. : 33-1 64 69 47 81 Fax : 33-1 64  
69 47 05 <http://cg.ensmp.fr>

# Remarques sur les calculs statistiques pour l'évaluation de la qualité de l'eau

Chantal de FOUQUET

ECOLE des Mines de Paris/ARMINES – Centre de Géostatistique

fouquet@cg.ensmp.fr

## 1. Deux notions fondamentales

Les méthodes statistiques de calcul d'indicateurs de la qualité de l'eau reposent sur deux hypothèses simplificatrices mais abusives.

- elles ne tiennent pas compte de la corrélation temporelle des mesures, mais supposent les valeurs non corrélées (ou indépendantes). Ceci amène à ne pas tenir compte des dates de mesure, mais seulement de leur nombre, dans les calculs de précision.

- il y a confusion entre valeurs estimées, seules connues, et valeurs vraies, qui restent toujours inconnues sauf pour les échantillons mesurés.

### 1.1. Corrélation temporelle

La corrélation temporelle des concentrations a été démontrée, par exemple dans l'étude ENSMP-ARMINES effectuée avec le Bureau BETURE pour le compte de l'IFEN en 1999-2000.

On rappelle le variogramme temporel des Nitrites pour l'un des bassins étudiés (figure 1.). A une composante périodique annuelle se rajoute une corrélation temporelle, composée de deux structures : l'une de portée inférieure à l'année, l'autre de portée manifestement pluri-annuelle.

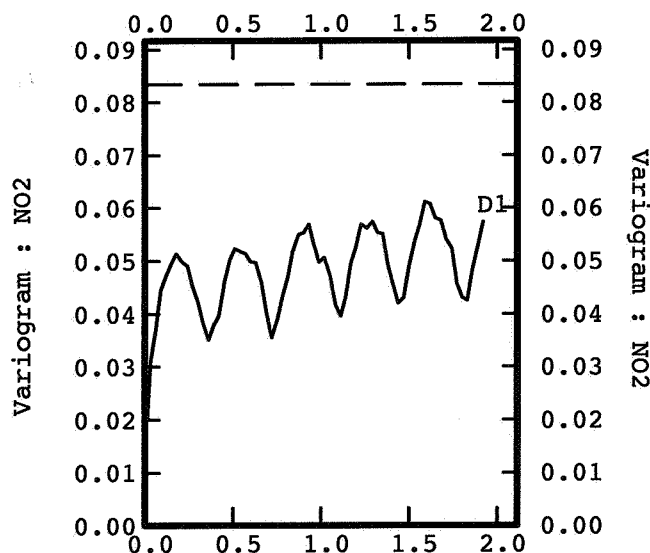


Figure 1. Variogramme temporel moyen des Nitrites. Bassin 2. En abscisse, le nombre de milliers de jours. La périodicité annuelle indique l'échelle temporelle.

En l'absence de corrélation temporelle, le variogramme serait pépitique, c'est-à-dire "plat". Ce n'est manifestement pas le cas.

Du fait de la corrélation temporelle, les mesures ne peuvent être considérées comme des tirages indépendants d'une variable aléatoire de loi statistique fixée mais inconnue. C'est pourtant sur cette hypothèse que reposent les calculs statistiques : deux concentrations aussi proches soient-elles dans le temps, admettent la même variabilité conjointe que si elles sont espacées d'un mois, d'un semestre ou d'une année. La variance de l'écart entre les deux concentrations est supposée ne pas dépendre de l'intervalle de temps entre les deux prélèvements.

Si l'on néglige momentanément la composante périodique, la corrélation temporelle signifie au contraire que deux prélèvements rapprochés dans le temps sont en moyenne plus proches que deux prélèvements espacés dans le temps. La précision des estimations de moyenne annuelle, par exemple ne dépend alors pas seulement du nombre des mesures, mais également de leurs dates.

En présence de corrélation temporelle, 12 prélèvements effectués 12 jours (ou semaines) successifs fourniront une estimation de la moyenne annuelle beaucoup moins précise que si ces prélèvements sont effectués tous les mois. Sous l'hypothèse d'indépendance des tirages, utilisée dans les calculs statistiques, la précision serait identique.

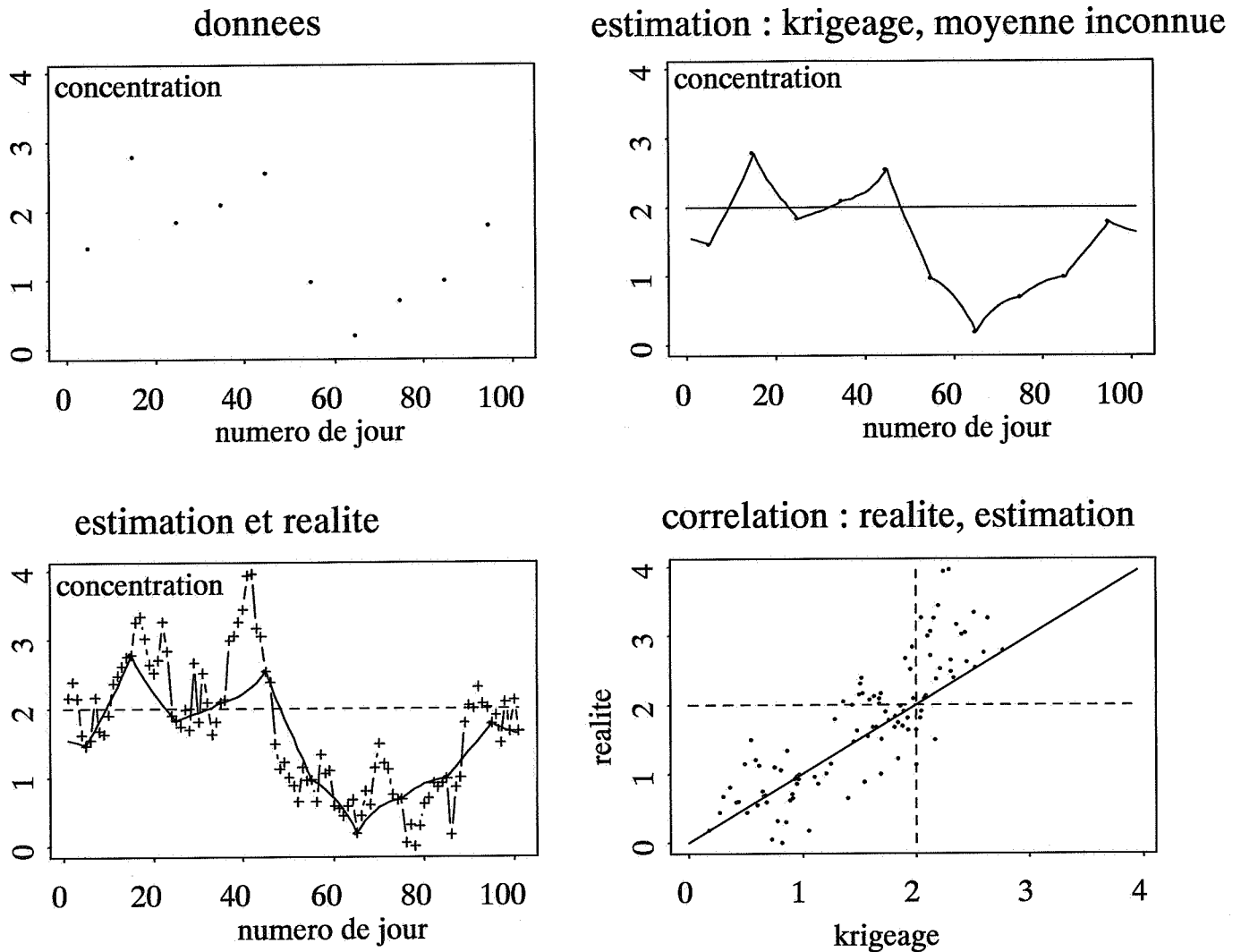
### 1.2. Valeur estimée et valeur vraie

Entre deux mesures, les concentrations réelles restent toujours inconnues. Seules des estimations de ces concentrations sont accessibles.

Quand on compare la courbe des valeurs **estimées** à un seuil (figure 2), cette courbe diffère de celle des concentrations **réelles**, qui est pourtant ce qui nous intéresse. Or l'écart entre valeur estimée et valeur réelle est d'autant plus important que l'information est peu précise, ou que la variabilité temporelle est importante.

On montre que l'histogramme des valeurs estimées diffère nécessairement de celui des concentrations réelles. La moyenne est identique (si l'estimation est sans biais), mais la variance et donc les quantiles sont différents. Sur la figure 2, la proportion des valeurs dépassant le seuil=3 est manifestement bien plus forte sur les valeurs vraies que sur les valeurs estimées. La concentration moyenne réelle des valeurs dépassant le seuil 2 est bien supérieure à la moyenne des valeurs estimées dépassant le seuil. Du seul fait que les concentrations réelles restent inconnues, en travaillant sur des estimations de ces concentrations, on se trompe sur dates de dépassement de seuil, et sur la moyenne de la concentration durant ces périodes critiques.

Calculer les quantiles directement à partir valeurs mesurées (SEQ-eau) reviendrait, entre deux mesures, à estimer la concentration par des segments de valeur constante. C'est évidemment moins précis que le krigeage utilisé ici.



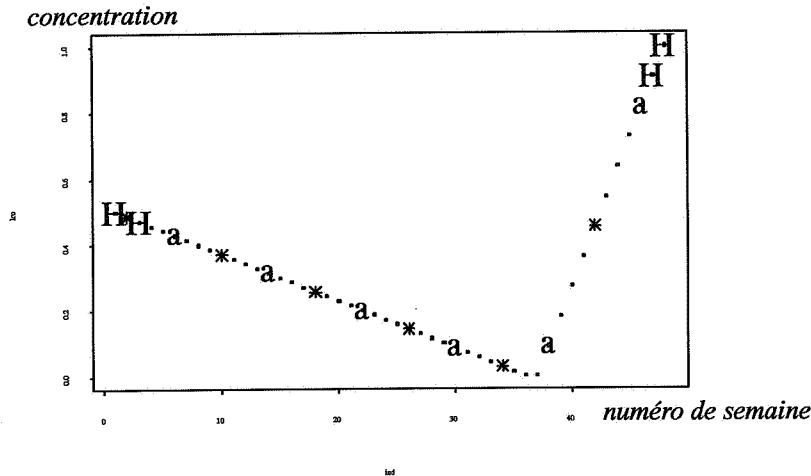
Le seuil critique est ici égal à 2.

Figure 2. Mesures, estimation et réalité

## 2. Quelques limitations des méthodes actuelles

Posons pour simplifier : 1 mois=4 semaines, 1an=12 mois=48 semaines.

La concentration en un paramètre, supérieure en hiver, décroît lentement de l'hiver à l'été, pour augmenter fortement en automne (figure 3). La concentration est mesurée tous les deux mois, les semaines 2 10 18 26 34 42.



Les "points" indiquent les mesures hebdomadaires. Les mesures bimestrielles sont notées par \*, les autres mesures mensuelles par un "a", et le resserrement hivernal par H.

**Figure 3. Variation de la concentration durant une année (exemple synthétique).**

## 2.1. La moyenne

Dans la figure 3, la moyenne est de 0.29 pour les mesures bimestrielles, de 0.31 pour les mesures mensuelles, et de 0.41 avec un renforcement de la surveillance en hiver (la moyenne hebdomadaire des 48 semaines est de 0.31). En fait, on pourrait artificiellement relever ou abaisser l'estimation de la moyenne annuelle, calculée comme la simple moyenne arithmétique des concentrations mesurées, en augmentant la fréquence en période de concentration forte ou faible, puisque ce calcul ne tient pas compte de la date des mesures.

Ce calcul pénalise ainsi la surveillance renforcée des périodes critiques.

En cas de renforcement des mesures en période forte ou faible, pour éviter une surestimation ou une sous-estimation, donc un biais, dans le calcul de la moyenne, il faut pondérer les valeurs en tenant compte de la fréquence variable des mesures.

Par ailleurs, en termes de précision, il n'est manifestement pas équivalent de regrouper les mesures ou de les répartir régulièrement au cours de l'année.

## 2.2. Calcul classique de la précision sur la moyenne

Soient  $n$  mesures  $Z_\alpha$ ,  $\alpha = 1, \dots, n$  au cours du temps  $T$ . La moyenne  $m$  est estimée par  $m^* = \frac{1}{n} \sum_{\alpha} Z_\alpha$  et la précision sur la moyenne est donnée par  $\text{Var } m^* = \frac{1}{n} D^2$  en notant  $D^2$  la variance expérimentale des données.

D'où vient ce calcul ? Supposons les  $Z_\alpha$  des Variables Aléatoires **indépendantes** (il suffit qu'elles soient **non corrélées**) et de même loi de probabilité, d'espérance  $m$  inconnue et de variance  $\sigma^2$ , c'est-à-dire  $\forall \alpha$ ,  $E[Z_\alpha] = m$  et  $E[(Z_\alpha - m)^2] = \sigma^2$ . Alors  $m^* = \frac{1}{n} \sum_{\alpha} Z_\alpha$  est un estimateur sans biais du paramètre moyenne  $m$  :

$$\begin{aligned}
 E[m - m^*] &= m - \frac{1}{n} \sum_{\alpha} E[Z_{\alpha}] \\
 &= m - \frac{1}{n} \sum_{\alpha} m \\
 &= 0
 \end{aligned}$$

Variance de l'erreur d'estimation : comme  $m$ , le paramètre inconnu, est une constante

$$\begin{aligned}
 \text{Var}[m - m^*] &= \text{Var } m^* \\
 &= \frac{1}{n^2} \sum_{\alpha} \text{Var } Z_{\alpha} \\
 &= n D^2 / n^2 \\
 &= D^2 / n
 \end{aligned}$$

Dans ce calcul, la variance de la somme est la somme des variances, puisque les  $Z_{\alpha}$  sont supposés non corrélés.  $D^2$  est l'estimateur de la variance, soit usuellement (pour tenir compte du fait que  $m$  n'est pas connu mais estimé par  $m^*$ ) :

$$D^2 = \frac{1}{n-1} \sum_{\alpha} (Z_{\alpha} - m^*)^2$$

Ce calcul ne **tient pas compte de l'implantation, régulière ou non, des mesures**. La précision est censée être identique avec six mesures bimestrielles dans une année, ou six mesures regroupées les six derniers jours de l'année! Cela résulte de l'hypothèse d'absence de corrélation des  $n$  mesures  $Z_{\alpha}$ .

Remarque : à cause de la corrélation temporelle, la variance expérimentale des mesures sur 6 jours successifs sera inférieure à celle de six mesures bimestrielles. Par suite, le calcul statistique indiquerait une meilleure précision dans le premier cas que dans le second!

Ce calcul ignore la corrélation temporelle des concentrations : en réalité, la concentration ne varie pas de façon erratique (le bruit blanc des physiciens), mais est corrélée temporellement (voir par exemple le variogramme temporel des nitrites, figure 1).

### Quelle "moyenne" ?

Plus important, si on estime le paramètre moyenne de la loi de probabilité, **on se trompe de variable!** En fait, on veut connaître la **moyenne temporelle**  $Z_T = \frac{1}{T} \int_T Z(t) dt$ , variable d'une année à l'autre, alors que si la concentration est supposée stationnaire, la moyenne statistique de la loi de probabilité est constante. En termes probabilistes, supposant la concentration stationnaire, on a :  $E[Z(t)] = E[Z_T] = m$

Relation entre paramètre moyenne  $m$  et moyenne temporelle : lorsque la concentration est supposée stationnaire (donc d'espérance constante), si l'intervalle de temps  $T$  devient très grand, alors la moyenne temporelle converge vers la moyenne statistique (pour les modèles probabilistes usuels).

Mais supposons une variation linéaire de la moyenne de la concentration au cours du temps :  $m(t) = E[Z(t)] = at + b$ , cf. localement, figure 3. La moyenne temporelle

$Z_T = \frac{1}{T} \int_T Z(t) dt$  de la concentration durant l'intervalle de temps T est encore définie. Mais quel est alors le sens du paramètre m supposée constant ? Le calcul  $m^* = \frac{1}{n} \sum_{\alpha} Z_{\alpha}$  est donc bien plutôt une estimation de  $Z_T$ , moyenne annuelle, que d'un hypothétique paramètre m.

### 2.3. La règle des 90 %

Pour évaluer la qualité dans des conditions critiques mais non exceptionnelles, la règle dite des 90% consiste à retenir 90% des mesures observées sur une période. Les valeurs en conditions critiques peuvent alors être comparées d'une année à l'autre.

Cette méthode pénalise la surveillance renforcée des périodes critiques.

Reprenons l'exemple de la figure 3. La concentration est mesurée tous les deux mois, les semaines 2 10 18 26 34 42. Avec six mesures, la règle du SEQ-eau, préconise de retenir la plus forte des concentrations, soit : 0.486.

Avec une mesure par mois (\* et a), le maximum mesuré est de 0.818, mais l'on retient alors la mesure de rang 11, soit à nouveau 0.486.

Avec 48 valeurs hebdomadaires, la valeur à retenir serait celle de rang 44, qui est ici la moyenne de 0.486 et 0.818. En supprimant le maximum des 12 mesures mensuelles, on a donc perdu de l'information.

Pour surveiller la période critique, on resserre maintenant les mesures en hiver, pour les semaines 1 3 47 48 (H), ce qui donne au total 16 mesures. La mesure maximale est égale à 1 ; on conserve la mesure de rang 15, soit 0.909.

Conclusion : cette règle pénalise l'observateur consciencieux, qui resserre les mesures en période critique. Il rajoute alors des valeurs fortes, et nécessairement le quantile augmente, puisque l'on ne tient pas compte de la date des mesures, mais seulement de leur nombre et de leur rang. Il faudrait "pondérer" les valeurs suivant l'intervalle de temps entre deux mesures : si on double le nombre de mesures durant une période, chacune devrait "compter" pour moitié moins, dans le calcul du rang ("poids" diminué de moitié).

Le rajout de mesures en période de basse concentration crée également un artefact. Pour diminuer le quantile à 90%, il suffit de rajouter suffisamment de mesures moyennes ou faibles.

### 2.4. Information portée par les valeurs supprimées

Considérons les deux séries de mesures mensuelles, qui diffèrent par la dernière valeur (figure 4) : 0.0286 0.0857 0.0909 0.143 0.2 0.257 0.314 0.371 0.429 0.455 0.486 0.818 et 1.718 pour la série se terminant par des valeurs fortes. Selon la règle actuelle des 90%, on supprime la 12ème valeur. La valeur critique (0.486) est identique pour les deux séries, alors que la dernière mesure indiquait une valeur nettement supérieure en fin d'année dans le deuxième cas.

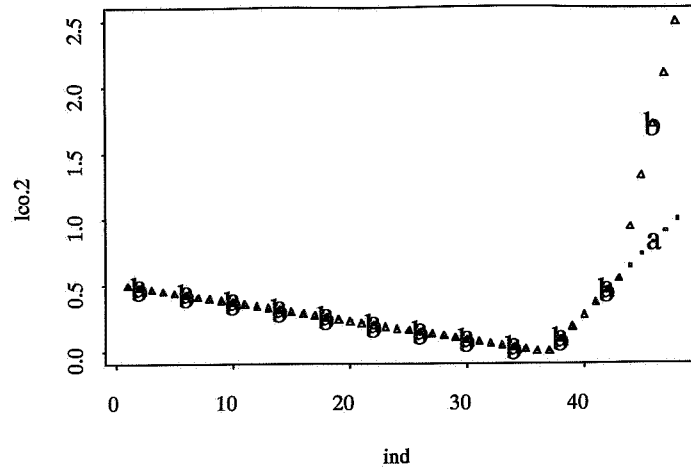


Figure 4. Information apportée par la mesure maximum.

Supprimer la valeur de rang 12 fournit, pour les deux séries, la même probabilité de dépassement du seuil égal à 1.2, alors que cette valeur est largement dépassée dans un cas, et ne l'est jamais dans l'autre. La valeur de rang 12 indiquait effectivement que, dans un cas, le seuil de 1.2 était dépassé.

Les valeurs les plus fortes fournissent une information sur la qualité en situation critique, mais non exceptionnelle.

Pour distinguer les périodes de valeurs fortes des situations exceptionnelles, on peut par exemple examiner les intervalles interquantiles.

### 3. Quelques notions de géostatistique

#### 3.1. Le krigage

C'est "le meilleur estimateur linéaire", au sens de la minimisation de la variance de l'erreur de mesure, donc suivant un critère de précision.

Soit  $Z$  le paramètre,  $x_\alpha$  les points expérimentaux. Les données sont les mesures aux points expérimentaux :  $Z(x_\alpha) = Z_\alpha$ .

On estime  $Z_0$ , qui est par exemple

- la valeur "ponctuelle" à un instant  $x$  :  $Z_0 = Z(x)$

- la moyenne durant un intervalle  $T$  :  $Z_0 = \frac{1}{T} \int_T Z(x) dx$

en gardant pour simplifier la même notation  $Z_0$ .  $Z_0$  est inconnu, son estimation est notée  $Z_0^*$ . Evidemment, on sait que  $Z_0^* \neq Z_0$ .

On impose :

-  $Z_0^*$  est, par simplicité, une combinaison linéaire des données :  $Z_0^* = c + \sum_{\alpha} \lambda_{\alpha}^0 Z_{\alpha}$

- le non biais, ni surestimation ni sous-estimation systématique. L'espérance de l'erreur d'estimation est nulle  $E[Z_0^* - Z_0] = 0$ ,  $E$  désignant la moyenne probabiliste.

- la “meilleure précision” :  $\text{Var}[Z_0^* - Z_0]$  minimum.

Cela conduit à un système linéaire déterminant les poids de krigeage  $\lambda_\alpha^0$ . Ces poids dépendent de ce que l’on estime (valeur ponctuelle, moyenne temporelle, voire même le paramètre statistique moyenne de la loi de probabilité). Ils dépendent aussi de l’implantation réciproque des données (redondance locale d’information, par exemple) et de la grandeur à estimer (mais non de sa valeur). Enfin, les poids dépendent de la corrélation spatiale ou temporelle.

Cette corrélation spatiale ou temporelle est décrite par le variogramme. Pour d’importantes questions d’inférence, on calcule le variogramme plutôt que la covariance.

Le krigeage multivariable est dit cokrigeage.

En l’absence de corrélation temporelle, le krigeage coïncide avec les calculs statistiques usuels.

La **variance de krigeage** dépend également de la grandeur à estimer (concentration “instantanée”, moyenne semestrielle ou annuelle, paramètre  $m$  de la loi de probabilité ; mais la variance de krigeage ne dépend pas de la valeur de cette grandeur), de l’implantation relative des mesures entre elles ainsi que par rapport à l’intervalle (ou la date) à estimer, et de la structure spatiale. Par conséquent, la variance de krigeage dépend de la date des mesures, et pas seulement de leur nombre.

La variance de krigeage peut donc être vue comme un indicateur de précision, ou de qualité de l’estimation.

### 3.2. D’autres estimateurs

Soit à prévoir si une valeur inconnue  $Z(x)$  dépasse ou non un seuil  $s$  fixé. Quelle que soit la méthode d’estimation de  $Z(x)$ , on commet des biais parfois très importants en comparant  $Z^*(x)$  à  $s$ . Si  $Z^*(x)$  est le krigeage, c’est plutôt “moins pire” que pour d’autres estimateurs. La solution consiste alors à estimer directement  $P(Z(x) > s)$ , à partir des données  $Z_\alpha$ . Pour cela, on fait appel à des classes un peu plus vastes d’estimateurs que les seules combinaisons linéaires des données, comme par exemple le krigeage disjonctif.

## 4. Exemples

Nous illustrons la variance d’estimation et le krigeage par des exemples synthétiques simplifiés. Nous considérons seulement deux exemples de variabilité spatiale : l’effet de pépité, qui correspond à l’absence de corrélation spatiale :  $\gamma(0) = 0$  ;  $h > 0$ ,  $\gamma(h) = \sigma^2$ , et le variogramme linéaire  $\gamma(h) = \omega h$ . On pose pour simplifier  $\omega = 1$ .

On estime la concentration moyenne sur une année, représentée par un segment de longueur  $T$ .

Pour comparer les résultats, on suppose que pour les deux variogrammes, la variance expérimentale d’un grand nombre de prélèvements effectués durant l’année serait identique. Cette variance “expérimentale” des mesures est la variance de dispersion.

La variance de dispersion d’un échantillon parmi  $n$ , implantés régulièrement au centre d’intervalles successifs de longueur  $T/n$  s’écrit, pour un variogramme linéaire de pente 1:

$$D^2(1/N) = \frac{T}{n} \frac{1}{3} \frac{n^2 - 1}{n}$$

$$= \frac{T}{3} \left(1 - \frac{1}{n^2}\right)$$

Pour un effet de pépité, la variance de dispersion d'un échantillon parmi  $n$  s'écrit (quelle que soit la longueur du champ et l'implantation des échantillons dans le champ) :

$$D^2(1/N) = \sigma^2 \left(1 - \frac{1}{n}\right)$$

On posera pour la suite :  $\sigma^2 = \frac{T}{3}$ . Pour simplifier, on prendra  $T=1$ .

#### 4.1. Précision et date des mesures

On compare l'estimation de la moyenne annuelle  $Z_T = \frac{1}{T} \int_0^T Z(t)dt$  dans les cas suivants :

- une mesure centrale en  $T/2$
- une mesure en début ou en fin d'année en  $t=0$  ou  $t=T$
- deux mesures en début et fin d'année ( $t=0$  et  $t=T$ )
- six mesures centrées sur des intervalles de longueur  $T/6$
- douze mesures centrées sur des intervalles de longueur  $T/12$ .

Dans le cas d'un effet de pépité, la date des mesures durant l'année  $n$  n'intervient pas, seul compte le nombre. Par suite la précision sur la moyenne annuelle est identique dans les deux premiers cas.

- Estimation d'un segment  $T$  par  $n$  mesures intérieures "ponctuelles", dans le cas d'un effet de pépité :  $\text{Var}(Z_T - Z_T^*) = \frac{\sigma^2}{n}$

Les résultats sont récapitulés au tableau suivant.

| dates des mesures | effet de pépité |         | variogramme linéaire |         |
|-------------------|-----------------|---------|----------------------|---------|
| $T/2$             | $1/3$           | $0.333$ | $1/6$                | $0.167$ |
| $t=0$ ou $t=T$    |                 |         | $2/3$                | $0.667$ |
| $t=0$ et $t=T$    | $1/6$           | $0.167$ | $1/6$                | $0.167$ |
| 6 bimestrielles   | $1/18$          | $0.056$ | $1/36$               | $0.028$ |
| 12 mensuelles     | $1/36$          | $0.028$ | $1/6 * 1/12$         | $0.014$ |

#### Variance d'estimation de la moyenne annuelle selon le variogramme

Dans le cas du variogramme linéaire, la variance d'estimation ne varie pas nécessairement en sens inverse du nombre des prélèvements. La date des mesures intervient également.

#### 4.2. Krigeage de la moyenne annuelle par deux prélèvements.

Cet exemple simpliste de l'estimation de la moyenne annuelle par seulement deux mesures permet de montrer comment le krigeage tient compte de la date des mesures, et comment, à même nombre d'informations, la précision varie suivant ces dates.

On dispose de deux échantillons, prélevés aux instants  $t$  et  $t'$ , avec  $0 < t < t' < T$ . On estime la moyenne annuelle par une combinaison linéaire des deux mesures :  $Z_T^* = \lambda Z_t + \lambda' Z_{t'}$

On cherche les “poids”  $\lambda$  et  $\lambda'$  assurant la meilleure précision. La variance d'estimation associée est la variance de krigeage  $\sigma_K^2$  (variance pour les poids optimaux).

- Effet de pépité : quelles que soient les dates  $t$  et  $t'$ , on a :  $\lambda = \lambda' = \frac{1}{2}$  et la variance d'estimation est comme on l'a vu :  $\sigma_{\text{est}}^2 = \frac{\sigma^2}{2}$ , d'où ici  $\sigma_{\text{est}}^2 = \frac{1}{6}$ .

On retrouve le résultat du calcul statistique classique, en divisant l'effet de pépité par le nombre de points de mesures.

- Variogramme linéaire : on considère les trois cas suivants

$t=0$  et  $t=T$ . Les deux poids sont égaux à  $1/2$  ; variance de krigeage (cf. plus haut)  
 $\sigma_K^2 = \frac{1}{6} \approx 0.167$

$t=1/3$  et  $t'=2/3$ . Les deux poids sont égaux à  $1/2$  ; variance de krigeage  $\sigma_K^2 = \frac{1}{18} \approx 0.056$ .  
Pour ce variogramme, à même nombre de mesures, cette configuration est plus précise que la précédente.

$t=0$  et  $t'=1/2$  (information durant le premier semestre seulement).  $\lambda = \frac{1}{4}$ ,  $\lambda' = \frac{3}{4}$ .  
L'information située au milieu du segment à estimer reçoit un poids supérieur à celui de l'information située à l'extrémité du segment. Le krigeage atténue la redondance d'information. Variance de krigeage  $\sigma_K^2 = \frac{50}{96} \approx 0.104$ . Cette variance est inférieure à celle obtenue avec l'un des deux points seulement, mais elle n'est pas deux fois plus faible.

Conclusion : Suivant l'implantation des deux mesures, la précision sur la moyenne annuelle est très différente. Lorsque l'information n'est pas implantée symétriquement par rapport au segment à estimer, la moyenne arithmétique n'est plus nécessairement optimale. Le krigeage permet de tenir compte des dates des mesures dans le calcul de l'estimateur (pondération optimale) ainsi que dans celui de la précision (variance d'estimation).

En résumé, les méthodes permettant une meilleure estimation, et fournissant une indication de la précision sont disponibles. Traduire ces méthodes en règles de calcul systématiques, donc “aveugles” n'est pas immédiat. Il est cependant important d'examiner sur des exemples réels les “performances” des procédés statistiques actuels.

### Références bibliographiques

G. Saporta. Probabilités, analyse des données et statistiques. Ed. Technip, 1990.

G. Matheron. La théorie des variables Régionalisées, et des applications. Les cahiers du Centre de Morphologie Mathématique, ENSMP, 1970.

### **Pour des améliorations aux calculs des comptes de l'eau**

Les calculs actuels de la moyenne annuelle de la concentration, de la précision associée à cette estimation, ainsi que des quantiles, reposent sur une hypothèse implicite : les concentrations mesurées sont considérées comme des tirages d'une variable aléatoire, indépendants les uns des autres. L'absence de corrélation des concentrations intervient explicitement dans le calcul de la variance d'estimation de la moyenne. A cause de l'indépendance supposée des concentrations, la date des différentes mesures n'est pas prise en compte dans ces calculs.

Or, les chroniques montrent clairement que les concentrations ne varient pas de façon totalement désordonnée, mais présentent des périodes de concentration plus ou moins forte ou faible. La corrélation temporelle des concentrations, qui traduit ce caractère simultanément erratique et structuré, a été en particulier démontrée par le calcul du variogramme temporel des concentrations, lors d'une étude effectuée pour l'IFEN en 1999. Pour les nutriments étudiés, à une composante périodique annuelle, se superposent différentes échelles de corrélation temporelle, variables de quelques semaines à plusieurs années.

La précision de l'estimation de la moyenne annuelle de la concentration dépend de l'importance relative des différentes échelles de variabilité temporelle, ainsi que de la densité de l'échantillonnage. En particulier, à cause de la corrélation temporelle, il n'est manifestement pas équivalent en terme de précision, d'effectuer un échantillonnage mensuel ou de regrouper ces 12 mesures, par exemple durant le seul mois de janvier. Or les calculs statistiques de précision sur la moyenne, où l'on divise la variance par le nombre de mesures, ne permettent aucune distinction entre ces deux échantillonnages. De plus, comme la dispersion des 12 mesures au cours d'un même mois est inférieure à celle des 12 mesures réparties dans l'année, les calculs actuels concluront, à tort, à une meilleure précision de la moyenne pour un échantillonnage concentré sur un mois.

La moyenne arithmétique des valeurs mesurées entraîne une augmentation artificielle de l'estimation de la moyenne annuelle en cas de resserrement de l'échantillonnage en période de fortes valeurs. Pour éviter ces biais, il serait nécessaire de pondérer les valeurs en fonction de la fréquence locale de l'échantillonnage.

Le calcul de quantiles d'après l'histogramme des mesures, sans considération de la date des prélèvements, pénalise là encore la surveillance des épisodes de valeurs fortes. Plutôt que de supprimer systématiquement les concentrations maximales dans le calcul des quantiles, ce qui conduit à une perte d'information, une distinction entre valeurs fortes ou exceptionnelles pourrait sans doute s'effectuer sur la base des écarts interquantiles.

Une réponse pertinente à ces questions complexes passe par des méthodes plus élaborées que les calculs "simples" actuels, mais dont les fondements méthodologiques sont fragiles.

Elaborée en contexte spatial, la géostatistique offre des réponses aux problèmes de l'estimation, précisément pour des variables dont les fluctuations sont simultanément erratiques et structurées. L'estimateur linéaire, sans biais et optimal en terme de précision, est le krigeage. La variance d'estimation associée quantifie la précision de l'estimation. Les estimateurs non linéaires (krigeage disjonctif, espérance conditionnelle) fournissent la probabilité de dépassement d'un seuil limite, ainsi que l'estimation de la moyenne des valeurs dépassant le seuil. La géostatistique, élaborée dès les années 1960 pour l'estimation linéaire, offre un formalisme cohérent pour améliorer en pratique la qualité des estimations.