VERS DES INDICATEURS GÉOSTATISTIQUES POUR LES COMPTES DE LA QUALITÉ DE L'EAU

Caroline Bernard-Michel & Chantal de Fouquet

Ecole des Mines de Paris, Centre de Géosciences-Géostatistique
35, rue Saint-Honoré. 77305 Fontainebleau Cedex.

Résumé

Les recommandations issues de la directive cadre européenne sur l'eau préconisent des indicateurs de qualité construits à partir des statistiques élémentaires de la moyenne annuelle et du quantile 90. Mais l'analyse exploratoire des mesures de concentrations en nutriments (nitrates etc.) par station montre que les hypothèses implicites servant de fondement à ces calculs ne sont pas vérifiées. Les concentrations présentent généralement une composante périodique annuelle, ainsi qu'une corrélation temporelle mise en évidence par les variogrammes expérimentaux. Comment définir alors la distribution dont on infère moyenne et quantile, et quelles hypothèses de stationnarité sont réellement nécessaires ?
La moyenne annuelle, définie comme une intégrale temporelle, est estimée par krigeage. La pondération obtenue, associée à l'interpolation linéaire de la fonction de quantile empirique, permet de réduire fortement le biais sur les quantiles ; mais cette estimation reste souvent imprécise.

Abstract

The recommendations stemming from the European "water framework directive" are to build quality indicators from elementary statistics on the annual mean and the 90 percentile.
But the exploratory analysis of the nutrient concentrations measurements (nitrates etc.) per station shows that the implicit usual hypotheses are not verified. The concentrations present an annual periodic component as well as time correlation, revealed by the experimental variograms. How to define then the distribution on which mean and quantile are inferred, and what hypotheses on stationnarity are really necessary ?
The annual mean value, defined as a time integral, is estimated by kriging. The corresponding weighting associated with a linear interpolation of the quantile function, allows reducing strongly the bias on the quantile. But its estimation remains often imprecise.

Mots-clés

environnement, qualité de l'eau, indicateurs, géostatistique, krigeage, moyenne annuelle, quantile, concentrations en nutriments

INTRODUCTION

En cohérence avec les recommandations issues de la directive cadre européenne, le système d'évaluation de la qualité de l'eau des cours d'eau (SEQ-Eau, 2003), vise à fournir un diagnostic précis sur l'aptitude de l'eau à différents usages, et à faciliter l'évaluation des évolutions interannuelles des substances examinées. Les préconisations actuelles pour le calcul de la moyenne et du quantile 90 renvoient aux calculs statistiques élémentaires :

- la moyenne annuelle est estimée par la moyenne arithmétique de l'échantillon, la variance d'estimation étant égale à la variance de l'échantillon divisée par l'effectif ;
- le quantile 90 est estimé par le quantile empirique.

Ces calculs reposent sur les hypothèses classiques mais implicites, injustifiées dans le cas des concentrations en nutriments en une station au cours du temps. Les concentrations sont en effet

supposées sans corrélation temporelle et statistiquement homogènes durant l'année. Or, à cause des variations de débit et du cycle de la végétation, de nombreuses substances présentent des variations saisonnières, avec des écarts parfois importants entre valeurs « estivales » et « hivernales ». D'autre part, les variogrammes montrent la présence d'une corrélation temporelle des concentrations en nutriments. Un échantillonnage saisonnier préférentiel peut alors fournir des estimations biaisées de la moyenne annuelle et des quantiles.

Nous examinons le calcul de la moyenne annuelle et du quantile 90 par station, en présence d'une composante périodique et d'une corrélation temporelle des concentrations.

RETOUR SUR LA "MOYENNE ANNUELLE"

Les données expérimentales confirment que les concentrations en nitrates « hivernales » sont plus élevées en "hiver" qu'en "été", à cause du ralentissement de l'activité végétale et d'un ruissellement plus important (Figure 1.a). Les variogrammes expérimentaux présentent généralement une composante temporelle structurée se rajoutant à une composante périodique annuelle (Figure 1.b).

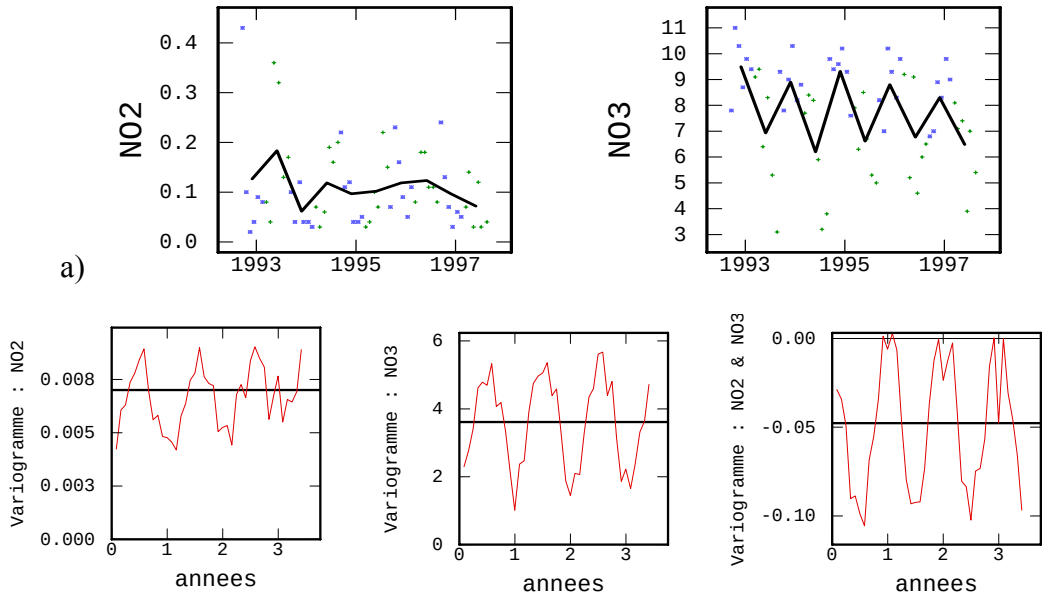


Figure 1. Concentrations en nitrites NO_2 et en nitrates NO_3 à la station 6027000 sur le Doubs. a) * mesures de septembre à février, + mesures de mars à août. La courbe correspond aux moyennes semestrielles. b) variogrammes temporels simples et croisés. Le variogramme croisé, négatif dans ce cas, indique l'anticorrélation temporelle des concentrations.

Pour l'estimation de la moyenne ou d'un quantile "annuel", les hypothèses d'indépendance et d'homogénéité ne sont cependant pas nécessaires. En effet, si $z(t)$ désigne la concentration en une station au cours du temps, la moyenne annuelle durant l'année $[t_0, t_0+T[$ est l'intégrale temporelle $z_T = \frac{1}{T}\int_{t_0}^{t_0+T} z(t)dt$. La version probabiliste de cette grandeur est l'intégrale stochastique $Z_T = \frac{1}{T}\int_{t_0}^{t_0+T} Z(t)dt$, qui reste définie même lorsque l'espérance de la Fonction Aléatoire Z n'est pas stationnaire. Lorsque Z admet une espérance m stationnaire, la relation entre ce paramètre et la moyenne annuelle recherchée est $E(Z_T) = m$: les concentrations "instantanées", et les valeurs annuelles fluctuent autour de leur espérance m, l'amplitude des fluctuations étant beaucoup plus réduites pour les valeurs annuelles que pour les valeurs "instantanées".

Appliqué sans précaution, le calcul statistique "standard" de la moyenne conduit ainsi à se tromper sur la grandeur à estimer.

A réalisation fixée, la moyenne annuelle peut être interprétée comme l'espérance de la variable aléatoire $z(U)$, U désignant une variable uniforme sur l'intervalle $[t_0, t_0 + T[$. Le quantile recherché est celui de cette variable $z(U)$.

Si l'on déconditionne par rapport à la réalisation, $Z(U)$ est une variable d'espérance Z_T et de variance $S^2 = \frac{1}{T}\int_{t_0}^{t_0+T} \left(Z(u) - Z_T\right)^2 du$. En modèle stationnaire ou intrinsèque, l'espérance de S^2 est égale à la variance de dispersion des concentrations ponctuelles durant l'année, qui s'exprime à l'aide du variogramme (Matheron,1965).

Les méthodes présentées ici sur des exemples ont été validées par comparaison des résultats aux valeurs théoriques, ou par simulations (Bernard-Michel et al., 2005).

KRIGEAGE DE LA MOYENNE ANNUELLE PAR STATION

En certaines stations, pour surveiller les fortes concentrations en nitrates, les mesures sont plus fréquentes en « hiver » (deux mesures par mois) qu'en « été » (une mesure par mois). L'échantillonnage étant préférentiel, la moyenne arithmétique fournit alors une estimation biaisée de la moyenne annuelle, facilement corrigeable par krigeage ou par une simple pondération par segment d'influence. Le signe du biais varie selon les substances : suivant les stations, la corrélation entre nitrates et nitrites est positive ou négative (comme à la station sur le Doubs, figure 1.).

La pondération par segment d'influence reste identique pour toutes les substances lorsque les mesures sont synchrones. La figure 2 montre un exemple de poids de krigeage pour l'estimation de la moyenne. Bien que les variogrammes des nitrites et des nitrates diffèrent fortement, les poids de krigeage présentent la même allure, sauf pour le premier et le dernier point. Pour les nitrates, les poids attribués aux mesures situées aux extrémités sont plus faibles, à cause de la prépondérance de la composante périodique. Au contraire dans le cas d'un "effet de pépite pur", c'est-à-dire en l'absence de corrélation temporelle, tous les poids seraient égaux et le krigeage de la moyenne coïnciderait alors avec l'estimateur statistique usuel.

Le krigeage fournit la variance d'estimation, pour la moyenne annuelle comme pour l'écart interannuel, tandis que la pondération par segment d'influence garantit la positivité des poids. Dans certaines configurations très irrégulières, certains poids de krigeage peuvent en effet devenir négatifs, ce qui peut être utilisé dans le cadre d'une automatisation des calculs pour détecter des anomalies de l'échantillonnage.

Dans les cas traités, la variance de krigeage de la moyenne annuelle est systématiquement inférieure à la variance statistique de l'espérance. En effet, la composante périodique rajoute de la variabilité et donc augmente la variance dans le calcul statistique standard, alors que cette périodicité est prise en charge par la composante en cosinus du variogramme dans le cas du krigeage (Bernard-Michel et al., 2005).

La figure 3 montre que les estimations varient sensiblement selon l'estimateur, moyenne "statistique" d'une part, krigeage ou segment d'influence d'autre part. Pour les nitrates la pondération corrige l'effet de l'échantillonnage préférentiel des fortes concentrations, bien visible sur l'estimateur classique à partir de 1990 (figure 3.b). Pour les nitrites, l'effet de la pondération varie suivant les années, la moyenne statistique étant le plus souvent supérieure (avant 1999) mais parfois inférieure (en 2001) aux estimations par pondération. Pour les nitrates, les poids de krigeage sont analogues à ceux par segment d'influence, et ces deux estimations sont analogues.

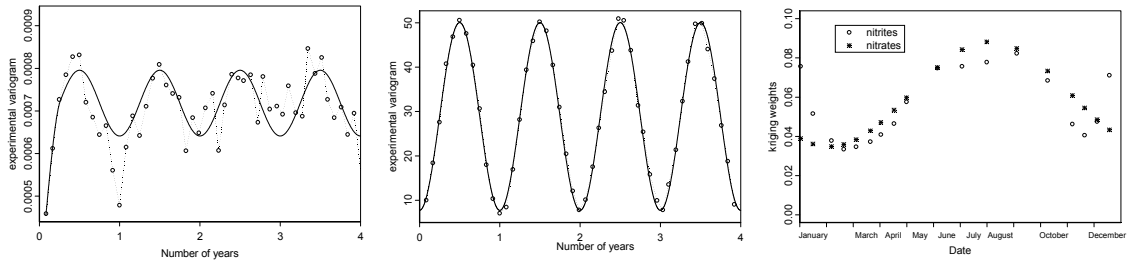
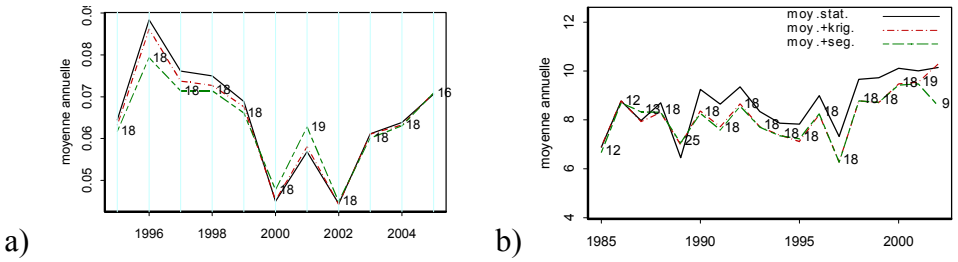


Figure 2. Variogramme temporel des nitrites et des nitrates pour la station 56000 sur la Loire, et poids de krigeage de la moyenne annuelle pour un échantillonnage préférentiel hivernal avec 18 mesures par an.



a) b)

Figure 3. Estimation de la moyenne annuelle à la station 26000 sur la Loire : moyenne statistique (trait continu), krigeage et pondération par segment d'influence. Le nombre de mesures annuelles est reporté. a) nitrites et b) nitrates.

Remarque : dans le modèle stationnaire ou intrinsèque strict, l'espérance des accroissements est supposée nulle sur le champ considéré, ici égal à une année Les hypothèses réellement nécessaires sont en fait bien plus faibles, car on vérifie que la variance de krigeage diffère peu du calcul approché en supposant l'indépendance des erreurs d'extension élémentaires d'une mesure à son "segment d'influence". L'hypothèse d'accroissements d'espérance nulle (ou d'espérance stationnaire des concentrations) ne porte plus alors que sur le plus grand segment d'influence, soit environ trois mois dans le cas très défavorable d'une mesure trimestrielle.

ESTIMATEUR EMPIRIQUE DU QUANTILE 90

La règle préconisée par le SEQ-eau pour le calcul du quantile 90 correspond à l'estimateur statistique usuel des quantiles par une fonction en escalier (Saporta,1990). Cet estimateur est connu pour être biaisé, le biais variant avec la taille de l'échantillon. Pour le quantile 90, la fonction en escalier introduit des discontinuités à chaque changement de dizaine de la taille de l'échantillon. Ceci rend délicate la comparaison entre stations n'ayant pas le même nombre de mesures, ou par station, lorsque ce nombre varie selon les années.

La littérature propose plusieurs autres méthodes d'estimation des quantiles, notamment la théorie des statistiques extrêmes, non adaptée dans ce cas. En effet, le quantile 90 est fort, mais non "extrême". Par ailleurs, les résultats asymptotiques supposent un grand nombre de données, alors que la fréquence d'échantillonnage varie généralement de 4 à 18 mesures annuelles par station.

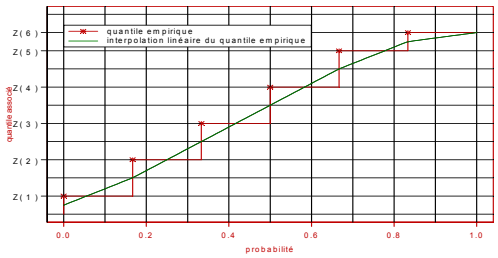


Figure 4. Fonction de quantile empirique et interpolation linéaire par morceaux

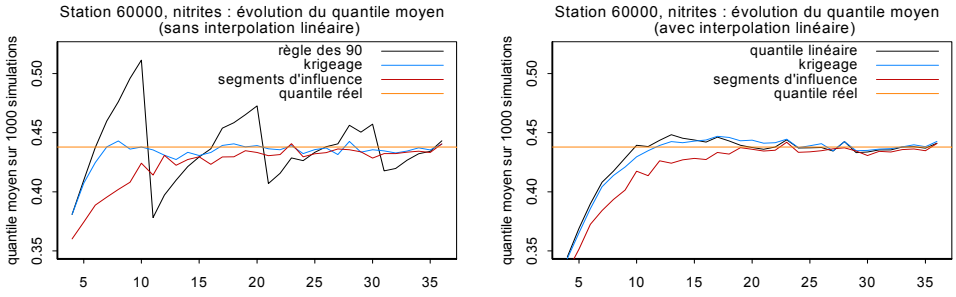
L'interpolation linéaire par morceaux de la fonction de quantile empirique (figure 4), associée

à la pondération des données par les poids de krigeage de la moyenne dans le cas d'un échantillon corrélé temporellement, fournit des résultats satisfaisants. En l'absence de corrélation, les quantiles ainsi calculés ont été comparés à la valeur théorique. En présence de corrélation, la comparaison a été effectuée sur simulations, pour des échantillonnages réguliers, irréguliers, puis préférentiels (Bernard-Michel et al.,).

La figure 5 présente les résultats obtenus *en moyenne* sur mille tirages d'une simulation conditionnelle des concentrations en nitrates en une station (pour se ramener à un cas "parfaitement connu" mais réaliste), pour un échantillonnage irrégulier. La règle actuelle correspond à la courbe la plus irrégulière (figure 5.a); l'interpolation linéaire de la fonction de quantiles atténue les discontinuités. La pondération réduit le biais en cas d'échantillonnage *préférentiel.* Avec moins de 20 mesures par an, les intervalles de confiance restent importants: même avec un échantillonnage régulier, 12 mesures sont insuffisantes pour une estimation précise du quantile 90, l'erreur relative restant très élevée (figure 5.b).

Lorsque les données sont peu nombreuses, l'estimation du quantile 90 dépend de la borne supérieure fixée pour la distribution, via le dernier segment de l'interpolation linéaire. Dans les ajustements présentés, cette borne coïncide avec le maximum de l'échantillon, ce qui provoque la sous-estimation observée sur la figure 5. D'autres choix sont possibles, comme par exemple utiliser les mesures de deux ou trois années consécutives, ainsi que le recommande le SEQ-eau lorsque les données sont trop peu nombreuses. Avec six mesures annuelles pendant deux ans, le problème de la "dernière" classe est ainsi éludé. Cet élargissement du "champ" atténue l'influence d'une année particulière, mais réciproquement, étend son influence à l'année suivante. Une autre solution consiste à mieux choisir la borne supérieure, par exemple en calant la variance de la distribution modélisée à la variance de dispersion des valeurs durant une année.

a)



b)

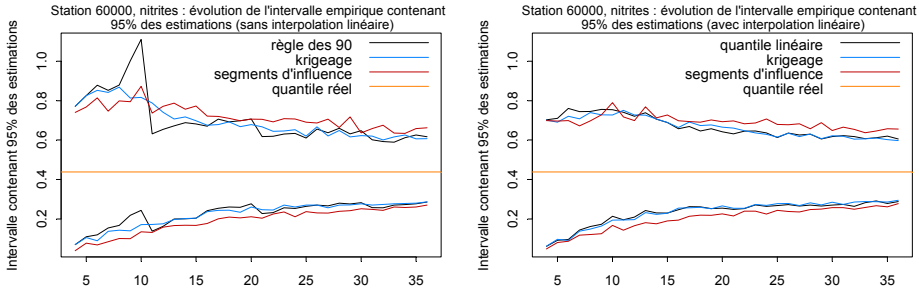
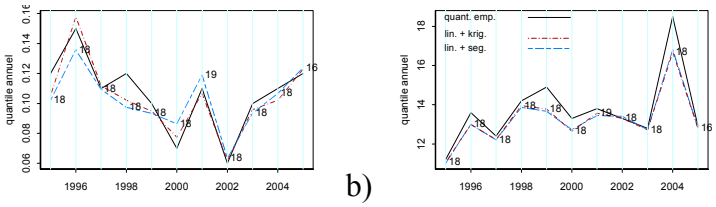


Figure 5. Estimation du quantile 90, après extraction d'un échantillon irrégulier sur simulation conditionnelle des concentrations en nitrites. A gauche, fonction de quantile empirique en escalier, et à droite, interpolation linéaire par morceaux decette fonction. a) moyenne des estimations, calculée sur 1000 simulations conditionnelles. b) intervalle de confianceà 95% empirique.

Comme pour la moyenne, la pondération réduit l'influence de l'échantillonnage préférentiel hivernal des fortes concentrations en nitrates (figure 6, même station qu'à la figure 3 pour l'estimation de la moyenne), alors que pour les nitrites, l'influence de cette pondération varie là encore selon les années. L'écart entre segment d'influence et krigeage reste plus marqué pour les nitrites. L'amplitude des écarts interannuels est généralement plus élevée pour l'estimateur classique du quantile que pour le krigeage. Enfin, les écarts interannuels sont parfois du même

ordre de grandeur que les écarts entre estimateurs pour une année donnée.



a) b)

Figure 6. Estimation du quantile 90, à la station 26000 sur la Loire : SEQ-eau (trait plein), et interpolation de la fonction de quantile empirique avec pondération par krigeage ou par segment d'influence. Le nombre de mesures annuelles est reporté. a) nitrites et b) nitrates.

CONCLUSION

Etablir des règles de calcul pour évaluer l'état de l'environnement (air, eau) nécessite de préciser les hypothèses sous-jacentes. En effet, de nombreux phénomènes naturels ou anthropisés présentent une certaine périodicité annuelle, ainsi qu'une corrélation spatiale ou temporelle. Or ces propriétés ne sont pas prises en compte dans les procédures statistiques classiques.

Par ailleurs, appliquer sans recul des procédures statistiques peut conduire à se tromper sur la grandeur à estimer. Ainsi que représente "l'espérance mathématique" usuellement inférée, lorsque l'hypothèse de stationnarité est mise en défaut ? La moyenne annuelle, clairement définie comme une intégrale temporelle, représente en fait l'indicateur recherché.

La procédure d'estimation a une influence pas du tout négligeable sur les indicateurs obtenus, selon que l'échantillonnage préférentiel est ou non corrigé par une pondération. Enfin, krigeage et segment d'influence fournissent des résultats parfois très voisins, mais pas systématiquement.

Enfin, la précision des estimations dépend de la variable à estimer, et varie fortement selon les hypothèses retenues. Un choix erroné des indicateurs et des modèles peut conduire à une évaluation erronée de la précision de ces indicateurs, voire à des préconisations d'échantillonnage inadaptées.

Remerciements

Ce travail a été effectué grâce à un financement du ministère en charge de l'environnement (Convention MEDD/ARMINES no. CV02000187). Les auteurs remercient Louis-Charles Oudin, préalablement à l'agence de l'eau Loire-Bretagne, pour le suivi actif de ces travaux.

Bibliographie

[1] (2003) Système d'évaluation de la qualité de l'eau des cours d'eau - SEQ-Eau. Rapport de présentation de la version 2

[2] Matheron G. (1965) Les variables régionalisées et leur estimation. Masson.

[3] Bernard-Michel C., de Fouquet C. (2005) Geostatistical indicators of waterway quality for nutrients. *In Geostatistics Banff 2004,* Leuangthong O. & Deutsch C. (Eds.). Springer.

[4] Bernard-Michel C., de Fouquet C. (2005). Estimating indicators of river quality by geostatistics. *In geostatistics for environmental applications - Geoenv 2004,* Renard P., Demougeot-Renard H. & Froidevaux R. (Eds). Springer.

[5] Saporta G. (1990) Probabilités, analyse des données, et statistique. Technip.

[6] Bernard-Michel C., de Fouquet C. Oudin L.-C. (2005). Calculs géostatistiques d'indicateurs des concentrations dans les cours d'eau *Rapport d'étude N-18/05/G*, Ecole des Mines de Paris, Fontainebleau.