

Indicateurs géostatistiques de la pollution des cours d'eau

Caroline Bernard-Michel

▶ **To cite this version:**

Caroline Bernard-Michel. Indicateurs géostatistiques de la pollution des cours d'eau. Sciences of the Universe [physics]. École Nationale Supérieure des Mines de Paris, 2006. English. NNT : . pastel-00001878

HAL Id: pastel-00001878
https://pastel.hal.science/pastel-00001878

Submitted on 6 Sep 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ECOLE DES MINES
DE PARIS

T H È S E

pour obtenir le grade de
Docteur de l'Ecole des Mines de Paris
Spécialité «Géostatistique»

par
Caroline BERNARD-MICHEL

Présentée et soutenue publiquement le 07 juillet 2006

INDICATEURS GEOSTATISTIQUES DE LA POLLUTION DANS LES COURS D'EAU

Directeur de thèse :Chantal de Fouquet

Jury

Mme	Chantal de FOUQUET	Directrice de thèse
M.	Denis MARCOTTE	Rapporteur
M.	Philippe RENARD	Rapporteur
M.	Ghislain de MARSILY	Président
M.	Michel MEYBECK	Examinateur
M.	Jean-Luc BADER	Examinateur

Table des matières

Remerciements 3

Résumé 7

Partie I Problématique 9

Chapitre 1 Indicateurs synthétiques de l'état des cours d'eau 11
1.1 Les indicateurs par station 11
1.1.1 Moyenne annuelle ou quantile? 11
1.1.2 Le SEQ-Eau 12
1.1.3 Vers des indicateurs géostatistiques 13
1.2 Les indicateurs le long des cours d'eau 14

Chapitre 2 Les données 17
2.1 Le réseau Français 17
2.2 Le référentiel : la BD Carthage 18
2.3 Le bassin Loire Bretagne 19
2.3.1 Présentation générale 19
2.3.2 Les paramètres étudiés 19
2.3.3 Les réseaux de mesure 20
2.3.4 L'échantillonnage temporel 22
2.3.5 La station « quai du roi » 26
2.4 Le bassin Rhin Meuse 26
2.4.1 Présentation générale 26
2.4.2 Le bassin de la Moselle 27
2.4.3 Les données 28
2.4.4 Surfaces drainées, occupation du sol 30
2.4.5 Implantation des données sous R 30
2.5 Chiffres significatifs 31

Partie II Indicateurs de qualité par station 33

Chapitre 3 Les indicateurs 35
3.1 Les calculs statistiques usuels 35
3.1.1 La moyenne annuelle 35
3.1.2 Le quantile 90 38
3.1.3 Problèmes liés à ces estimateurs 40
3.2 Quels indicateurs estimer ? 44

Chapitre 4 Estimation de la moyenne 49
4.1 Estimation géostatistique: le krigeage 49
4.2 Une simplification : les segments d'influence 51
4.3 Réflexions sur les pondérations 53
4.4 Modélisation de la série temporelle ? 57
4.5 Conclusion 59

Chapitre 5 Estimation du quantile 61
5.1 Bibliographie 61
5.2 Interpolation linéaire du quantile empirique 62
5.3 Tests de la méthode pour des variables aléatoires indépendantes 62
5.3.1 Loi uniforme : calcul théorique du biais 63

5.3.2 Evaluation du biais par simulation 64
5.3.3 Généralisation aux quantiles de tout ordre 68
5.3.4 Problème de la première et de la dernière classe 69
5.3.5 Estimateurs des quantiles par anamorphose 69
5.3.6 Probabilité de dépassement de seuil 73
5.4 Pondération des données pour des variables corrélées 73
5.5 Exemples 75
5.6 Conclusion 78

Chapitre 6 Validation des méthodes 79
6.1 Comment valider les méthodes ? 79
6.2 Simulations non conditionnelles 80
6.2.1 Les simulations 80
6.2.2 Echantillonnage, validation 82
6.2.3 Les résultats 83
6.3 Comparaison de différents échantillonnages sur des chroniques reconstituées 86
6.4 Etude par station 95
6.5 Conclusion 99

Chapitre 7 Application au réseau RNB Loire-Bretagne 101
7.1 Modélisation des variogrammes 101
7.2 Statistiques globales sur le réseau 102
7.2.1 Par année entre 2002 et 2004 103
7.2.2 Estimation sur trois années (2002-2004) 107
7.2.3 Evolution à long terme (1985-2005) 108
7.3 Analyse par région 110
7.4 Conclusion 111

Chapitre 8 Application au réseau RNB du bassin de la Moselle 113

Partie III Indicateurs de qualité entre stations 117

Chapitre 9 Position du problème et premières analyses expérimentales sur les données 119
9.1 Problèmes posés le long des réseaux 119
9.1.1 Choix de la distance 120
9.1.2 Stationnarité 125
9.1.3 Conditions de cohérence aux confluences 127

Chapitre 10 Modélisation des flux et des débits le long des cours d'eau 129
10.1 Bibliographie 129
10.1.1 Approche Agences de l'eau 130
10.1.2 Modélisation géostatistique. 131
10.2 Modèle général 134
10.3 Application aux débits spécifiques 136
10.3.1 Définitions 136
10.3.2 Modèle pour les débits spécifiques 137
10.3.3 Inférence du modèle 139

Conclusion 153

Bibliographie 157

Annexe A : Méthodes géostatistiques, variogramme et krigeage 165

Annexe B : Calculs des indicateurs sur trois années et entre 1985 et 2005 173

Annexe C : Publication Comptes Rendus Géosciences 179

Remerciements

Dans un premier temps, je remercie les membres du jury d'avoir accepté d'étudier avec attention mon travail. Merci beaucoup à Ghislain de Marsily dont j'ai pu constater la gentillesse lors de nos échanges par mail et qui m'a conseillée sur le choix de mon jury. Merci à mes rapporteurs Philippe Renard et Denis Marcotte, qui ont accepté la lourde tâche de corriger mes nombreuses erreurs et proposé des améliorations judicieuses dans mon rapport. Merci à Michel Meybeck que je n'ai pas eu l'occasion de connaître pendant ma thèse, c'est dommage…et enfin merci à Jean-Luc Bader de l'Agence de l'eau Loire Bretagne. Je tenais vraiment à ce qu'un « acteur » de l'eau soit présent à ma thèse et il n'a pas été facile de trouver un volontaire ! Donc un grand merci.

Je remercie Louis Charles Oudin de l'Agence de l'eau Loire Bretagne, qui est avec Chantal à l'origine de cette thèse. Ca a été un plaisir de travailler avec Louis Charles qui parle si bien de l'eau ! Il est toujours captivant. J'ai parfois même regretté qu'il ne vienne pas soutenir à ma place ! Je remercie aussi Danielle Maupas pour son aide tout au long de cette thèse.

Bien évidement, je remercie Chantal, qui a eu confiance en moi et m'a permis de travailler sur un sujet de thèse, à mon avis difficile car sans précédent, mais vraiment très intéressant. Je ferai bref en disant que j'ai été très contente de faire cette thèse avec Chantal, que je trouve intéressante, gentille, sensible et que tout simplement j'aime bien.

Merci à Karine qui est venue m'aider un mois pendant l'été. J'ai rarement vu quelqu'un d'aussi efficace ! Et puis très sympa en plus !

Merci à tout le centre de géostatistique. D'abord à Jean-Paul Chilès qui m'a accueillie dans son laboratoire et a tout fait pour que je suive cette thèse dans les conditions les plus favorables : prolongations de contrat, nombreux congrès, participation à la vie du centre. J'ai eu la chance de profiter de beaucoup d'expériences enrichissantes. Merci à Françoise, Nathalie et Isabelle, toujours prêtes à aider et vraiment adorables. Enfin merci à tous les chercheurs du centre de géostatistique avec qui j'ai été contente de travailler, discuter, partir en congrès…Un clin d'œil à Michel Poulain du CIG pour avoir assisté à toutes mes présentations et m'avoir conseillée de manière toujours très positive.

Merci à tous les acteurs de l'eau, de l'IFEN (A. Blum, F. Bertrand) ou des agences Loire Bretagne (J. Durocher, N. Baratin, A. Maupas, B. Bara, O. Coulon) et Rhin Meuse (G. Demortier, C. Conan, N. Siefert). La plupart ne m'ont jamais rencontrée et pourtant ils m'ont beaucoup aidée à acquérir des données, ce qui parfois leur demandait du travail. Je pense en particulier aux personnes de l'IFEN et de l'agence Rhin Meuse avec qui nous n'avions même pas de partenariat. Récupérer des données n'est pas une tâche facile mais ça fait vraiment plaisir de voir que la majorité des personnes que j'ai contactées m'ont répondu avec gentillesse et aidé dans mes recherches.

Je remercie tous les thésards de Fontainebleau et j'en ai vu passer ! Les anciens comme Alex et Nico avec qui on a bien profité des 3h de transport par jour pour discuter, mes compagnons de bureau, il y en a eu…ca tourne dans ce bureau ! Christophe (le doyen du bureau), Marco (je te pardonne pour la fenêtre et la lumière), Franck, le bioman isolé et Pierre (de passage dans le bureau ☹, on a bien rigolé, hein ?). Il y a aussi les autres, Mathieu, Marie, Martial…les cfsg, les thésards du cmm , du cig, de géophysique (je ne t'oublie pas François !), du cours d'espagnol…tout plein de monde vraiment sympa que je remercie sans les citer un à un !

Je remercie Michèle Desenfant du LNE, qui m'a donné envie de faire cette thèse. Merci Michèle d'avoir facilité mon départ du LNE. C'est vraiment une preuve d'intelligence d'avoir réagi ainsi et j'étais vraiment contente que tu viennes à ma soutenance.

Enfin et c’est le plus important pour moi, merci à ma famille. Merci à mes parents qui m’ont toujours laissée faire mes choix, mauvais ou bons, qui ne m’ont jamais poussée dans mes études (c’est sûrement pour ça que j’ai mis tellement longtemps), mais surtout merci pour leur soutien et leur amour. Merci à mes grands frères Gilles et Bruno dont je suis si fière, Gilles pour me remonter les bretelles quand je me plains un peu trop (maintenant ca va, hein ?), Bruno, pour me remonter le moral quand je perds confiance. Merci pour tout, j’ai toujours su qu’être votre petite sœur est la meilleure position dans la famille.

Pour terminer, je dédie cette thèse à mon grand père paternel qui serait tellement fier de voir mon nom à coté du sien dans l’annuaire de l’école de Mines, lui qui désespérait d’être le seul scientifique de la famille.

A grand-papa

Résumé

La qualité chimique des cours d'eau est mesurée par un réseau de stations constitué progressivement depuis une quarantaine d'années, qui fournit une base de données très riche, mais très hétérogène.

Le système d'évaluation de la qualité de l'eau préconise actuellement de synthétiser les mesures par station par des indicateurs statistiques tels que la moyenne arithmétique et le quantile 90. Ces calculs reposent sur deux hypothèses implicites mais erronées : l'indépendance des mesures et la stationnarité des concentrations durant l'année. En effet, les concentrations en nutriments (nitrates, nitrites et orthophosphates) présentent généralement des variations saisonnières et les variogrammes expérimentaux mettent en évidence une corrélation temporelle. Dans une première partie, nous examinons les biais et les incertitudes des indicateurs actuels. Le krigeage prend en compte l'irrégularité de l'échantillonnage et les corrélations temporelles dans l'estimation et dans le calcul d'incertitude associé. Une simplification par segments d'influence est proposée. Le biais du quantile empirique est réduit par une simple interpolation linéaire. L'apport de ces méthodes est étudié théoriquement et par simulation sur le bassin Loire Bretagne qui présente des stratégies d'échantillonnage très différentes selon la station et l'année.

Pour interpoler ces indicateurs le long des cours d'eau, la question se pose ensuite de modéliser leur corrélation spatiale. Or les modèles usuels de covariance, développés pour des espaces euclidiens, ne sont plus nécessairement valables sur une structure arborescente. Un modèle général de fonctions aléatoires le long d'un réseau hydrographique a donc été développé. En tout point du réseau, les cours d'eau sont considérés comme la combinaison de filets « élémentaires » définis par les chemins de l'ensemble des sources à l'exutoire. Les questions d'indépendance des fonctions aléatoires entre filets et de leur stationnarité sont discutées et l'inférence de ce modèle est examinée sur le bassin de la Moselle.

Abstract

In order to assess river quality, different parameters such as nutrients concentrations are measured in different monitoring stations, setting up a very important but heterogeneous database.

The French evaluation system of water quality recommends to summarize the information contained in these measurements by a few statistical indicators such as the annual mean of concentrations or the 90% quantile. They are estimated using the classical statistical inference based on hypothesis proved to be incorrect: time correlations and seasonal variations are ignored. Actually, in France, nitrate concentrations are generally higher in winter. Biases and confidence intervals can be reduced by kriging or segments of influence and a linear interpolation of the empirical quantile is proposed. Methods are analyzed theoretically and experimentally on the Loire Bretagne basin.

Estimating indicators along a stream network then requires specific models of random functions because usual covariance models are no longer valid on such structures. We propose a global model of random functions along a tree graph introducing the concept of "elementary thin streams", defined by the whole set of paths between sources and outlet. At each point of the network, the river is considered to be the linear combination of these streams on which one dimensional stationary random functions are defined. An application to water discharge on the Moselle Basin (north-east of France) is presented

Partie I
Problématique

Afin d'évaluer le niveau des concentrations en différentes substances polluantes, des indicateurs statistiques (moyenne annuelle, quantile 90) sont préconisés par les agences de l'eau dans le cadre du SEQ-Eau (système d'évaluation de la qualité de l'eau) et de la Directive Cadre Européenne. Le calcul de ces indicateurs repose sur des hypothèses d'indépendances temporelles et spatiales des mesures qui s'avèrent souvent inexactes, avec comme conséquences des biais dans les estimations et dans l'évaluation de la précision.

Des estimateurs géostatistiques sont proposés pour prendre en compte les corrélations temporelles (par station) ou spatiales (entre stations). Le calcul du quantile 90 nécessite un retour sur la modélisation de l'histogramme.

L'échantillonnage étant très variable selon les stations, le bassin Loire Bretagne est retenu comme cadre pour l'étude de la corrélation temporelle. La corrélation spatiale des débits et flux de concentration porte sur le bassin Rhin Meuse, et plus particulièrement de la Moselle, pour lequel les débits sont disponibles et dont l'échantillonnage présente un caractère plus systématique.

Chapitre 1

Indicateurs synthétiques de l'état des cours d'eau

Lutter contre la pollution des cours d'eau nécessite de pouvoir évaluer la qualité de l'eau et son évolution, et plus précisément les flux ou les concentrations en différentes substances (nutriments, pesticides…).

Pour ce faire, de nombreux paramètres physico-chimiques ou biologiques sont mesurés, et des outils de synthèse statistique ont été mis en place pour exploiter les résultats. Les concentrations mesurées sont converties en un indice de qualité ou « indicateur », permettant de juger de la qualité de l'eau pour une substance. Cet indice, calculé par station, peut être la moyenne annuelle des concentrations ou le quantile 90. La question se pose ensuite d'une interpolation entre stations pour fournir une image réaliste de l'ensemble du réseau.

Le détail des outils statistiques actuels est décrit dans divers rapports des agences de l'eau. Il n'existe pas encore de document normalisé pour l'ensemble de la France (Agence de l'eau et du Ministère de l'Environnement, 1994 ; Agence de l'eau Loire Bretagne, 2002 ; Wallin et al., 2002 ; Littlejohn et al, 2002).

Ce manuscrit étant destiné à deux publics parfois distincts : les géostatisticiens et les spécialistes de la qualité de l'eau, nous avons jugé nécessaire de détailler à la fois le contexte de la qualité de l'eau et les méthodes géostatistiques, dans l'objectif de faciliter la lecture de chacun.

1.1 Les indicateurs par station

1.1.1 Moyenne annuelle ou quantile?

Certaines substances (les graisses notamment) deviennent dangereuses lorsque la concentration accumulée dans l'organisme atteint un certain seuil. Ce qui détermine le risque pour la santé est alors la dose absorbée sur une certaine durée, dont les moyennes annuelles peuvent rendre compte. C'est le cas de certains micropolluants. De même, les flux de concentrations en polluant sont généralement synthétisés par des moyennes annuelles (Moatar and Meybeck, 2005 ; Meybeck et al., 2003) .

Dans le SEQ-Eau, le problème général de la pollution des eaux est plutôt abordé du point de vue des risques toxiques ou écotoxiques aigus (L.-C. Oudin, communication personnelle). Pour ces polluants à effet immédiat, le danger vient des concentrations pouvant entrainer la mortalité d'espèces vivantes. Pour caractériser l'occurrence des risques élevés, les spécialistes de l'eau retiennent comme indice de qualité la concentration maximale, susceptible de provoquer un effet toxique aigu. Mais comme cette « pointe » peut être exceptionnelle, voire être entachée d'une erreur de mesure, il semble plus judicieux de retenir comme indicateur le quantile 90, c'est-à-dire la concentration « instantanée ») dépassée dans 10% des cas, soit 10% du temps en un point fixé.

Pour les nutriments examinés ici (nitrates, nitrites, phosphates…), l'indicateur retenu est le quantile 90.

D'autres indicateurs plus complexes pourraient bien évidement être étudiés (Pereira et al., 2000)

1.1.2 Le SEQ-Eau

Utilisé au niveau national, le SEQ EAU a été instauré suite à la loi sur l'eau de 1992, dans le contexte de la directive cadre Européenne (2000) (Simonet, 2001 ; Agence de l'eau Loire Bretagne, 2002). Il vise à fournir un diagnostic précis sur l'aptitude de l'eau à la vie dans les cours d'eau et à différents usages. A partir des mesures physico-chimiques et bactériologiques, des indices sont construits par famille de paramètres, permettant un classement sur une échelle de qualité et un suivi de l'évolution interannuelle.
Par substance ou « paramètre », on retient comme prélèvement le plus déclassant dans un ensemble de mesures annuelles ou pluriannuelles, la valeur dépassée dans au moins 10% des prélèvements; c'est la règle dite « des 90% » qui correspond à l'estimation du quantile 90 pour cette période. Cet indicateur est ensuite converti en indice de qualité variant entre 0 et 100.
Une altération est définie par un groupe de paramètres. Pour chaque altération, la qualité de l'eau, déterminée par l'indice de qualité le plus défavorable, est codée par une couleur, allant du bleu pour le meilleur au rouge pour le pire (FIG.1-1).

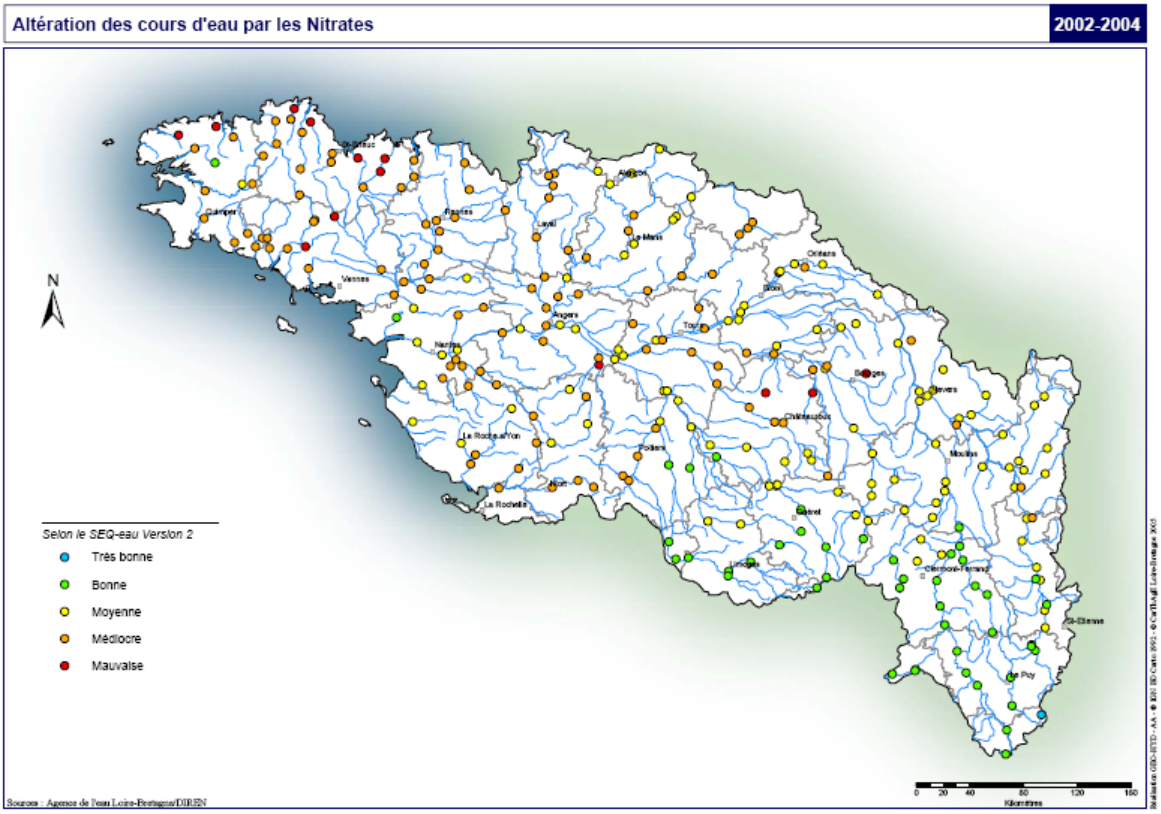


FIG.1-1 : Exemple de carte ponctuelle des qualités SEQ-Eau calculée sur la période 2002/2004 pour l'altération des nitrates. *Source : Agence de l'eau Loire Bretagne.*

1.1.3 Vers des indicateurs géostatistiques

Ces indicateurs statistiques, moyenne annuelle et quantile 90 des concentrations, sont actuellement calculés par les statistiques classiques, qui reposent sur deux hypothèses fondamentales mais implicites:

- Les valeurs mesurées aux différentes dates sont des tirages de variables aléatoires de même loi de probabilité.
- Ces tirages sont indépendants les uns des autres.

Or pour certaines substances, la concentration présente des variations systématiques au cours de l'année. A cause du cycle annuel de la végétation et du ruissellement, les concentrations en nitrates sont généralement plus élevées en hiver (FIG. 1-2, à gauche), l'amplitude des variations saisonnières pouvant représenter la majeure partie de la variabilité temporelle des concentrations.

Inversement, il arrive en certaines stations que les nitrites (FIG. 1-2, à droite) ou les orthophosphates soient systématiquement plus faibles en hiver. L'hypothèse d'une même distribution statistique des concentrations par station au cours du temps se révèle alors inadaptée.

D'autre part, une étude antérieure (de Fouquet and Bez, 2001), confirmée par les variogrammes expérimentaux (exemple FIG. 4-1) montre que l'hypothèse d'indépendance des concentrations par station au cours du temps, qui revient à considérer les valeurs mesurées comme des tirages indépendants les uns des autres, est erronée. Or en présence de corrélation temporelle, ou de variations saisonnières systématiques, un échantillonnage préférentiel induit des biais dans l'estimation de la moyenne annuelle et des quantiles.

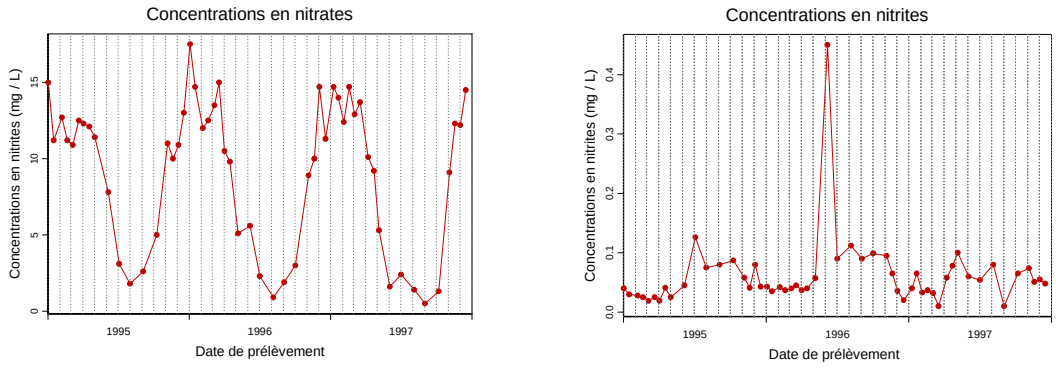


FIG. 1-2 : Concentrations en nitrates et en nitrites sur la Loire à Villandry. Station 56000 de 1995 à 1997. Les variogrammes temporels sont présentés en FIG. 4-1.

Enfin, le codage de l'aptitude en classes de couleurs, ou la "règle des 90%", sont des opérations pour lesquelles il est nécessaire de distinguer les valeurs réelles des indicateurs et leurs estimations disponibles à partir des mesures. Assimiler un indicateur à son estimation peut entraîner des biais importants, en particulier pour la comparaison à une valeur de référence ; ces biais sont d'autant plus marqués que l'estimation est peu précise. Quel peut alors être l'apport de la géostatistique, élaborée pour résoudre la question de l'estimation (linéaire et non linéaire) dans le contexte des variables régionalisées, c'est-à-dire présentant une corrélation spatiale ou temporelle (Matheron, 1965 ; 1969 ; 1970 ; 1973) ?

Classiquement, on propose d'estimer la moyenne annuelle par krigeage. Ceci nécessite de redéfinir pertinemment l'indicateur « moyenne annuelle ». Sur les réalisations, la moyenne annuelle est une intégrale temporelle. Dans le modèle, c'est la version probabiliste de cette grandeur qui sera estimée et non plus l'espérance mathématique de la variable aléatoire dont les tirages sont donnés par les mesures.

En vue d'une application systématique, une simplification de l'implémentation a été demandée. Le krigeage a donc été comparé à son approximation par segments d'influence, facilement automatisable.

Le calcul du quantile 90 ou plus généralement d'un quantile nécessite de préciser la variable aléatoire dont on modélise la distribution. Plusieurs méthodes de calcul de la fonction de quantile ont d'abord été comparées. La corrélation temporelle des mesures est ensuite classiquement prise en compte par une pondération des données. Par cohérence avec l'estimation de la moyenne, les poids de krigeage de la moyenne annuelle sont retenus. L'effet de l'approximation par segments d'influence est ensuite examiné.

L'importance des corrections proposées est étudiée expérimentalement sur différentes parties du bassin Loire Bretagne.

Remarque : la « moyenne » annuelle peut être définie de deux façons différentes, selon que l'on tient compte ou non des débits :

- Le premier point de vue, retenu pour l'indicateur statistique, peut être décrit comme celui du « buveur d'eau », qui prélèverait la même quantité tout au long de l'année. Si $z(t)$ désigne la concentration instantanée et T l'année commençant à l'instant t_0, la moyenne annuelle s'écrit $\frac{1}{T}\int_{t_0}^{t_0+T} z(t)\,dt$.
- Le deuxième point de vue revient à considérer une retenue remplie par le cours d'eau durant une année. Si $d(t)$ désigne le débit instantané du cours d'eau, la moyenne dans la retenue au bout d'une année est alors $\frac{1}{D_T}\int_{t_0}^{t_0+T} d(t)z(t)\,dt$, D_T désignant le débit annuel total $D_T = \int_{t_0}^{t_0+T} d(t)\,dt$. C'est l'approche par « flux », non considérée dans cette première partie, mais dont il faut tenir compte pour modéliser les concentrations le long d'un réseau hydrographique à cause des mélanges aux confluences.

1.2 Les indicateurs le long des cours d'eau

Une fois les indicateurs estimés par station, se pose la question de leur interpolation le long des cours d'eau, afin d'obtenir une estimation de la qualité en dehors des points expérimentaux. Ceci nécessite de modéliser la corrélation spatiale des concentrations. Les fonctions aléatoires ne sont alors plus définies sur des espaces euclidiens mais sur des structures arborescentes, et dans ce cas, les modèles de covariance usuels ne sont plus nécessairement valides. Il est donc nécessaire de développer de nouveaux modèles.

Un modèle général de covariance sur un réseau hydrographique a été proposé dans lequel la fonction aléatoire en tout point est considérée comme la combinaison linéaire de fonctions aléatoires définies sur des filets élémentaires.

Une application est effectuée sur le bassin de la Moselle pour laquelle les données sont acquises de façon plus systématique et qui présente un réseau plus dense que le bassin Loire Bretagne.

L'étude spatiale des indicateurs se limitera à la modélisation de la covariance des débits spécifiques et des flux annuels moyens. En particulier, on cherchera à contrôler les hypothèses du modèle sur les données. La question de l'inférence du modèle sera discutée, le nombre de stations étant toujours très faible quel que soit le bassin considéré.

Une fois la covariance modélisée, l'interpolation au fil de l'eau peut s'effectuer classiquement par les méthodes géostatistiques usuelles tel que le krigeage. Cependant, il n'est pas réaliste de fournir des cartes linéaires de débits ou flux sans faire appel à des modèles hydrologiques et biochimiques évolués tels que PEGASE, SENEQUE ou DECLIC qui prennent en compte un nombre considérable de relations avec des variables exogènes.

Chapitre 2

Les données

L'application pratique des méthodes géostatistiques a pour but d'évaluer les améliorations proposées, mais aussi d'examiner la faisabilité de leur mise en œuvre systématique. Ces aspects pratiques nous amènent à revenir sur l'organisation des réseaux de collecte des données : découpage administratif du territoire en agences de bassins, référentiel hydrographique. Deux cas particuliers sont détaillés, le « bassin » Loire-Bretagne et celui de la Moselle qui dépend de l'agence Rhin Meuse. Les caractéristiques de l'échantillonnage des nutriments sont précisées pour chacun.

2.1 Le réseau Français

Le réseau hydrographique français se divise en 6 grandes zones géographiques (FIG. 2-1) qui regroupent les différents bassins versants. Toutes les gouttes de pluie qui tombent dans un bassin versant se rejoignent pour former une rivière qui débouche sur un fleuve ou dans la mer.

Ces bassins versants sont gérés depuis 1964 par les agences de l'eau dont la mission principale est d'aider financièrement et techniquement les opérations d'intérêt général au service de l'eau et de l'environnement du bassin. Un des objectifs principaux des agences est d'évaluer la qualité des cours d'eau par rapport à différentes pollutions et son évolution dans le temps.



FIG. 2-1: Découpage du réseau hydrographique français en 6 agences.

Source : Agence de l'eau Seine Normandie

2.2 Le référentiel : la BD Carthage

Avec plus de 500 000 kilomètres de cours d'eau de plus de 1km, soit une densité moyenne de 1km de cours d'eau par km^2, la France possède un réseau hydrographique important. Un système de repérage spatial des milieux aquatiques, la BD Carthage, a été créé en 1994 par les agences de l'eau et le ministère chargé de l'environnement, en partenariat avec l'institut géographique national : (SANDRE, 2002). Les cours d'eau de la BD Carthage sont représentés par une succession de tronçons linéaires renseignés par différents attributs telles que leur largeur, leur nom, les abscisses curvilignes des nœuds de début et de fin de tronçons, leurs points kilométriques etc. Ils sont affectés d'un code hydrographique résultant du découpage de la France en bassins versants (circulaire de février 1991). Chaque bassin hydrographique géré par les agences est découpé en quatre partitions par aires hydrographiques décroissantes :

- la région hydrographique (ordre 1), qui correspond aux groupements de bassins versants
- le secteur hydrographique (ordre 2)
- le sous secteur hydrographique (ordre 3)
- la zone hydrographique (ordre 4)

Au final, l'ensemble du territoire français est donc découpé en zones hydrographiques, dont les limites s'appuient sur celles des bassins versants élémentaires. Ces zones sont codées en rapport aux zones d'ordres supérieurs qu'elles traversent. Un exemple de découpage en zones hydrographiques pour le bassin Loire Bretagne est présenté FIG. 2-2.

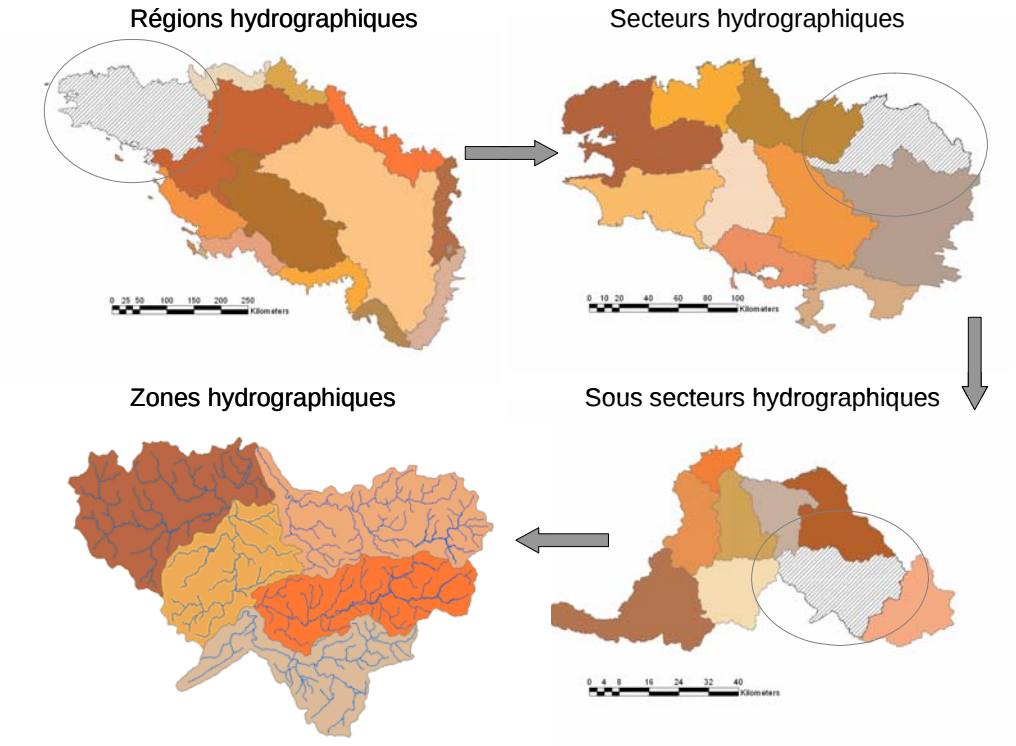


FIG. 2-2 : Découpage en zones hydrographiques du bassin Loire-Bretagne

2.3 Le bassin Loire Bretagne

2.3.1 Présentation générale

Le bassin Loire Bretagne, d'une superficie de 155 000 km^2 représente 28% du territoire national. Il est composé principalement des bassins versants de la Loire et de ses affluents, des bassins côtiers vendéens et bretons et de la Vilaine. Cet ensemble de bassins constitue un réseau hydrographique de 135 000 km avec des régimes hydrologiques très contrastés et des contraintes économiques et environnementales marquées : rural, il concentre les deux tiers de l'élevage français (majoritairement en Bretagne), 50% des productions céréalières nationales (Poitou-Charentes) et une activité industrielle orientée dans la production agro-alimentaire (à l'ouest et au centre du bassin). C'est aussi une région de pêche et de tourisme (FIG. 2-3)



FIG. 2-3 : Carte de l'occupation du sol sur le bassin géré par l'agence Loire Bretagne et drains principaux du réseau hydrographique.

2.3.2 Les paramètres étudiés

Les paramètres suivants sont examinés:

- ✓ Les débits aux stations de mesures.
- ✓ Les concentrations en nitrates qui forment à eux seuls une altération du SEQ-Eau. Depuis 20 ans, les concentrations en nitrates ont beaucoup augmenté. Elles sont particulièrement importantes à l'ouest du bassin, en Bretagne et dans la Vilaine. Elles sont dues à l'importance des cultures, aux développements des élevages et aux engrais utilisés. Depuis 1990, le rythme

d'augmentation moyenne des concentrations ralentit, en raison de l'évolution du coût des engrais et d'une sensibilisation des éleveurs. Cependant, si dans l'ensemble les teneurs en nitrates se stabilisent, elles restent très importantes sur certains points de prélèvement.

- ✓ Les nitrites qui font partie de l'altération des matières azotées, sont essentiellement liés aux rejets domestiques et industriels. Ces polluants ont significativement régressé depuis les années 80, mais restent importants dans certains secteurs.
- ✓ Les orthophosphates font partie de l'altération des matières phosphorées, principalement liée aux rejets ponctuels urbains, industriels et d'élevages. Leur étude met en évidence de nombreux foyers de pollution mais globalement une amélioration est constatée depuis les années 90.

Les unités des variables étudiées dans l'ensemble du document sont rappelées dans le tableau TAB. 2-1. Afin de ne pas alourdir les légendes des graphiques et tableaux, elles seront omises par la suite.

Nitrates	mg (NO3) / L (milligrammes NO3 par litre)
Nitrites	mg (NO2) / L
Orthophosphates	mg (PO4) / L
Débits instantanés	m^3 / s (mètre cube par seconde)
Flux de concentration	g / s
Distances	km (kilomètres)

TAB. 2-1 : unités des variables étudiées

Les unités des variogrammes, variances, ou des autres grandeurs découlent de ce tableau : par exemple un variogramme est exprimé en mg^2/L^2.

2.3.3 Les réseaux de mesure

L'ensemble des mesures chimiques de 1985 à 2005 pour ce bassin est aujourd'hui disponible sur le site de l'agence http://www.eau-loire-bretagne.fr/. Ces données ont été collectées sur des stations provenant de trois réseaux de mesure: le réseau national de bassin (RNB), les réseaux départementaux et les zones d'action renforcées (ZAR). Durant ces 21 années, les mesures pour les nitrates, nitrites et orthophosphates sont issues de plus de 1300 stations de mesures que nous désignerons par la suite sous le nom de réseau RNB complet (FIG. 2-4). Beaucoup de ces stations sont peu informées puisque 55% d'entre elles comportent moins de 12 mesures par an. D'autre part, certaines stations ont très rapidement été fermées. De manière générale, le réseau de mesures sur le bassin Loire Bretagne est bien informé spatialement mais l'échantillonnage est très hétérogène. Ainsi, non seulement pour une même station, la fréquence d'échantillonnage varie d'une année à l'autre, mais pour une même année, les stratégies d'échantillonnage diffèrent aussi sur l'ensemble du réseau. En fait, ce qui concerne la qualité de l'eau est en évolution rapide. Les premières mesures de polluants ont été effectuées dans les années 1970 en certaines stations étant amenées à disparaître et être remplacées. Peu à peu, l'implantation des stations et les stratégies d'échantillonnage évoluent. Cette évolution est nécessaire

et positive, mais elle rend difficile l'étude de l'évolution des concentrations, puisque les supports d'étude (stations, échantillons) diffèrent énormément.

Actuellement, le huitième programme d'intervention couvre la période de 2002 à 2006. Dans ce programme, un ensemble de stations RNB représentatives ont été sélectionnées sur l'ensemble de la France pour définir un réseau plus homogène sur lequel est appliqué le SEQ-Eau. Pour l'agence Loire Bretagne, 269 stations RNB de référence ont été choisies. Cet ensemble de stations est noté RNB8. Pour illustrer l'influence des méthodes, les stations du RNB8 seront étudiées pour deux périodes :

- entre 2002 et 2004 (RNB8-2002)
- entre 1985 et 2005 (RNB8-1985)

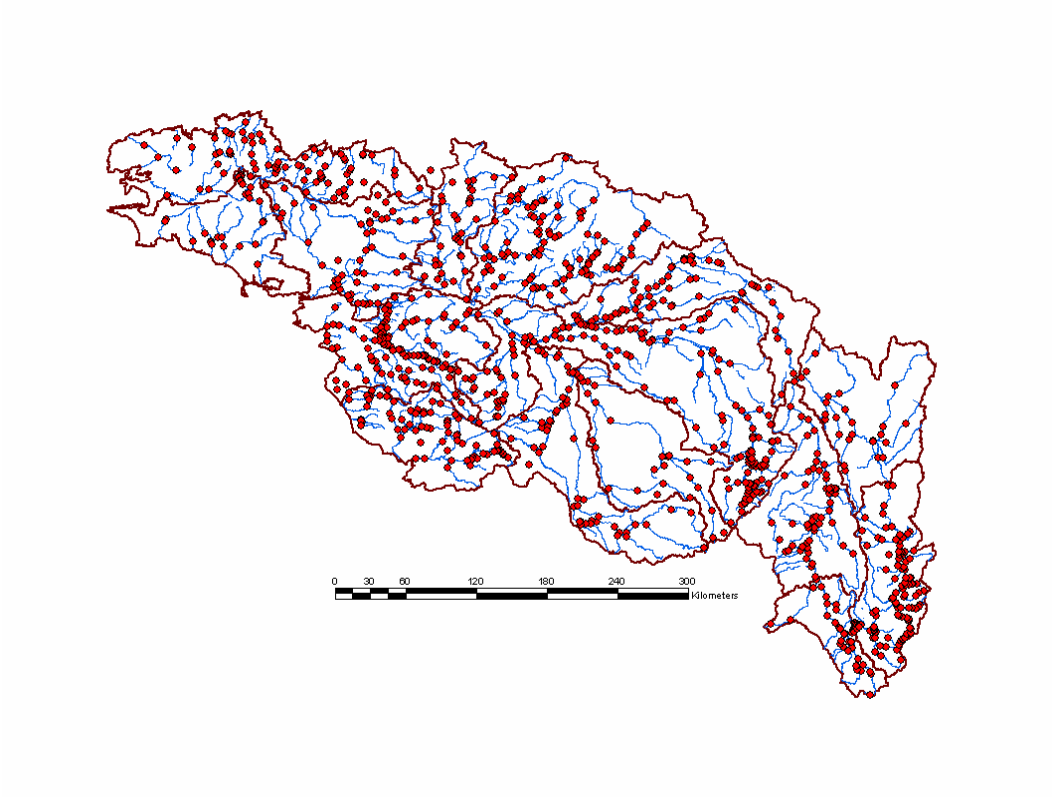


FIG. 2-4 : Ensemble des stations du réseau national de données (RNB complet) ayant fonctionné au moins une année entre 1985 et fin 2005.

Dans les études de l'agence Loire Bretagne, les résultats sont présentés soit pour la totalité du bassin, soit par région hydrographique. Ainsi 6 commissions géographiques sont étudiées (FIG. 2-5) :

- Vilaine et côtiers bretons (Bretagne nord, Bretagne sud, vilaine)
- Allier-Loire amont (Allier amont, Allier aval, Loire à l'amont de Villerest, Loire de Villerest au bec d'Allier)
- Loire aval et côtiers vendéens (Côtiers vendéens, Loire angevine, Loire aval)
- Vienne-Creuse (Vienne, Creuse)
- Mayenne-Sarthe-Loir (Mayenne, Loir et Sarthe)

- Loire moyenne et affluents (Cher et Indre, Loire moyenne en amont du bec de Vienne)

Ce découpage est conservé pour les applications.

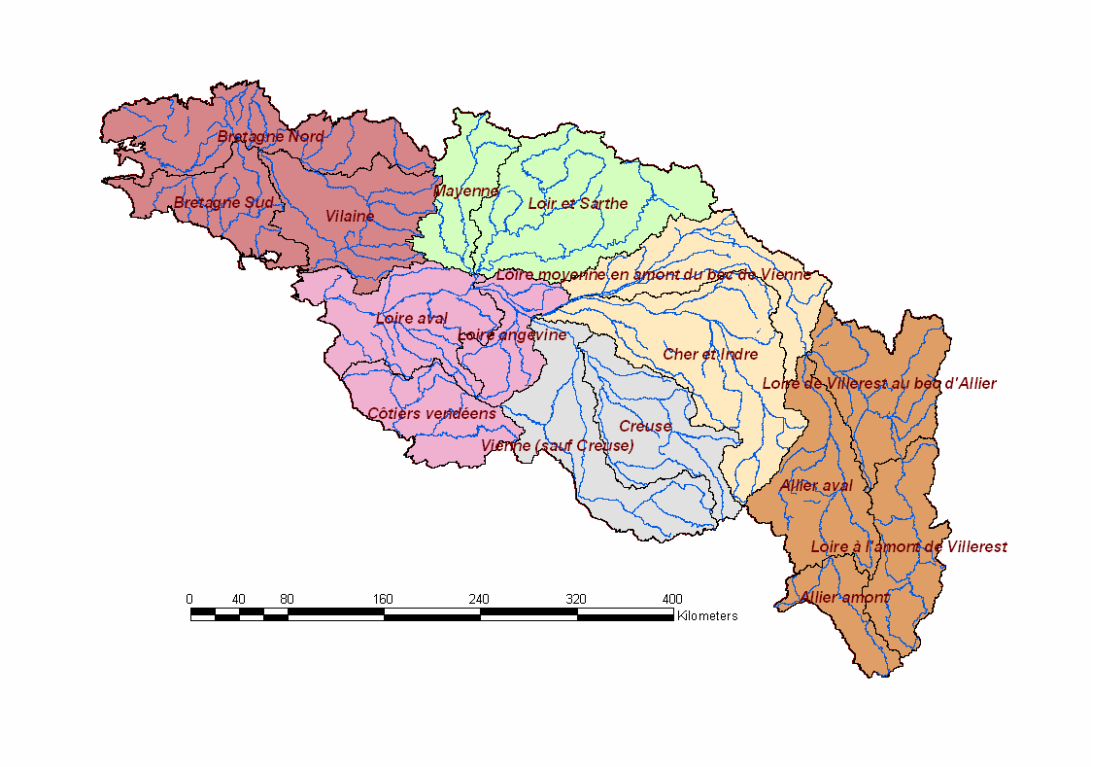


FIG. 2-5 : Carte des bassins versants composant le bassin Loire-Bretagne. Groupement des bassins en 6 commissions géographiques.

2.3.4 L'échantillonnage temporel

La fréquence d'échantillonnage varie considérablement avec les années, mais aussi suivant les stations de mesures. Dans le système d'évaluation de la qualité des cours d'eau, un nombre minimum de prélèvements annuels est requis pour la règle des 90%. Pour les nitrates, une mesure par trimestre au minimum est nécessaire. Pour les nitrites et les orthophosphates, quatre prélèvements répartis entre mars et octobre sont nécessaires. Cependant, lorsque l'utilisateur estime que les données sont représentatives de la situation critique annuelle, ces règles peuvent être adaptées. De manière générale, 3 types d'échantillonnage se présentent :

- Des échantillons avec moins de 12 mesures par an. Dans la majorité des cas, les mesures sont réparties irrégulièrement durant la période estivale, avec pour conséquence des biais dans les estimations (FIG. 2-7, A1, B1, B4) : les nitrates étant plus élevés en hiver, un échantillonnage concentré en été sous estime les indicateurs. Pour le réseau RNB8-1985, lorsque la fréquence d'échantillonnage est de 6 mesures par an, l'ensemble des mesures est en été (mai à octobre). Avec 4 mesures en été (juin à septembre) et deux en hiver (mars, décembre), les stations du réseau RNB8-2002 présentent un échantillonnage plus pertinent (FIG. 2-7, A1). De tels échantillonnages résultent de compromis destinés à réduire les dépenses. En général, le nombre de mesures a été réduit à 6 par an sur les affluents aux cours d'eau importants lorsque l'essentiel de la pollution est liée à des rejets ponctuels avec un effet de dilution en hautes eaux. Seules les mesures durant la

période critique des basses eaux sont alors conservées. En cas de fortes concentrations en nitrates, deux mesures ont été ajoutées en hautes eaux, soit 8 au total par an. Cette restriction budgétaire conduit à des estimations peu fiables, ce dont les agences sont conscientes. Elles tentent d'améliorer les indicateurs en calculant la qualité de l'eau sur 3 ans, passant ainsi à un calcul sur 18 mesures au lieu de 6.

- Des échantillons de 12 mesures par an, en général une mesure par mois (FIG. 2-7, A2, B2). Les mesures sont majoritairement régulières. En Loire Bretagne, elles représentent la majorité (65%) des échantillons du réseau RNB8-2002.
- Des échantillons comportant plus de 12 mesures par an. Ces échantillons sont majoritairement irréguliers avec une augmentation de la fréquence des mesures en hiver, ce qui crée des biais dans les estimations. Par exemple on trouve un certain nombre de stations avec 18 mesures par an (6 en été, 12 en hiver), ce qui a pour effet de surestimer les indicateurs pour les nitrates (FIG. 2-7, A3, B3). Ces stations ont été conçues pour la mesure des flux. Les flux (produit des débits et des concentrations) calculés en hautes eaux sont élevés et très imprécis du fait de la faiblesse des concentrations. Il a donc été décidé d'augmenter la fréquence des prélèvements en hautes eaux.

L'histogramme du nombre de mesures par an sur le réseau RNB8-2002 (FIG. 2-6) montre que dans la majorité des cas, 6, 12 ou 18 prélèvements sont effectués par an. Les 16 % restants sont des cas pour lesquels la fréquence d'échantillonnage diffère (exemples FIG. 2-7, B5 et B6), soit 31 fréquences répertoriées. Ces échantillons présentent pour la majorité des irrégularités d'échantillonnage non systématiques ou non caractérisables à cause du faible nombre de stations concernées.

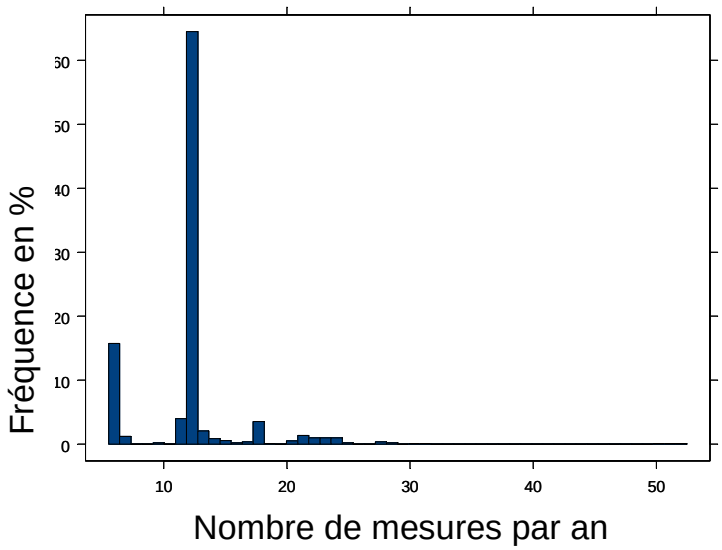
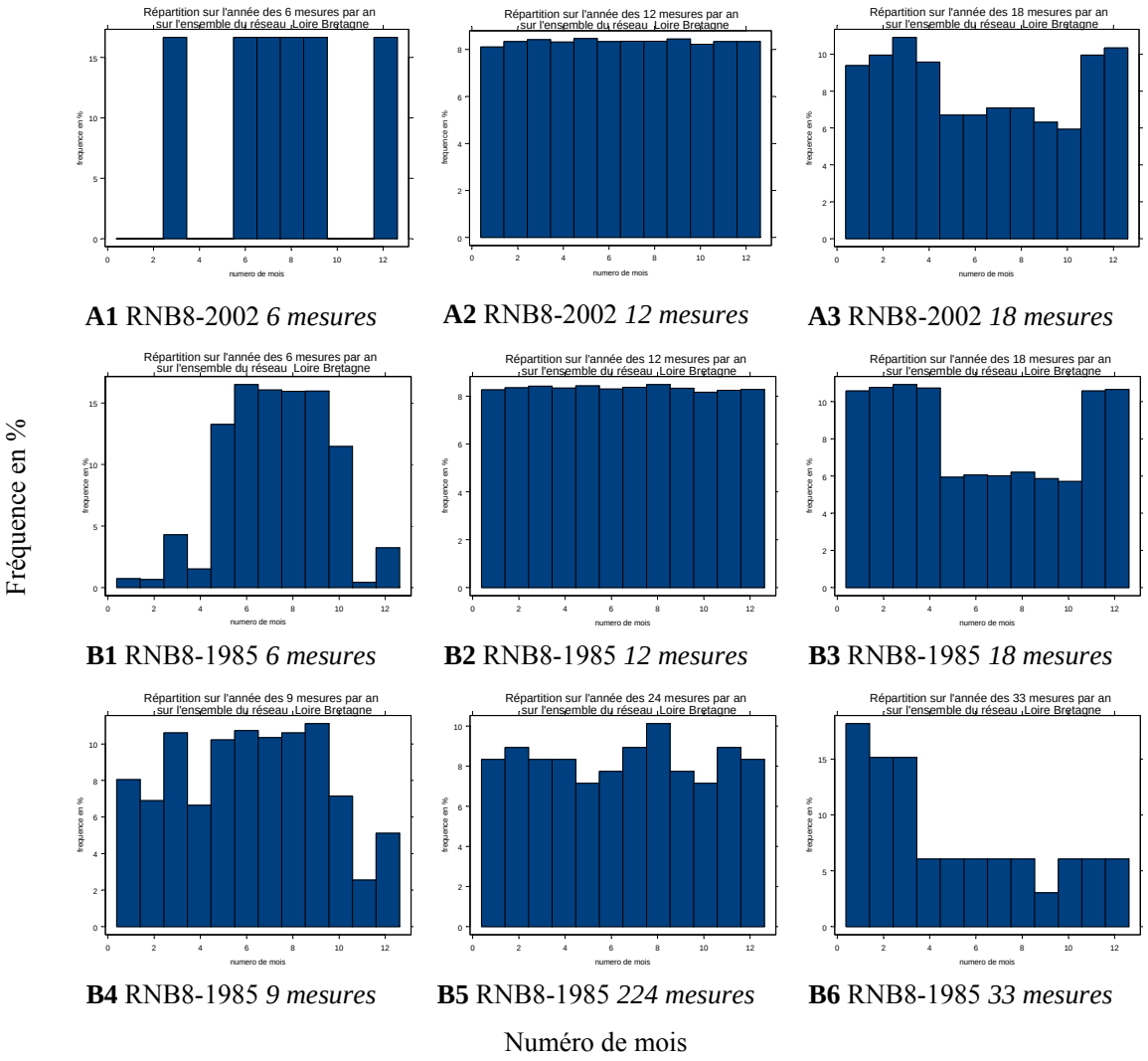


FIG. 2-6 : Histogramme de la fréquence d'échantillonnage (nombre de mesures par an) sur le réseau RNB8-2002. En abscisse, le nombre de mesures par an, en ordonnée, leur fréquence.



A1 RNB8-2002 *6 mesures* **A2** RNB8-2002 *12 mesures* **A3** RNB8-2002 *18 mesures*

B1 RNB8-1985 *6 mesures* **B2** RNB8-1985 *12 mesures* **B3** RNB8-1985 *18 mesures*

B4 RNB8-1985 *9 mesures* **B5** RNB8-1985 *224 mesures* **B6** RNB8-1985 *33 mesures*

FIG. 2-7 : Répartition mensuelle des prélèvements en fonction du nombre de mesures annuelles. Les figures A1, A2 et A3 portent sur le réseau RNB8-2002. Les figures B1, B2, B3, B4, B5 et B6 portent sur le réseau RNB8-1985. En abscisse, le numéro du mois, en ordonnée, sa fréquence.

Le tableau TAB. 2-2 et la carte FIG. 2-8 montrent l'hétérogénéité spatiale et temporelle. Pour les stations du RNB8, la stratégie d'échantillonnage varie selon la région (FIG. 2-8). En Vendée et en Bretagne du sud, c'est à dire au sud ouest du bassin, les échantillons sont en général plus nombreux qu'ailleurs et comportent plus de 12 mesures par an. A l'inverse, au nord ouest du bassin, en Bretagne du nord, Mayenne, Loir et Sarthe, moins de 12 mesures sont effectuées par an. Enfin, à l'est du bassin, les mesures sont généralement régulières. La fréquence des mesures a des conséquences importantes sur les estimations. Ainsi, à l'ouest du bassin, à dominance agricole (FIG. 2-3), la fréquence des mesures est augmentée alors que c'est précisément dans cette région que les concentrations sont élevées. Les indicateurs risquent donc d'être fortement surestimés.

Le tableau TAB. 2-2 montre que le nombre de mesures annuelles varie par région en fonction du temps. Par exemple, en Vienne et en Creuse, entre 1985 et 2005, 49% des stations présentent moins de

12 mesures par an. Entre 2002 et 2005, seulement 17% des stations restent sous-échantillonnées. Globalement, au cours du temps, le nombre de stations pour lesquelles moins de 12 mesures par an sont disponibles diminue passant approximativement de 37% à 21% pour les stations du RNB8 de 1985 à aujourd'hui. Inversement, le nombre de stations présentant plus de 12 mesures par an a augmenté.

RNB8-1985 RNB8-2002	**< 12**		**12**		**>12**	
Loire Bretagne (bassin entier)	***37%***	21%	***52%***	65%	***11%***	14%
Mayenne-Sarthe-Loir	***34%***	29%	***52%***	57%	***14%***	14%
Vilaine et côtiers bretons	***45%***	26%	***45%***	56%	***10%***	18%
Vienne-Creuse	***49%***	17%	***44%***	77%	***7%***	6%
Loire Moyenne et affluents	***41%***	14%	***45%***	67%	***14%***	19%
Allier-Loire-Amont	***43%***	18%	***50%***	76%	***7%***	6%
Loire aval et côtiers vendéens	***41%***	25%	***55%***	52%	***4%***	23%

TAB. 2-2: Pourcentage de stations en fonction du nombre de mesures annuelles pour chacune des commissions géographiques. Les statistiques sont effectuées pour l'ensemble des stations RNB8-1985(en gras) et pour les stations RNB8-2002. Les années non mesurées sont écartées des calculs.

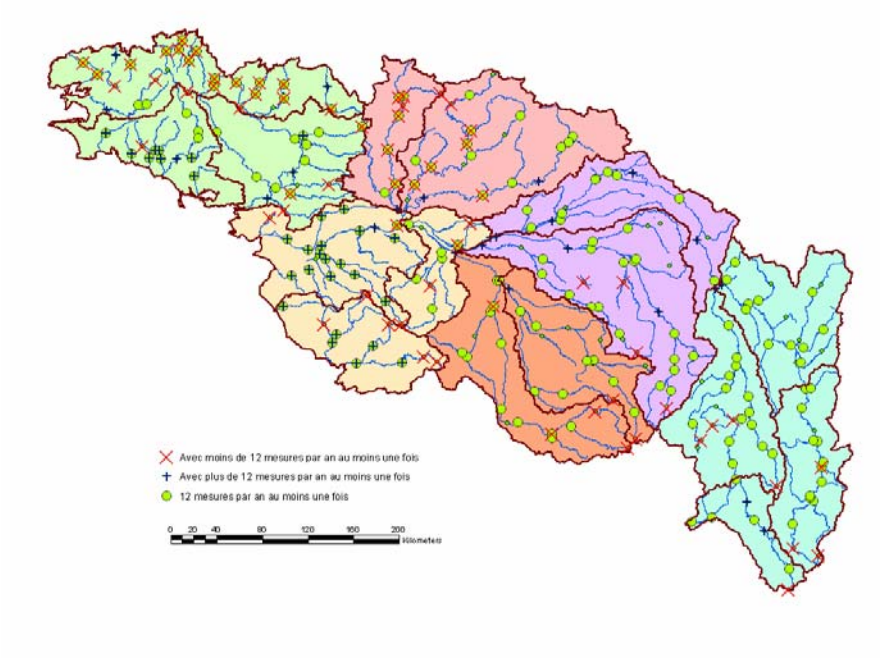


FIG. 2-8 : Répartition des fréquences d'échantillonnage sur l'ensemble des stations RNB8-2002. Chaque type d'échantillonnage est reporté sur la carte s'il a été effectué au moins une fois sur les 3 années.

2.3.5 La station « quai du roi »

La station « quai du roi » à Orléans, présente un intérêt particulier pour notre étude. C'est la station la mieux informée du bassin Loire Bretagne pour les nitrates avec 172 mesures en 1985. Elle présente des variations saisonnières marquées et des moyennes annuelles relativement homogènes au cours du temps (FIG. 2-9). Par la suite, de nombreux exemples et développements seront présentés pour cette station dans l'idée que les mesures étant quasi journalières, la moyenne annuelle et le quantile peuvent être supposés connus. En réalité, l'échantillonnage de cette station étant parfois préférentiel, nous la complèterons par simulation pour se ramener à un échantillon de 365 mesures. Le détail de cette pratique sera décrit dans le paragraphe (6.3)

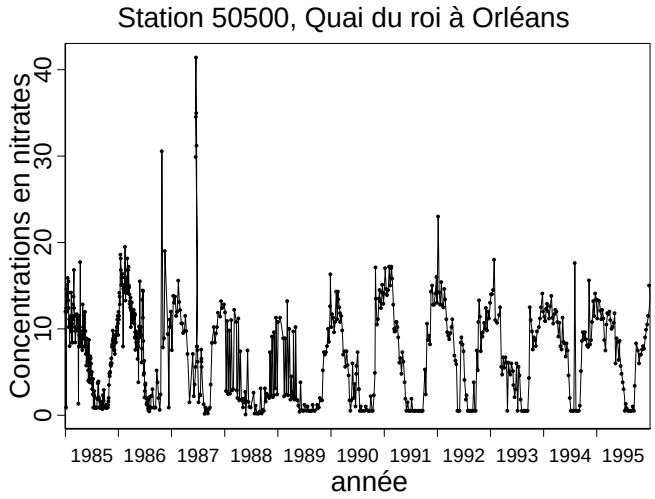


FIG. 2-9 : Chronique sur 11 ans de la concentration en nitrates. Station 50500, Quai du Roi à Orléans.

2.4 Le bassin Rhin Meuse

2.4.1 Présentation générale

Avec une superficie de 32 700 km^2, le bassin Rhin Meuse est composé de trois bassins versants drainés principalement par la Moselle, le Rhin et la Meuse (FIG. 2-10).

Certaines de ces rivières traversent plusieurs pays. Lorsque la source est en Allemagne ou en Suisse, des données manquent pour l'étude des corrélations spatiales. C'est le cas du Rhin qui prend sa source à l'est de la Suisse. La Meuse quant à elle, prend sa source en France au plateau de Langres. Elle traverse le sud de la Belgique et les Pays Bas où elle se mêle au Rhin. Mais certaines des sources du réseau hydrographique qu'elle engendre se situent en Belgique, ce qui comme pour le Rhin est embarrassant dans le cas d'une étude spatiale des débits ou concentrations. Nous avons retenu le bassin de la Moselle qui prend sa source dans les Vosges et se jette dans le Rhin en Allemagne. Son réseau hydrographique se compose d'une centaine de confluences importantes résultant de l'intersection des drains considérés par l'agence de l'eau Rhin Meuse comme principaux dans un objectif de qualité. L'occupation du sol sur l'ensemble de ces bassins est principalement urbaine et forestière (FIG. 2-10). La tendance globale est à l'amélioration pour le bassin Rhin Meuse, sauf pour certains secteurs dans les sous-bassins de la Moselle et de la Sarre, mais surtout de la Meuse. Le SEQ-

Eau montre que les deux polluants les plus préoccupants sur le bassin sont les nitrates et le phosphore. Bien que les détériorations liées à ces polluants restent importantes, une amélioration est constatée, notamment pour le phosphore, suite à la mise en œuvre de traitements spécifiques des effluents urbains, à la réduction du phosphore dans les lessives et à la mise aux normes des bâtiments d'élevage.

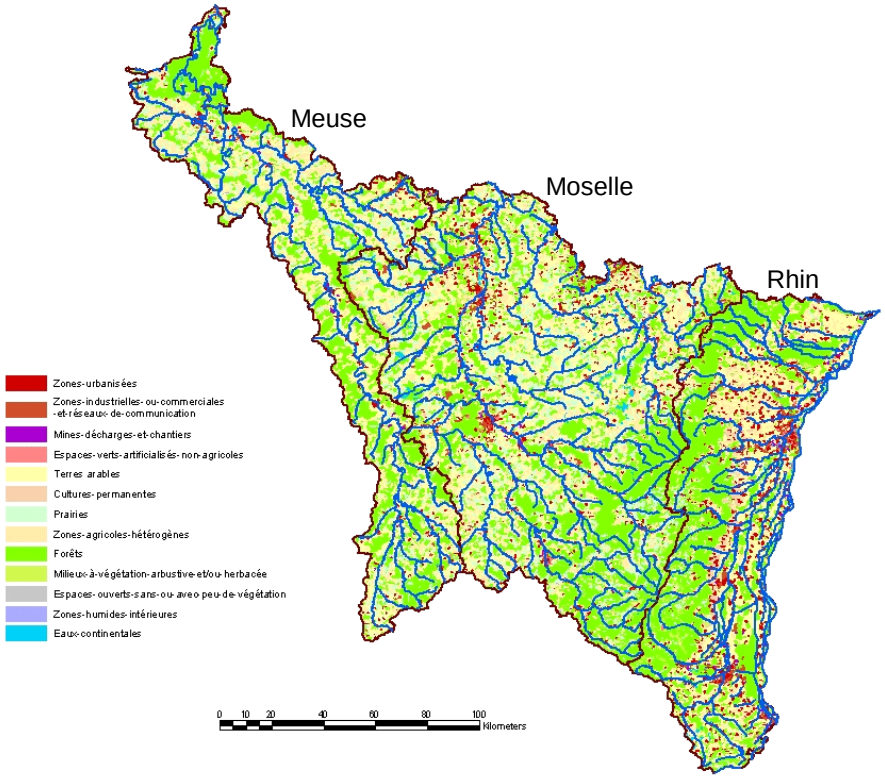


FIG. 2-10 : Bassins de la Moselle, du Rhin et de la Meuse et leurs drains principaux. Occupation du sol sur l'ensemble des bassins gérés par l'agence de l'eau Rhin Meuse

2.4.2 Le bassin de la Moselle

Le bassin de la Moselle comporte deux réseaux hydrographiques indépendants : celui drainé par la Sarre et celui drainé par la Moselle. Pour l'étude des corrélations spatiales, le réseau constitué par la Sarre est écarté, le nombre de stations étant trop faible. L'étude est donc restreinte au réseau hydrographique constitué par les drains principaux de la Moselle. Cet ensemble de cours d'eau représente environ 2500 kilomètres de linéaire, avec une centaine de confluences et une centaine de stations informées pour les nitrates et les débits (FIG. 2-11).

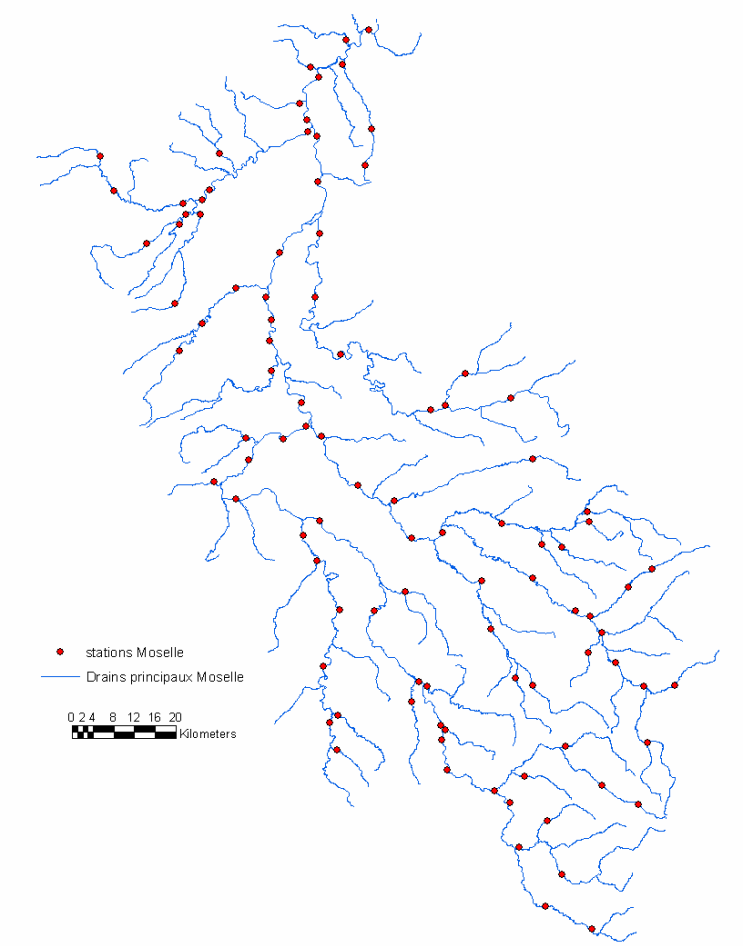
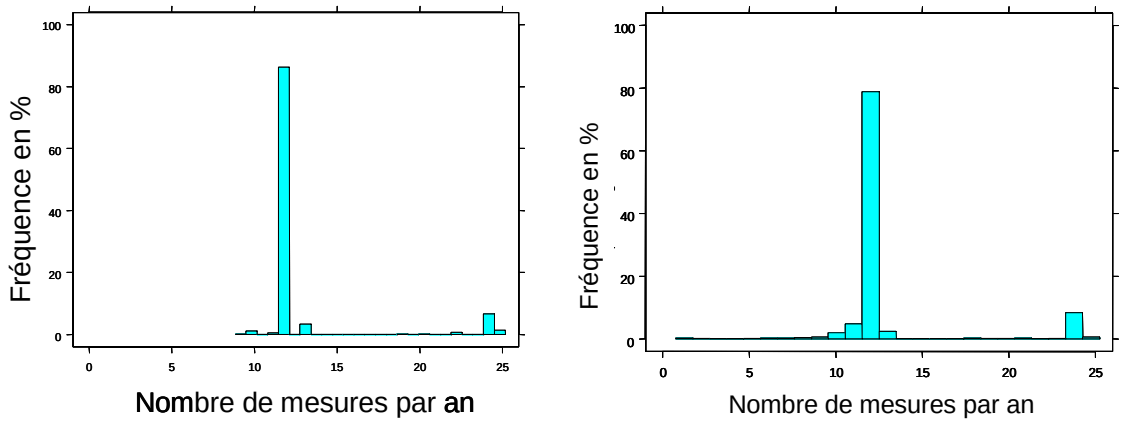


FIG. 2-11 : Ensemble des drains principaux et des rivières contenant au moins une station sur le réseau hydrographique drainé par la Moselle

2.4.3 Les données

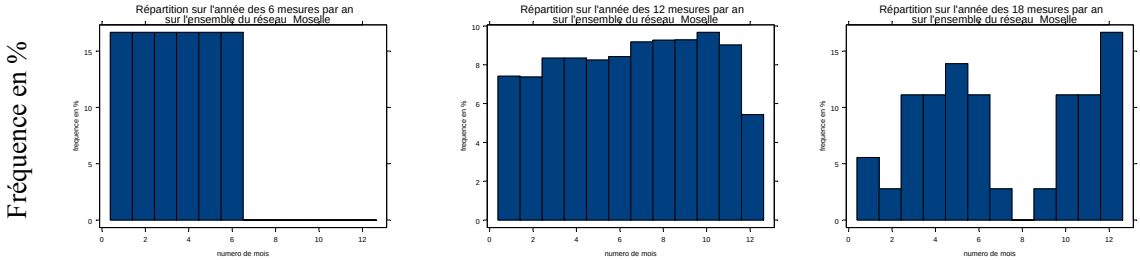
Par rapport au bassin Loire Bretagne, le bassin de la Moselle présente de nombreux avantages. La densité des stations est plus forte avec 99 stations informées pour une surface de 15 405 km2 contre 269 stations étudiées en Loire Bretagne pour une superficie de 155 000 km^2, soit environ quatre fois plus de stations relativement à la surface. Ensuite, les mesures sont fréquentes et plus régulières, en général 12 par an, parfois 24 (FIG. 2-12). En 10 ans, seulement 6% des stations ne respectent pas ce schéma d'échantillonnage (12 ou 24 mesures par an) pour les mesures de polluants, et 13% pour les débits. En Loire Bretagne, environ 35% des 269 stations RNB8-2002 n'ont pas un échantillonnage régulier de 12 ou 24 mesures par an entre 1985 et 2005. Enfin, la majorité des mesures de concentration est associée à un débit instantané, ce qui permet de calculer le flux.

Il existe sur l'ensemble de la France des stations hydrométriques pour lesquelles les mesures de débits sont journalières voire horaires. Mais ces stations de mesures de débit diffèrent de celles des mesures en nutriments. Elles sont aussi moins nombreuses. C'est pourquoi nous avons préféré retenir les données provenant des stations RNB. L'ensemble des mesures entre 1992 et 2003 est disponible sur le site de l'agence Rhin Meuse http://www.eau-rhin-meuse.fr/.



a) Concentrations en nitrates b) Débits

FIG. 2-12 : Histogramme du nombre de mesures par an pour l'ensemble des stations du réseau hydrographique drainé par la Moselle entre 1992 et 2003. En abscisse, le nombre de mesures par an, en ordonnée, leur fréquence.



6 mesures *12 mesures* *18 mesures*

Numéro de mois

FIG. 2-13 : Répartition des prélèvements sur une année en fonction de la fréquence d'échantillonnage. Sur l'ensemble des stations du réseau hydrographique drainé par la Moselle, entre 1992 et 2003, pour une fréquence d'échantillonnage fixée, on calcule le pourcentage de mesures effectuées pour chacun des mois de l'année. En abscisse, le numéro du mois. En ordonnée, sa fréquence.

Les concentrations en nitrates et les débits sont mesurés aux mêmes dates à parfois un jour ou deux près. La stratégie d'échantillonnage est alors globalement identique pour les nitrates et les débits.

Sur 12 ans, la répartition des mesures se résume de la manière suivante (FIG. 2-13):

- Dans le cas de moins de 10 mesures par an, soit environ 2% des cas, les prélèvements sont effectués les 6 premiers mois de l'année.
- Entre 10 et 12 mesures par an, soit 86% des cas, les mesures sont en général espacées de 30 jours en moyenne, avec cependant des mesures moins fréquentes en début et fin d'année, ce qu'il faudra prendre en compte dans l'estimation des moyennes annuelles des débits ou des nitrates.

- De 13 à 23 mesures par an, soit 4% des cas, l'échantillonnage est irrégulier et préférentiel. Le caractère préférentiel diffère selon la taille de l'échantillon. Contrairement au bassin Loire Bretagne, un échantillon de 18 mesures par an ne correspond pas ici à un échantillonnage préférentiel en hiver.
- Avec 24 mesures par an, soit 8% des cas, l'échantillonnage est plutôt régulier mais quelques mesures manquent en février et décembre.

2.4.4 Surfaces drainées, occupation du sol

Pour ce bassin, les relations entre les débits et les surfaces drainées, ou les concentrations et l'occupation du sol, ont été quantifiées à l'aide des données CORINE LAND COVER 1990, disponibles auprès de l'institut français de l'environnement (IFEN). Cette base de données répertorie en Europe tous les éléments géographiques de plus de 25 ha (eau, bois, villes, etc.). Elle a vu le jour avec l'acquisition d'images satellites pour une échelle de 1 :100 000e. Elle repose sur une nomenclature à 3 niveaux et 44 postes répartis selon 5 grands types d'occupation du territoire :

- Territoires artificialisés
- Territoires agricoles
- Forêts et milieux semi-naturels
- Zones humides
- Surfaces en eau

Dans le cadre de la mise en œuvre des réseaux européens de suivi des eaux continentales Eurowaternet (Leonard, 1999), chaque station a été affectée des caractéristiques de la zone hydrographique qui la contient : pourcentage d'agriculture cumulé à l'amont, de forêt…Ces informations qui correspondent à des approximations ont été recalculées à petite échelle dans le modèle PEGASE en milieu de thèse. Pour chaque station, la distance à la source, l'altitude, la surface drainée, l'occupation des sols répartie entre territoires agricoles, zones urbaine, forêts de feuillus et de conifères, prairies et un attribut « divers » sont ainsi disponibles.

2.4.5 Implantation des données sous R

L'ensemble des données a été exporté dans le logiciel R pour le traitement statistique. Ainsi la structure arborescente de la Moselle a été implantée à l'aide de la librairie « graph » qui facilite le travail sur graphes. Les drains principaux ainsi que l'ensemble des stations ont été exportés « manuellement » reconstituant ainsi le graphe défini par les points de confluences, les distances et les arêtes ayant pour poids les distances entre nœuds. Ces manipulations ont été rendues inévitables par manque d'informations et d'outils adaptés pour un traitement sur SIG. Ainsi, le chaînage des cours d'eau de la BD Carthage n'a été disponible qu'en milieu de thèse. Ces manipulations auront toutefois permis de déceler la présence de quelques erreurs dans les bases de données

La figure FIG. 2-14 représente un extrait de la structure d'arbre formée par les principaux cours d'eau du réseau de la Moselle. Les stations sont codées par le nombre à 7 chiffres, qui est leur code RNB. Les points de confluences et les sources sont codés par des nombres allant de 1 à 162, « 1 » étant l'exutoire de la Moselle sur le territoire français.

Les informations telles que les mesures des débits, concentrations en nitrates, orthophosphates, occupations du sol, surfaces de bassins versants drainés ont été intégrées sous forme de listes ou de tableaux.

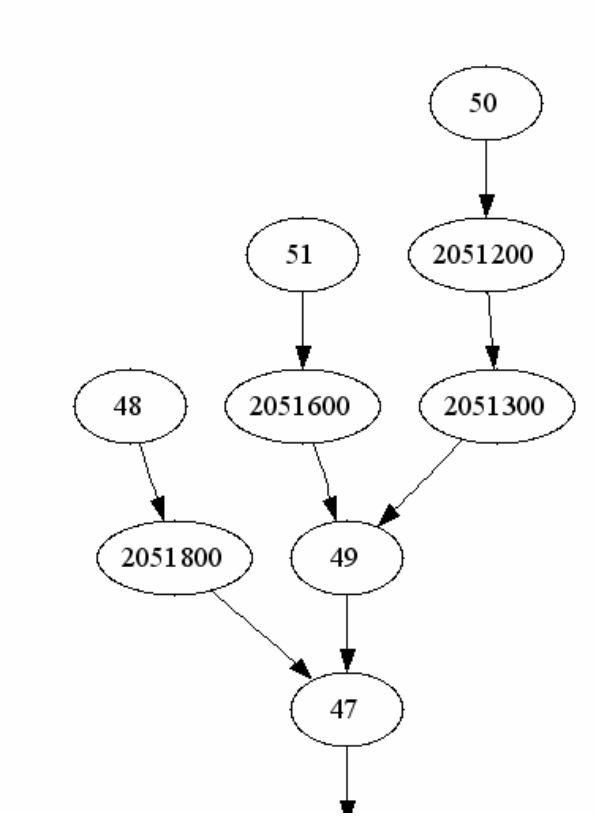


FIG. 2-14 : Extrait de la structure de graphe implantée dans le logiciel R. Réseau hydrographique drainé par la Moselle.

2.5 Chiffres significatifs

Un problème particulier a retenu notre attention sur les bases de données. Lors du premier envoi des données de concentrations en nitrates par l'agence Loire Bretagne, il est apparu que les mesures ne comportaient pas le même nombre de chiffres significatifs et que les quatre derniers chiffres après la virgule étaient identiques pour un grand nombre d'entre elles. Après consultation, nous avons déduit que les mesures fournies par l'agence de l'eau avaient été transformées lors de leur introduction dans la base de données.

Effectivement, la conversion a été effectuée selon le protocole suivant :

- mg N/l en mg NO_3/l division par 0.226
- mg N/l en mg NO_2/l division par 0.304
- mg P/l en mg PO_4/l division par 0.326

Pour homogénéiser la base de données, les agences arrondissent désormais toutes les mesures à deux chiffres après la virgule. Les mesures n'ont donc pas le même nombre de chiffres significatifs selon le paramètre choisi. En particulier, pour les concentrations en nitrites qui peuvent être très faibles, il arrive pour certaines stations que les mesures soient toutes arrondies à une même valeur x et conduit à une variance nulle, donc un intervalle de confiance nul pour la moyenne annuelle.

Partie II
Indicateurs de qualité par station

Nous examinons deux indicateurs de la concentration par station : la moyenne annuelle et le quantile 90. Le calcul de ces indicateurs repose sur une hypothèse d'indépendance temporelle des mesures, ce qui conduit à des biais, en particulier dans le cas d'un échantillonnage préférentiel. La solution proposée consiste à pondérer les données, soit par une méthode géostatistique telle que le krigeage de la moyenne temporelle, soit par une méthode géométrique telle que les segments d'influence. Ces deux pondérations sont comparées sur des exemples. Par ailleurs, le quantile empirique est un estimateur biaisé, même dans le cas de mesures indépendantes. Nous proposons de réduire ce biais par la simple fonction d'interpolation linéaire de la fonction de quantile, des modèles plus compliqués comme le passage par une anamorphose n'améliorant pas significativement les résultats.

Le biais et les incertitudes de chacune des méthodes sont examinés théoriquement et par simulation.

Enfin, les méthodes proposées sont comparées aux méthodes actuelles sur l'ensemble du réseau Loire Bretagne.

Chapitre 3

Les indicateurs

Les calculs d'indicateurs préconisés par les agences font implicitement appel à des outils statistiques classiques.

Nous revenons dans ce chapitre sur les hypothèses de validité de ces calculs dont on montre qu'elles ne sont pas en accord avec les données.

D'autres estimateurs sont alors proposés, fondés sur un formalisme géostatistique.

3.1 Les calculs statistiques usuels

A la base de toute étude statistique, se trouve une population dont on souhaite caractériser certaines propriétés. Généralement on ne dispose que d'un sous-ensemble pour caractériser cette population. Dans le cas de la pollution des cours d'eau, on veut estimer par station la moyenne ou le quantile 90 des concentrations en nitrates sur une année. Dans ce cas, la population est constituée par l'ensemble des concentrations « instantanées » en nitrates, ce qui constitue une population de taille infinie. On estime donc ces valeurs inconnues à l'aide d'un échantillon constitué de mesures effectuées par les agences à divers moments de l'année. Ces mesures peuvent être horaires, journalières, hebdomadaires, mensuelles…Les méthodes statistiques usuelles reposent sur un présupposé : les valeurs expérimentales sont tirées dans la population complète suivant un échantillonnage aléatoire simple, c'est-à-dire suivant *n* tirages équiprobables et indépendants les uns des autres. Les valeurs observées $(z_1, z_2, ..., z_n)$ sont alors représentées par des variables aléatoires et il en est de même des résumés numériques calculés sur ces valeurs.

La statistique classique repose sur une double interprétation: les *n* valeurs observées constituent *n* réalisations indépendantes d'une même variable aléatoire Z ou de façon équivalente, les *n* valeurs observées constituent un seul tirage de *n* variables aléatoires indépendantes et de même loi $(Z_1, Z_2, ..., Z_n)$. La population initiale est donc supposée homogène, c'est-à-dire constituée d'un grand nombre d'éléments obéissant à une même loi de probabilité. En particulier, la moyenne qui représente l'espérance mathématique et le quantile 90 sont supposés identiques pour tous les éléments de la population initiale.

Ce chapitre rappelle l'estimation classique de la moyenne et de son intervalle de confiance ainsi que celle d'un quantile d'ordre quelconque. Nous proposons ensuite des solutions pour améliorer les estimations lorsque les hypothèses sous jacentes ne sont pas respectées.

3.1.1 La moyenne annuelle

3.1.1.1 Estimateur : la moyenne empirique

On se propose donc d'estimer la moyenne, au sens de « l'espérance mathématique » d'une loi de probabilité, qui est la moyenne des valeurs possibles pondérées par leur probabilité (Saporta, 1990 ; Gaudoin, 2001a ; Bernard-Michel and de Fouquet, 2002)

Soit un échantillon $X_1, X_2....X_n$ tel que $X_1, X_2....X_n$ sont des variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X_i) = m$ et de variance $Var(X_i) = \sigma^2$.

L'estimateur usuel de la moyenne est $\bar{X}_n$:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (3\text{-}1)$$

Sous l'hypothèse d'homogénéité, $E(X_i) = m$, l'estimateur de la moyenne est sans biais :

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m \quad (3\text{-}2)$$

et la variance de l'erreur d'estimation commise vaut :

$$Var(\bar{X}_n - m) = Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n} \quad (3\text{-}3)$$

La variance de la somme est égale à la somme des variances sous l'hypothèse d'indépendance des X_i. Lorsque le nombre de données augmente, la variance de l'erreur d'estimation tend vers 0. L'estimation est de plus en plus précise : la moyenne expérimentale converge en moyenne « quadratique » vers le paramètre « espérance mathématique » recherché.

Dans ce calcul, la variance σ^2 de la population totale est supposée connue. En pratique, on dispose seulement de la variance expérimentale $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ des données, qui est un estimateur biaisé de la variance σ^2 :

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

On approche alors la variance d'estimation théorique de la moyenne m par $\frac{\sigma^2}{n} \simeq \frac{S^2}{n-1}$.

3.1.1.2 Intervalle de confiance

La précision de l'estimation est mesurée par la variance d'estimation. On préfère parfois encadrer la valeur inconnue, ici le paramètre m, par un intervalle dans lequel elle a « de fortes chances » de se trouver. C'est ce que traduit la notion d'intervalle de confiance.

Un intervalle de confiance de seuil (ou niveau de signification) $\alpha \in [0,1]$ pour un paramètre θ, est un intervalle aléatoire I tel que $P(\theta \in I) = 1 - \alpha$ où α représente la probabilité que le paramètre θ n'appartienne pas à l'intervalle *I*, c'est-à-dire la probabilité qu'on se trompe en disant que $\theta \in I$. Cette probabilité est choisie assez faible.

Remarque : Les intervalles de confiance suscitent souvent des erreurs d'interprétation et des abus de langage. Si I est un intervalle de confiance à 95% d'un paramètre θ, on ne peut pas en déduire qu'il y a 95% de chance que θ se trouve dans l'intervalle. En fait, si l'on recommence 100 tirages d'un échantillon de n données, et qu'on calcule à chaque fois l'intervalle de confiance associé, on obtient alors 100 intervalles de confiance différents ; en moyenne, θ sera dans 95 de ces intervalles.

Hypothèses de calcul

Pour encadrer la moyenne, on cherche un intervalle encadrant de la forme $\left[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon\right]$. Par définition de l'intervalle de confiance de risque α, ε doit être tel que :

$$P\left(\bar{X}_n - \varepsilon \le m \le \bar{X}_n + \varepsilon\right) = 1 - \alpha$$

c'est à dire

$$P\left(\left|\bar{X}_n - m\right| \le \varepsilon\right) = 1 - \alpha$$

Les calculs supposent les X_i de loi normale, mais aucun test n'est préconisé dans le guide européen (Littlejohn et al., 2002) pour vérifier cette hypothèse. Fausse en toute rigueur, car une concentration est toujours positive et bornée, cette hypothèse n'est pas fondamentale si on dispose d'un grand nombre de données ; on peut alors utiliser le théorème central limite et construire un intervalle de confiance « asymptotique ».

Pour le calcul, on distingue deux cas :

- l'écart type de la loi est connu
- l'écart type est inconnu

Bien que peu réaliste, la première hypothèse est cependant utilisée dans le guide européen ; une discussion détaillée est présentée dans (Bernard-Michel and de Fouquet, 2002).

Notons $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de la loi normale centrée réduite (FIG. 3-1), c'est-à-dire la valeur dépassée dans seulement $\left(\frac{\alpha}{2} * 100\right)\%$ des cas.

L'intervalle de confiance au risque α de la moyenne est :

- $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$ pour un écart type supposé connu, avec $z_{1-\frac{\alpha}{2}}$ quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ d'une variable gaussienne réduite.
- $\left[\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$ pour un écart type supposé inconnu avec $t_{n-1,1-\frac{\alpha}{2}}$ quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ d'une variable de Student à (n-1) degrés de liberté.

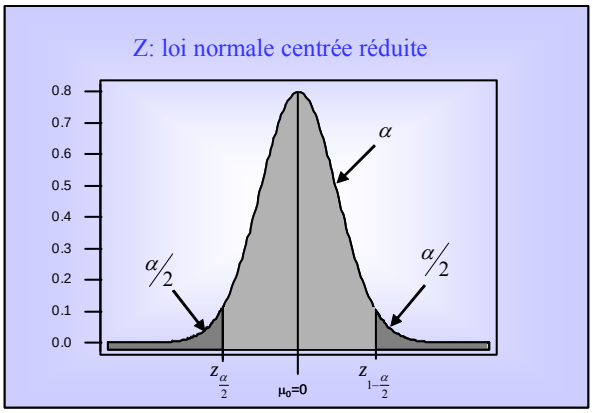


FIG. 3-1 : quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ d'une loi normale centrée réduite

3.1.2 Le quantile 90

3.1.2.1 Estimateur : le quantile empirique

La fonction de répartition d'une variable aléatoire Z est en général définie par:

$$F(z) = P(Z \leq z) \quad (3\text{-}4)$$

Cependant, on trouve également dans la littérature (Saporta, 1990 ; Borovkov, 1987) une autre définition de la fonction de répartition :

$$F(z) = P(Z < z) \quad (3\text{-}5)$$

C'est une fonction monotone croissante, continue à gauche, telle que $F(-\infty) = 0$ et $F(+\infty) = 1$.

Les deux définitions sont équivalentes pour les lois à densité continues. En revanche, pour des lois discrètes, les valeurs des fonctions de répartition diffèrent aux points de discontinuité. La définition (3-5) a été retenue pour la suite, car d'une part elle correspond au calcul de la règle des 90% appliquée par les agences de l'eau, mais d'autre part, elle fournit aux points de discontinuité une estimation plus pessimiste du quantile, ce qui est préféré dans l'étude des pollutions.

Le quantile d'ordre p est défini comme le nombre :

$$q_p = \inf\{x : F(x) > p\} \quad (3\text{-}6)$$

La fonction q est appelée fonction de quantile ; c'est l'inverse de la fonction de répartition.

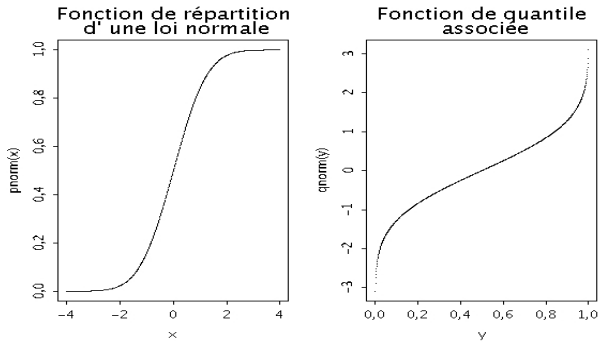


FIG. 3-2 : fonction de répartition et fonction de quantile de la loi normale centrée réduite.

En pratique, la loi n'est accessible qu'à travers un échantillonnage. La fonction de répartition est alors estimée par la fonction de répartition empirique, définie par :

$$F_n(Z) = \frac{1}{n}\sum_i I_{(Z(x_i)<z)} \text{ où les mesures sont les } Z(x_i) \quad (3\text{-}7)$$

Le quantile q_p est estimé par le quantile empirique $\hat{q}_p$, défini par :

$$\hat{q}_p = Z_{(i)} \text{ pour } \frac{i-1}{n} \leq p < \frac{i}{n} \quad (3\text{-}8)$$

où $Z_{(1)}, Z_{(2)}, ..., Z_{(n)}$ est la statistique d'ordre des variables aléatoires $Z_1, Z_2, ..., Z_n$, c'est à dire plus simplement les mesures classées par ordre croissant. Un exemple de quantile empirique est donné FIG. 3-3. Dans le cas particulier où l'ordre du quantile est égal à 90%, l'estimation par le quantile empirique correspond exactement à la règle des 90% du SEQ-Eau (Agence de l'eau Loire Bretagne, 2002). Ces calculs reposent donc sur l'hypothèse d'indépendance des tirages et d'homogénéité de la

loi de probabilité de la population totale. Cet estimateur peut donc être biaisé lorsque ces hypothèses ne sont pas vérifiées. D'autre estimateurs du quantile sont bien évidement envisageables. En particulier, une interpolation linéaire du quantile empirique sera présentée au Chapitre 5.

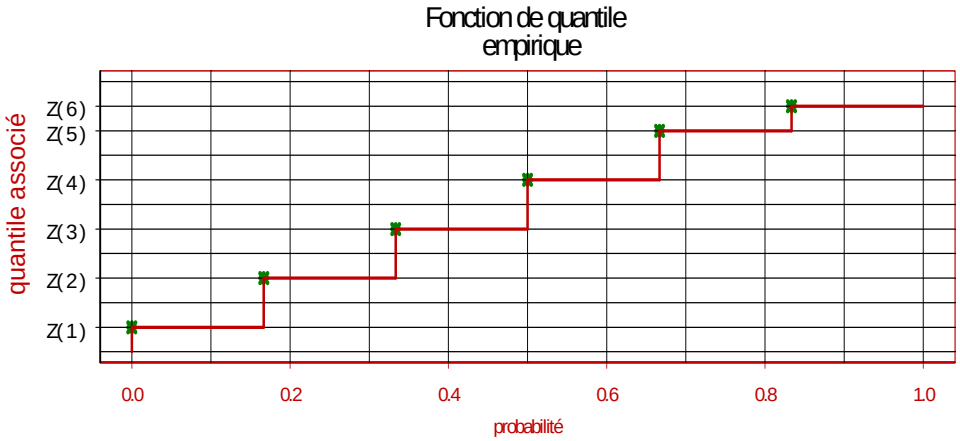


FIG. 3-3 : Fonction de quantile empirique. Sur ce graphe, les Z(i) sont les statistiques d'ordre des six variables aléatoires considérées.

3.1.2.2 Intervalle de confiance du quantile empirique

Les deux méthodes les plus courantes pour calculer l'intervalle de confiance du quantile sont les suivantes (Gaudoin, 2001b):

Méthode 1 :

Soient :

- $(X_i)_{i\in\{1,n\}}$ échantillon de variables aléatoires de même loi de moyenne *m* et d'écart type σ
- $(X_i)_{i\in\{1,n\}}$ indépendantes

alors on montre que $\hat{q}_p$ suit une loi normale $N\left(q_p, \frac{\sqrt{p(1-p)}}{f_X(q_p)}\right)$ avec f_X densité de X.

D'où un intervalle de confiance au risque α :

$\left[\hat{q}_p - \frac{z_{1-\frac{\alpha}{2}} s_n}{\sqrt{n}}, \hat{q}_p + \frac{z_{1-\frac{\alpha}{2}} s_n}{\sqrt{n}}\right]$ avec s_n estimateur de $\frac{\sqrt{p(1-p)}}{f_X(q_p)}$ et $z_{1-\frac{\alpha}{2}}$ quantile $\left(1-\frac{\alpha}{2}\right)$ de la loi $N(0,1)$.

En général, on ne peut pas calculer cet intervalle de confiance car il nécessite la connaissance et l'existence de la densité de la loi.

De plus cet intervalle de confiance étant asymptotique n'est valable que pour un très grand nombre de données. La démonstration utilise en effet le théorème central limite.

Méthode 2 :

C'est sûrement la meilleure méthode car on calcule un intervalle de confiance non asymptotique et les hypothèses sont nettement moins lourdes.

Soient :

- $(X_i)_{i\in\{1,n\}}$ échantillon de variables aléatoires soumises à une même loi de moyenne *m* et d'écart type σ
- $(X_i)_{i\in\{1,n\}}$ indépendantes

S'il existe *i* et *j* tels que $\sum_{k=i}^{j-1} C_n^k p^k (1-p)^{n-k} = 1-\alpha$, alors $[X_{(i)}, X_{(j)}]$ est un intervalle de confiance de $\hat{q}_p$ au risque α.

Cette méthode est inapplicable lorsque les données sont peu nombreuses et en particulier pour calculer un intervalle de confiance du quantile 90. En général, pour un faible effectif, ce quantile 90 correspond à la plus forte valeur mesurée ou à l'avant dernière. On ne peut donc pas calculer d'intervalle de confiance à partir des mesures puisqu'on ne dispose pas de valeur plus forte que l'estimation elle même.

3.1.3 Problèmes liés à ces estimateurs

3.1.3.1 Biais de l'échantillonnage préférentiel

L'hypothèse d'homogénéité de la population initiale revient à supposer la loi des concentrations inchangée selon la saison. L'hypothèse d'indépendance entre les tirages, ou d'absence de corrélation, revient à dire qu'une donnée (une mesure à une date fixée) n'apporte aucune information sur le tirage suivant. Or ces deux hypothèses sont erronées, ce qui a des conséquences importantes en cas d'échantillonnage préférentiel, ou seulement irrégulier.

En de nombreuses stations, les concentrations en nitrates sont plus élevées en « hiver ». Augmenter la fréquence des mesures à cette période augmente automatiquement la moyenne expérimentale, ainsi que les quantiles : à même nombre de mesures, la proportion des valeurs supérieures à une valeur fixée augmente si l'on mesure plus fréquemment les valeurs élevées.

En l'absence de tout modèle probabiliste, considérons la courbe $y = 1 + \cos\frac{2\pi t}{T}$ pour $0 \le t \le T$ de moyenne égale à 1, le minimum à 0, le maximum à 2 et le quantile 90 à 1.951. Avec 12 échantillons régulièrement répartis durant la période T, l'écart entre la moyenne et son estimation est faible, et même nul dans l'exemple présenté (FIG. 3-4 et TAB. 3-1). Mais si l'on augmente la fréquence des mesures pour $0 \le t \le \frac{T}{4}$ ou $\frac{3T}{4} \le t \le T$, alors la moyenne expérimentale des données est augmentée arbitrairement.

Avec un échantillonnage régulier aux 12 dates $\frac{T}{24}, \frac{3T}{24}, ..., \frac{23T}{24}$, la moyenne est estimée sans erreur ; le quantile 90, égal à la 11ème valeur par ordre croissant, est estimé à 1.939. Avec 6 mesures supplémentaires en début et en fin d'année, la moyenne estimée augmente à 1.151, le quantile 90, estimé par l'avant-dernière valeur, restant ici égal à son estimation à partir de 12 échantillons.

Par ailleurs, l'estimation des quantiles par une fonction en escalier introduit des discontinuités en fonction du nombre d'échantillons. Rajouter des données comportant plutôt des valeurs fortes peut

provoquer une diminution du quantile estimé, à cause de l'augmentation de l'effectif, et par suite du changement de rang de la donnée retenue (TAB. 3-1). Selon le cas, le quantile estimé à partir des 18 mesures est supérieur, égal ou inférieur à celui estimé à partir des 12 mesures. Ainsi, le quantile réel peut aussi bien être surestimé ou sous-estimé.

ordre	quantile réels	12 mesures		18 mesures	
		rang	estimation	rang	estimation
90	1.951	11	1.939	17	1.939
85	1.891	11	1.939	16	1.907
80	1.809	10	1.768	15	1.818

TAB. 3-1: Estimation du quantile par la règle des 90%

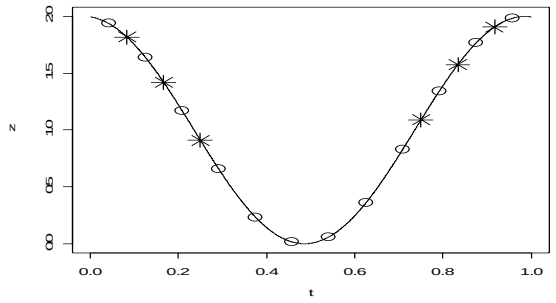


FIG. 3-4 : Exemple d'échantillonnage préférentiel des fortes valeurs

Selon les stations, les concentrations en nitrites et orthophosphates présentent des variations saisonnières dans le même sens que celles des nitrates, ou opposées. L'échantillonnage préférentiel des nitrates entraîne alors une surestimation ou une sous-estimation selon les cas. Pondérer les valeurs expérimentales permet de corriger ces biais, en attribuant un poids plus important à des valeurs espacées, et un poids plus réduit lorsque la fréquence des mesures augmente. Différentes pondérations sont envisageables. La pondération par « segment d'influence », facile à automatiser, sera présentée comme une simplification du krigeage.

En dehors des variations systématiques (fluctuation saisonnière, « tendance » pluriannuelle), on constate qu'une mesure à une date donnée apporte généralement de l'information sur les valeurs aux dates voisines. Contrairement au modèle du dé idéal, pour lequel la valeur observée lors d'un lancer ne fournit aucune information sur celle obtenue lors du lancer suivant, les valeurs mesurées ne sont pas réparties « au hasard pur » au cours du temps. Dans ce cas, une mesure à une date donnée ne serait alors pas plus représentative des valeurs durant la semaine de mesure qu'à six mois d'intervalle. Cette liaison entre les concentrations à des dates proches est décrite par la « corrélation temporelle ». En général, en-dehors des fluctuations saisonnières, le degré de liaison entre deux concentrations diminue lorsque l'intervalle de temps augmente. Supposons qu'après une mesure indiquant une forte concentration, l'intervalle de temps entre les prélèvements soit réduit. En l'absence de toute fluctuation saisonnière cette fois, on observerait encore un échantillonnage préférentiel conduisant à

surestimation de la moyenne. Là encore, une pondération atténuant l'influence des valeurs rapprochées permettra de corriger le biais.

En pratique, l'estimation géostatistique corrige les effets des irrégularités de l'échantillonnage, sans avoir à faire la part entre les variations saisonnières et la corrélation temporelle.

3.1.3.2 Variance d'estimation

En présence de corrélation temporelle, les variances d'estimation calculées par statistique sont erronées puisque les covariances temporelles entre les mesures à différents instants sont supposées nulles. L'intervalle de confiance annoncé par la méthode s'avère donc inexact.

Prenons l'exemple d'une chronique pour laquelle les mesures sont journalières et supposons que l'indicateur recherché est la moyenne arithmétique des concentrations sur ces 365 jours notée m_{th}.

Si on extrait un échantillon de numérique de taille *n* $x_1, x_2, ...x_n$, alors :

- m_{th} est alors estimé en statistique classique par $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et la variance d'estimation annoncée est $\sigma_E^* = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Elle dépend de la taille de l'échantillon et de la dispersion des mesures autour de l'indicateur estimé.
- L'erreur d'estimation est $\overline{x} - m_{th}$

Supposons maintenant que l'on dispose de N échantillons tous de taille *n*. On note alors $x_1^j, x_2^j, ...x_n^j$ l'échantillon j:

- A partir de l'échantillon *j*, on estime m_{th} par $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$ et la variance d'estimation associée annoncée par statistique est : $\sigma_{E,j}^{2*} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(x_i^j - \overline{x}\right)^2$.
- L'erreur d'estimation pour chaque échantillon est $\overline{x}_j - m_{th}$ avec $j \in \{1,..n\}$.

Alors le biais de l'estimateur de la moyenne peut être approché expérimentalement par $\overline{X} - m_{th}$ où $\overline{X} = \frac{1}{N}\sum_{j=1}^{N}\overline{x}_j$ est la moyenne sur tous les échantillons des estimateurs. Dans le reste du document, nous désignerons par « biais expérimental » cette grandeur.

De même, sur ces *N* échantillons, la variance d'estimation est approchée expérimentalement par $\frac{1}{N}\sum_{j=1}^{N}\left(\overline{x}_j - m_{th}\right)^2$. Approchant la variance des erreurs, cette grandeur est désignée comme le carré moyen résiduel et sa racine comme l'écart type résiduel.

Pour vérifier la pertinence du calcul du biais et de la variance d'estimation d'une méthode, il est intéressant de comparer ces quantités expérimentales à celles annoncées par la théorie. Dans le cas de la variance d'estimation, il faudrait comparer les *N* variances d'estimations annoncées $\sigma_{E,j}^{2*}$ au carré

moyen résiduel. Une méthode simplifiée consiste à comparer la moyenne de ces variances $\frac{1}{N}\sum_{j=1}^{N}\sigma_{E,j}^{2*}$ qui devrait être proche du carré moyen résiduel.

Nous verrons à l'aide d'un exemple sur la station « Quai du roi » que cette condition n'est pas vérifiée pour l'estimation statistique.

3.1.3.3 Biais du quantile empirique

Il est connu que le quantile empirique est un estimateur biaisé, même pour des données indépendantes (Gaudoin, 2001b). Cela signifie que si l'on disposait par exemple d'une centaine d'échantillons de 12 mesures, et que l'on appliquait la règle des 90% à chacun de ces échantillons, alors ces estimations seraient en moyenne différentes de la valeur théorique. De plus, ce biais dépend fortement de la taille de l'échantillonnage. Il n'est donc pas recommandé de comparer deux quantiles calculés avec un nombre de mesures différent. En particulier, l'application de la règle des 90% à l'ensemble des stations RNB est à interpréter avec précaution puisque le nombre de mesures diffère par site.

Mathématiquement, le biais de la méthode des 90% est calculable si on connaît la loi des statistiques d'ordre, donc la loi des mesures elles mêmes.

Loi des statistiques d'ordre :

Soient $X_1, X_2, ...X_n$ des variables aléatoires indépendantes de fonction de répartition *F*, de loi f continue, alors, $X_{(i)}$ admet pour densité :

$$\forall x \in IR,\ f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i} f(x) \quad (3\text{-}9)$$

On en déduit les moments des statistiques d'ordre :

$$E(X_{(i)}) = \int x \cdot f_{X_{(i)}}(x)dx \quad (3\text{-}10)$$

Ces expressions n'étant généralement pas explicites, on utilise généralement l'approximation numérique suivante:

$$\forall i \in \{1,....n\}, E(X_{(i)}) \approx F^{-1}(\frac{i}{n+1}) \quad (3\text{-}11)$$

De même, la variance de la statistique d'ordre peut être approchée par l'expression suivante :

$$Var(X_{(i)}) \simeq \frac{i(n-i+1)}{(n+1)^2(n+2)f^2\left[F^{-1}\left(\frac{i}{n+1}\right)\right]} \quad (3\text{-}12)$$

Biais du quantile :

On obtient facilement une valeur approchée du biais du quantile puisque chaque quantile correspond à une statistique d'ordre. Le biais d'un quantile se déduit donc de l'espérance de la statistique d'ordre correspondante. Par exemple pour 10 mesures, le quantile 90 correspond à la statistique d'ordre 10, pour 12 mesures, à la statistique d'ordre 11 etc. Pour une loi uniforme entre 0 et 1, l'espérance et la variance théorique du quantile 90 se calculent aisément en fonction de la taille de l'échantillonnage.

- De 1 à 10 mesures : $E(\hat{q}_{90}) = \frac{n}{n+1}$, $Var(\hat{q}_{90}) = \frac{n}{(n+1)^2(n+2)}$
- De 11 à 20 mesures : $E(\hat{q}_{90}) = \frac{n-1}{n+1}$, $Var(\hat{q}_{90}) = \frac{2(n-1)}{(n+1)^2(n+2)}$
- De 21 à 30 mesures : $E(\hat{q}_{90}) = \frac{n-2}{n+1}$, $Var(\hat{q}_{90}) = \frac{3(n-2)}{(n+1)^2(n+2)}$
- De 31 à 40 mesures : $E(\hat{q}_{90}) = \frac{n-3}{n+1}$, $Var(\hat{q}_{90}) = \frac{4(n-3)}{(n+1)^2(n+2)}$

La figure FIG. 3-5 présente le calcul théorique de l'espérance et de la variance du quantile 90 en fonction de la taille de l'échantillonnage. Le quantile 90 pour une loi uniforme entre 0 et 1 est égal à 0.9, on en déduit que le biais est très important pour moins de 10 mesures et présente des discontinuités pour une taille d'échantillonnage multiple de 10. En général, le quantile est nettement sous estimé bien que la règle des 90% soit asymptotiquement sans biais d'après le théorème de Mosteller (Saporta, 1990). Son utilisation n'est donc justifiée que lorsque les observations sont nombreuses, ce qui n'est jamais le cas pour les mesures de concentrations dans les rivières en une station par année. Comme la loi des mesures n'est pas connue a priori, il n'est pas possible de prévoir le biais par un calcul théorique.

Au chapitre 5, des méthodes de corrections empiriques sont examinées.

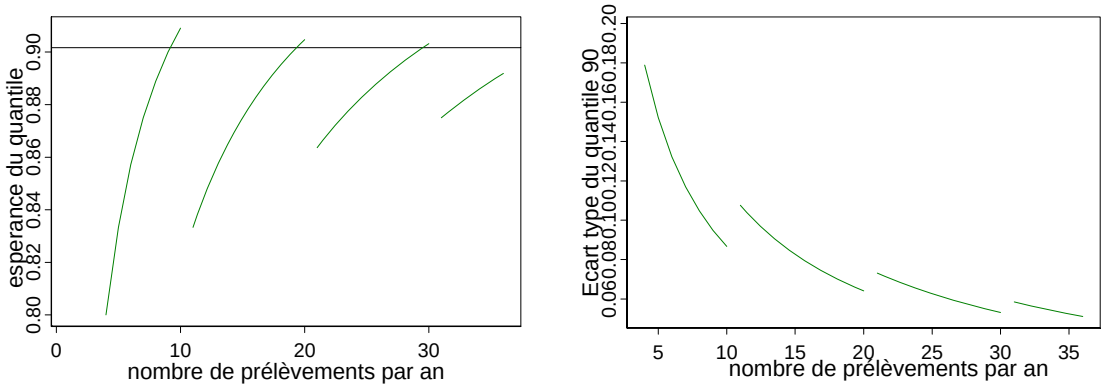


FIG. 3-5 : Courbes théoriques de l'espérance et de l'écart-type d'un quantile 90 estimé par le SEQ-Eau en fonction du nombre de prélèvements. Loi uniforme.

3.2 Quels indicateurs estimer ?

Supposons la concentration parfaitement connue, car mesurée exhaustivement durant une année (une mesure chaque minute, par exemple). Pour l'année j considérée, la concentration peut être décrite par une fonction $z(t)$, avec $t_j \leq t \leq t_j + T$. La moyenne de la concentration durant l'année j est définie comme la moyenne temporelle $z_j^{an} = \frac{1}{T}\int_{t_j}^{t_j+T} z(t)dt$. Son estimation à partir des n mesures effectuées durant l'année revient à approcher cette intégrale par une somme discrète. Si les mesures sont

régulièrement réparties durant l'année, l'intégrale $I = \frac{1}{T}\int_{t_j}^{t_j+T} z(t)dt$ est approchée par la moyenne $\frac{1}{n}\sum_{i=1}^{n} z(t_i)$, ce qui revient à attribuer la même pondération $\frac{1}{n}$ (la même « largeur ») à toutes les mesures. Lorsque les mesures sont irrégulièrement espacées, l'intégrale I peut être approchée par une somme de rectangles de largeur variable : c'est la méthode des segments d'influence. Le krigeage fournit alors une pondération « optimale », dépendant de la régularité de la courbe $z(t)$.

La notion de moyenne annuelle durant l'année j reste parfaitement définie en dehors de toute hypothèse d'homogénéité de la concentration durant l'année. Cette moyenne annuelle de la concentration est compatible avec une périodicité annuelle comme avec une « tendance » pluriannuelle marquée.

L'hypothèse d'homogénéité de la moyenne apparaît alors imposée par la méthode de calcul « statistique », plutôt que par la physique du phénomène. En effet, le calcul statistique usuel consiste à estimer la moyenne au sens de l'espérance mathématique de la loi de probabilité. Pour pouvoir inférer ce paramètre statistique, on pose naturellement l'hypothèse que les différents tirages disponibles sont tous issus de la même loi, et donc que les concentrations par station admettent la même espérance au cours du temps.

Par la suite nous noterons Z_T la moyenne temporelle et *m* l'espérance de la loi.

Deux définitions de la « moyenne »

Quelle est la relation entre ces deux « moyennes », le paramètre espérance mathématique d'une loi de probabilité, objet du calcul statistique usuel, et la moyenne temporelle ? Pour les lois de probabilité usuelles, le paramètre espérance mathématique est la limite de la moyenne temporelle calculée sur un très grand nombre d'années, en l'absence de périodicité ou de « dérive » pluriannuelle.

Supposons durant une année les 365 concentrations journalières parfaitement connues. Alors, la moyenne annuelle sur l'année, égale à la moyenne arithmétique des 365 valeurs journalières, est parfaitement connue. Sa variance d'estimation devrait donc être nulle. Or le calcul statistique indique que cette variance d'estimation est égale à $\frac{S^2}{n-1}$, S^2 désignant la variance expérimentale des 365 valeurs, évidemment non nulle. En effet, si l'estimation de l'espérance mathématique s'améliore quand le nombre de données augmente, elle reste toujours inconnue, même à partir de 365 valeurs journalières, 730 (durant 2 ans), ou 1095 (durant 3ans).

En l'absence de « dérive » à moyen ou long terme, supposer par exemple l'espérance mathématique constante durant un siècle signifie que les concentrations oscillent autour de cette « moyenne ». Mais durant le siècle, les 100 moyennes annuelles seront (généralement) toutes différentes. Elles oscilleront autour de l'espérance, l'amplitude de ces oscillations étant très atténuée par rapport à celle des valeurs journalières.

La transcription des calculs statistiques classiques conduit ainsi à « se tromper » sur la variable à estimer, et par suite sur la méthode d'estimation appropriée.

Quelques notions de modélisation

Si la moyenne (l'espérance de la loi) est variable durant l'année, alors il en est de même de l'histogramme expérimental. Comme l'espérance, la variance n'est pas nécessairement homogène durant l'année.

Finalement, quelle est la loi de probabilité dont on cherche à connaître certaines caractéristiques : espérance (qui serait alors égale à la moyenne annuelle de la concentration), variance, quantile ?

Un point de vue peut être le suivant : considérons l'ensemble des concentrations au cours de l'année. Si l'on prend au hasard une mesure durant l'année, à un instant aléatoire θ uniforme sur $[t_0, t_0 + T]$ on obtient une variable aléatoire $z(\theta)$. La loi de probabilité dont on cherche à connaître la moyenne, la variance et le quantile est celle de cette variable aléatoire $z(\theta)$. Les paramètres de la loi de cette variable aléatoire se calculent en remarquant que θ admet la densité $\frac{dt}{T}$ sur l'intervalle $[t_0, t_0 + T]$:

- espérance: $E[z(\theta)] = \frac{1}{T}\int_{t_0}^{t_0+T} z(t)dt$, notée z^{an} ;
- variance : $\mathrm{var}\, z(\theta) = \frac{1}{T}\int_{t_0}^{t_0+T} (z(t) - z^{an})^2\, dt$;
- quantile 90 : c'est la valeur dépassée durant l'année pendant 10% du temps.

  Ce quantile q_{90} est défini par $\frac{1}{T}\int_{t_0}^{t_0+T} 1_{z(t)>q_{90}}\, dt = 0.10$, $1_{z(t)>q_{90}}$ désignant l'indicatrice de la concentration z à la coupure q_{90} : $1_{z(t)>q_{90}} = 1$ lorsque $z(t) > q_{90}$ et 0 sinon.

Lorsque θ est uniforme sur $[t_0, t_0 + T]$, à un intervalle de temps τ fixé (par exemple 30 jours), les concentrations $z(\theta)$ et $z(\theta + \tau)$ sont deux variables aléatoires qui ne sont plus indépendantes mais corrélées, le niveau de corrélation dépendant de l'intervalle de temps τ

Ce point de vue est celui de la géostatistique transitive (Matheron, 1970). La concentration est décrite comme une variable régionalisée, c'est-à-dire une variable présentant une certaine structure temporelle. Cette variable est échantillonnée à des dates supposées aléatoires, par exemple tous les 30 jours à partir d'une origine aléatoire (entre le 1er et le 30 janvier de chaque année).

En pratique, pour mener les calculs, certaines hypothèses de stationnarité restent nécessaires, mais elles sont beaucoup moins strictes que dans la statistique classique.

Le point de vue utilisé en pratique est encore un peu différent (Matheron, 1970). La concentration $z(t)$, décrite comme une variable régionalisée (une fonction numérique définie sur $[t_0, t_0 + T]$) est considérée comme le résultat d'un tirage possible parmi beaucoup d'autres. L'ensemble des tirages possibles, c'est-à-dire l'ensemble des variables régionalisées admettant la même structure spatiale, constitue une Fonction Aléatoire, notée $Z(t)$. On se proposera par exemple d'estimer la moyenne annuelle $z^{an} = \frac{1}{T}\int_{t_0}^{t_0+T} z(t)dt$, vue comme un tirage de la variable aléatoire $Z^{an} = \frac{1}{T}\int_{t_0}^{t_0+T} Z(t)dt$. Les concentrations aux dates t et $t + \tau$ sont considérées comme un tirage des deux variables aléatoires corrélées $Z(t), Z(t + \tau)$. Là encore, certaines hypothèses de stationnarité, moins strictes que pour la modélisation statistique, restent nécessaires. Par exemple, les hypothèses de stationnarité temporelle portent sur les écarts $Z(t + \tau) - Z(t)$ et non plus sur les concentrations

$Z(t)$; on montre (Matheron, 1970) que la méthode d'estimation proposée reste valide même en supposant la variance de $Z(t)$ lentement variable dans le temps.

En résumé, l'apport essentiel de la géostatistique est le suivant :

- les concentrations sont désormais considérées comme (des tirages de variables) corrélées temporellement ;
- les hypothèses de stationnarité sont allégées par rapport au modèle statistique ;
- la moyenne annuelle à estimer est l'intégrale temporelle $z^{an} = \frac{1}{T}\int_{t_0}^{t_0+T} z(t)dt$, et non plus le paramètre « espérance mathématique » d'une loi de probabilité (pour marquer la différence, on utilisera aussi l'expression de valeur annuelle de la concentration).
- Le quantile 90 annuel à estimer est le quantile 90 d'une variable aléatoire tirée uniformément dans $[t_0, t_0 + T]$. Il est défini par q_{90} tel que $\frac{1}{T}\int_{t_0}^{t_0+T} 1_{z(t)>q_{90}}\, dt = 0.10$, $1_{z(t)>q_{90}}$ désignant l'indicatrice de la concentration z à la coupure q_{90}.

Dans la suite *T* est discrétisé en 365 jours.

Chapitre 4

Estimation de la moyenne

La notion de support pour les mesures doit ici être introduite. En théorie, les concentrations sont supposées instantanées. En pratique, les mesures ne sont pas prélevées instantanément, cette opération étant de l'ordre de quelques minutes. Quel support faut-il donc retenir pour caractériser le moment de la mesure ? La minute ? L'heure ? La journée ? Ce choix n'est pas sans conséquence sur les calculs de variance et de quantiles, mais cette question n'est pas prise en compte par la suite. Compte tenu du faible nombre de mesures par an, nous assimilons chaque mesure à une mesure journalière.

4.1 Estimation géostatistique: le krigeage

Le krigeage (Chilès and Delfiner, 1999 ; Emery and Arnaud, 2000) permet de tenir compte non seulement des irrégularités d'échantillonnage mais aussi de la corrélation temporelle. L'avantage de cette méthode est qu'elle propose le meilleur estimateur linéaire au sens de la minimisation de la variance d'estimation. Les biais des échantillonnages préférentiels sont corrigés et les intervalles de confiance annoncés sont plus précis que ceux prédits par statistique classique.

En pratique, la variable estimée est la moyenne temporelle $Z_T = \frac{1}{T}\int_T Z(t)dt$. Les poids optimaux sont calculés par krigeage ordinaire à moyenne inconnue, et ont leur somme égale à 1. Ainsi les données avec leurs pondérations sont cohérentes avec l'estimation de la moyenne annuelle. Une introduction à l'analyse variographique et au krigeage est reportée en annexe A.

De manière pratique, le krigeage est un peu plus compliqué à implanter que les statistiques classiques puisqu'il demande le calcul du variogramme expérimental et sa modélisation (par l'utilisateur ou automatiquement - un programme semi-manuel a été mis au point à cet effet) par station. Mais il permet la cohérence avec l'estimation « optimale » de la moyenne annuelle recherchée et le calcul de la précision de cette estimation.

Un exemple de poids de krigeage de la moyenne annuelle, pour un échantillonnage préférentiel avec 18 mesures par an, est présenté en FIG. 4-1. Bien que les variogrammes des nitrites et ceux 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 sont égaux et le krigeage de la moyenne coïncide alors avec l'estimateur statistique usuel.

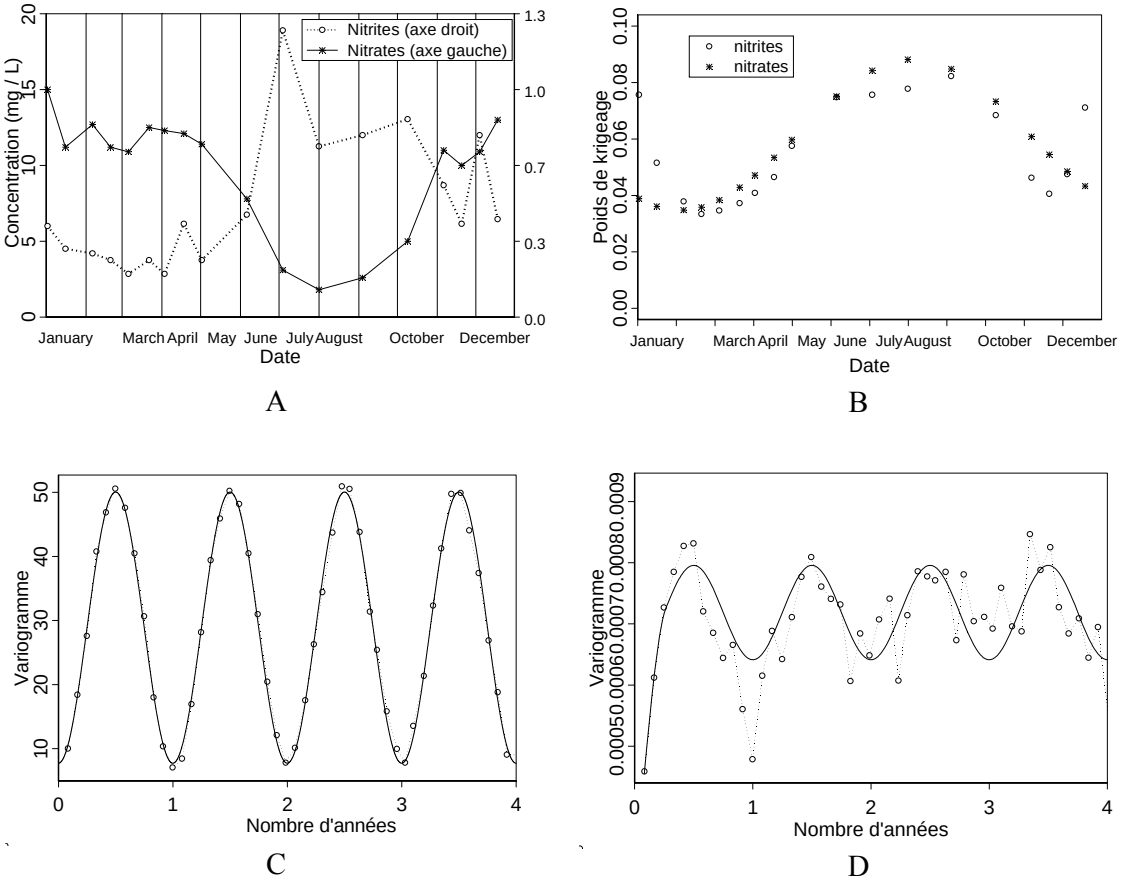


FIG. 4-1: A) Concentrations en nitrates et nitrites en 1995, station 56000 sur la Loire. B) Poids de krigeage pour la moyenne annuelle, échantillonnage préférentiel avec doublement de la fréquence des mesures en période hivernale. C) Variogramme temporel des nitrates. D) Variogramme des nitrites.

Dans les estimations, seules les données de l'année à étudier ont été retenues. Il semblerait logique d'utiliser les mesures de l'année précédente puisque les mesures sont corrélées dans le temps. De plus, en beaucoup de stations, les mesures sont peu nombreuses, il est donc intéressant d'utiliser les mesures voisines afin d'étudier la précision des estimations.

L'influence du voisinage a été étudiée sur la station 50500 à Orléans. Des dates régulièrement espacées ont été retenues de 1985 à 1987 (soit une mesure tous les 15 du mois). L'étude des poids en fonction du voisinage montre que les poids accordés aux mesures antérieures ou postérieures à l'année estimée sont significatifs. Ils participent jusqu'à 10% dans l'estimation de la moyenne, et la précision est améliorée jusqu'à 12%. Cependant, compte tenu de la périodicité, les résultats sont complexes. Jusqu'à un voisinage de six mois, la moyenne estimée diminue sans cesse. Pour un voisinage d'un mois, les concentrations ajoutées sont celles de décembre et janvier, on pourrait donc s'attendre à une augmentation de la moyenne puisque ce sont des valeurs fortes. Cela tient au fait que les mesures en milieu d'année (en été) prennent alors un poids plus important. A partir d'un voisinage de 6 mois, les poids redeviennent sensiblement égaux sur l'année à estimer. Le gain de précision annoncé par le krigeage est sensible. Il parait donc justifié de retenir des mesures extérieures à l'année courante pour en estimer la moyenne (par exemple de 6 mois). Pour l'année en cours, on ne dispose généralement

que des mesures de l'année précédente, et certaines stations présentent des variations très importantes d'une année à l'autre. La prise en compte de mesures voisines est donc à considérer avec prudence.

Année + i mois	+ 0	+ 1	+ 2 s	+ 3	+ 4	+ 5	+ 6	+ 7	+ 8	+ 9	+ 10	+ 11	+ 12
Moyenne	9.00	8.96	8.96	8.81	8.71	8.60	8.57	8.58	8.56	8.52	8.58	8.60	8.62
Ecart type d'estimation	0.50	0.49	0.48	0.48	0.47	0.46	0.45	0.45	0.45	0.45	0.44	0.44	0.44

TAB. 4-1 : Effet du voisinage sur l'estimation de la moyenne annuelle et de l'écart type d'estimation associé.

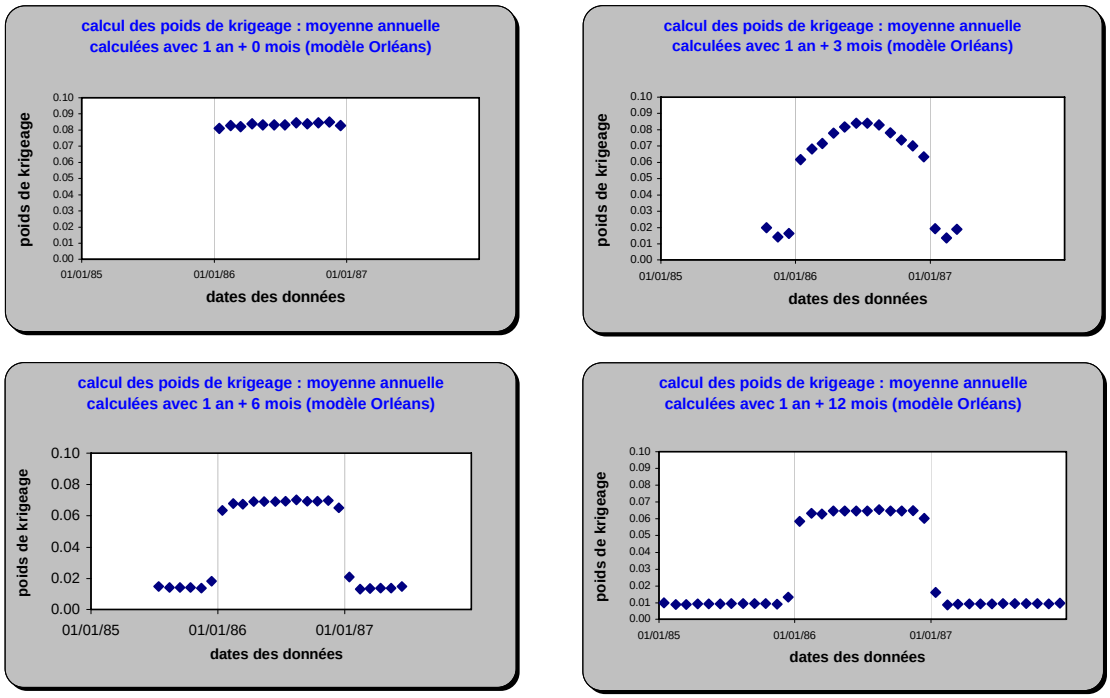


FIG. 4-2 : Poids de krigeage de l'estimation de la moyenne annuelle en 1986 à Orléans, nitrates.

4.2 Une simplification : les segments d'influence

Dans l'exemple de la figure FIG. 4-1 B, les poids de krigeage sont sensiblement égaux à 1/12 (=0.083) pour les mesures mensuelles et à 1/24 (= 0.042) lorsque l'échantillonnage est doublé en hiver.

A la demande des agences de l'eau, une simplification du krigeage est proposée, la pondération par segments d'influence, dans laquelle les poids sont proportionnels à l'intervalle de temps séparant les dates de mesure. Le principe de cette méthode est le suivant :

On trace les médiatrices des segments joignant une date de prélèvement à la date voisine et on affecte à chaque mesure un poids égal au segment qui l'entoure divisé par 365,25 (voir figure FIG. 4-3). La borne inférieure du premier segment et la borne supérieure du dernier coïncident avec les extrémités

de la période totale (l'année) considérée. Cette méthode est bien adaptée pour un resserrement des mesures à n'importe quelle période car elle affecte des poids faibles aux mesures rapprochées. Elle présente cependant deux inconvénients :

- Dans le cas d'un dispositif « fermé », c'est-à-dire avec des mesures situées aux extrémités du tronçon, les poids affectés à la première et à la dernière mesure sont calculés différemment des autres. Cependant il est possible que cela n'entraîne pas une grosse différence dans l'estimation, en particulier lorsque les mesures sont nombreuses. Si le dispositif est centré, c'est-à-dire si les mesures sont effectuées au milieu de chaque mois (de chacun des segments), cet inconvénient disparaît.
- Cette méthode ne tient pas compte de la corrélation temporelle, mais seulement de l'irrégularité de l'échantillonnage. Pour des mesures irrégulières, il n'y a pas de raison particulière pour que cette pondération soit optimale.

Le principal avantage de cette pondération reste sa facilité de mise en œuvre.

Ecrivons mathématiquement l'estimateur de la moyenne une fois les poids p_i calculés.

- Si $X_1, X_2 .... X_n$ sont des variables aléatoires réelles indépendantes et de même loi, d'espérance $E(X_i) = m$ et de variance $Var(X_i) = \sigma^2$.

Alors, un estimateur pondéré de la moyenne est $\overline{X}_n$:

$$\overline{X}_n = \sum_{i=1}^{n} p_i X_i \quad (4\text{-}1)$$

Sous l'hypothèse d'homogénéité, $E(X_i) = m$ pour chacune des variables, l'estimateur de la moyenne est sans biais :

$$E\left(\overline{X}_n\right) = E\left(\sum_{i=1}^{n} p_i X_i\right) = \sum_{i=1}^{n} p_i \times E(X_i) = m \quad (4\text{-}2)$$

La variance d'estimation de la moyenne annuelle X_T se calcule suivant la relation classique en géostatistique :

$$Var(X_T^{\ *} - X_T) = -\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} p_\alpha p_\beta \gamma(t_\alpha - t_\beta) - \overline{\gamma}(T,T) + 2\sum_{\alpha=1}^{n} p_\alpha \frac{1}{T}\int_T \gamma(t_\alpha - t)dt$$

où $\overline{\gamma}(T,T)$ désigne la valeur moyenne du variogramme entre deux points, dont l'un parcourt l'année T et l'autre parcourt l'estimateur (voir en fin d'annexe A).

L'estimation de la valeur de cette variance nécessite le calcul du variogramme temporel.

En vue d'une application systématique, peut-on simplifier le calcul de cette variance ?

Une solution consisterait à calculer et ajuster automatiquement le variogramme pour en déduire la variance par la relation précédente, mais dans ce cas, il n'y aurait plus d'intérêt particulier à utiliser les segments d'influences, le krigeage pouvant être appliqué.

La question a été posée d'une approximation de cette variance par le calcul de la variance statistique d'estimation de l'espérance, tenant compte de ces poids. Nous donnons donc dans la suite le calcul de cette variance, qui correspond à une approximation de la variance d'estimation de la moyenne annuelle (au sens de l'intégrale temporelle) dans le cas d'un variogramme pépitique de palier σ^2. Par

abus de language, ce calcul est présenté comme « variance d'estimation par segments d'influence » dans la suite. Il est approché par $\sigma^2 \sum_{i=1}^{n} p_i^2 \simeq \frac{S^2}{1 - \sum_{i=1}^{n} p_i^2}$.

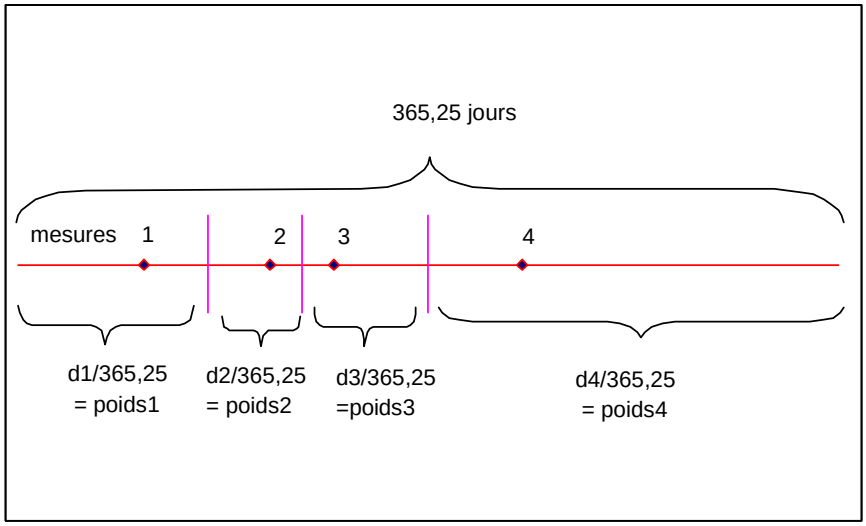


FIG. 4-3 : Méthode des segments d'influence

4.3 Réflexions sur les pondérations

Les atouts et inconvénients respectifs du krigeage ou de la pondération par segments d'influence sont ici examinés au travers d'exemples de concentrations en nitrate.

L'atout principal des segments d'influence est la facilité d'implantation; cette méthode corrige efficacement les biais causés par un échantillonnage augmenté sur une période donnée. Cependant, elle ne prend pas en compte les corrélations temporelles ni dans l'estimation de la moyenne annuelle, ni dans le calcul de variance d'estimation qui lui a été associé. De plus, contrairement au krigeage, elle ne fournit pas une variance d'estimation minimale.

Le krigeage s'adapte à la fois dans le cas d'échantillonnages préférentiels et irréguliers en fournissant des poids optimaux. Dans le cas d'un échantillonnage régulier, tous les poids sont en pratique égaux. Les inconvénients sont d'une part l'implémentation difficile car il nécessite la modélisation d'un variogramme par station et la résolution d'un système de krigeage. D'autre part, les poids obtenus peuvent être négatifs. Les poids deviennent négatifs, par exemple dans le cas de six mesures annuelles, toutes estivales ou de configurations anormales. Des poids négatifs sont a priori admissibles, mais associés à des valeurs élevées de la concentration, ils pourraient exceptionnellement fournir des estimations négatives, ce qui n'a jamais été observé dans les nombreux exemples traités. En revanche, dans ce cas, nous verrons par la suite que la méthode proposée pour le calcul du quantile 90 devient impossible.

Sur l'ensemble des stations RNB traitées (269 au total, et 21 ans de mesures), tous les cas de poids négatifs rencontrés correspondent à un échantillonnage inférieur à 8 mesures par an, donc pour lequel toutes les mesures sont concentrées sur une même période et ne permettent absolument pas d'estimer de façon pertinente la moyenne annuelle ou le quantile. En pratique, augmenter la composante

pépitique du variogramme suffit à rendre positifs les poids de krigeage. Mais dans une procédure automatique, signaler les poids de krigeage négatifs permettrait de détecter certaines configurations problématiques de l'échantillonnage. Considérons par exemple la station 100000 sur le Thouet à Missé et regardons l'évolution des moyennes annuelles sur 18 ans, calculée par statistique, krigeage ou segments d'influence (voir FIG. 4-4, gauche). Les moyennes par statistiques sont systématiquement inférieures aux moyennes pondérées, ce qui provient de l'échantillonnage préférentiel en été. Une ligne verticale est ajoutée sur la figure FIG. 4-4 à gauche pour les années pour lesquelles les poids de krigeage sont positifs, aucune ligne n'étant reportée s'ils sont négatifs. Entre 1991 et 1996, les poids de krigeage sont négatifs. Ces années correspondent à un échantillonnage entièrement concentré en été. Pour les années restantes, les mesures sont mieux réparties et les poids de krigeage positifs.

Les années pour lesquelles les poids et les estimations sont étudiés sont reportées à la figure FIG. 4-5 et au tableau TAB. 4-2.

Comme le montre le tableau TAB. 4-2, si les deux pondérations (krigeage et segments d'influence) sont proches, l'écart type d'estimation annoncé par le krigeage est bien plus faible.

Sur cet exemple, le krigeage permet effectivement d'alerter sur la configuration des données qui a surtout des conséquences importantes dans le calcul des quantiles.

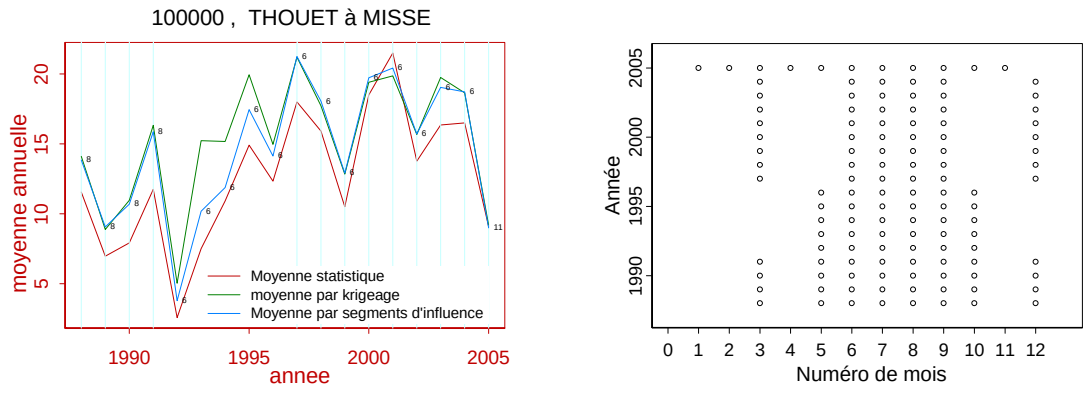


FIG. 4-4 : A gauche ; évolution de la moyenne annuelle des concentrations en nitrates par statistique, moyenne pondérée par segments d'influence, krigeage. A droite, mois de prélèvements par année.

		Statistique	Krigeage	Segments d'influence
1993	Moyenne annuelle	7.5	15.22	10,18
	Ecart type d'estimation	3.03	1.98	3.57
2003	Moyenne annuelle	16.35	19.75	19.03
	Ecart type d'estimation	5.09	4.99	10.82

TAB. 4-2 : Moyenne annuelles en 1993 et 2003 et écarts types d'estimation estimés par statistique, krigeage et moyenne pondérée par segments d'influence

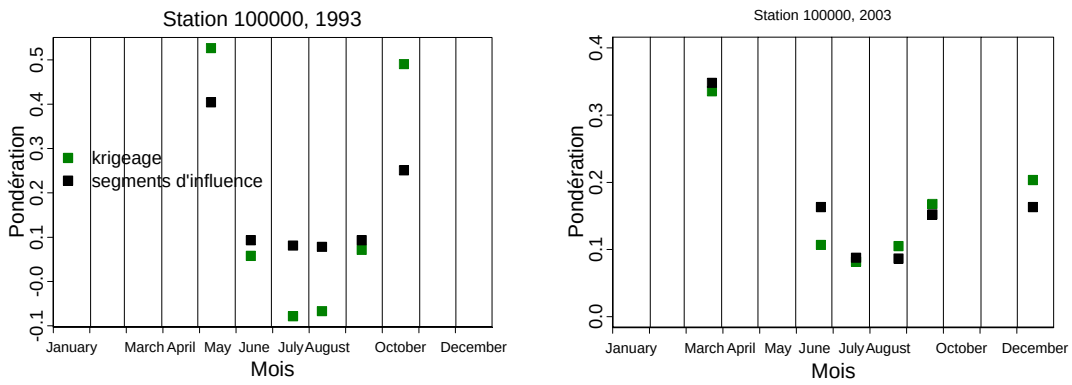


FIG. 4-5 : A gauche : poids de krigeage et segments d'influence, station 100000 en 1993. A droite : poids de krigeage et segments d'influence, station 100000 en 2003

Considérons maintenant (FIG. 4-6) deux stations pour lesquelles les configurations d'échantillonnage recouvrent l'essentiel de celles rencontrées sur le bassin. Pour la station 215500 ont été extraites 3 années de concentrations en nitrates avec respectivement 6, 9 et 12 mesures par an. Les chroniques, poids et estimations sont présentées FIG. 4-6 (gauche). Pour la station 216000, ont été extraites 3 années de concentrations en nitrates avec respectivement 12, 15 et 18 mesures par an. Les chroniques, poids et estimations sont présentées FIG. 4-6 (droite). Les conclusions sont les suivantes :

- Avec 6 mesures par an regroupées en été, certains poids de krigeage sont négatifs. De plus, les estimations de la moyenne risquent d'être sous estimées quelle que soit la méthode utilisée, puisque les concentrations sont faibles en été.
- Avec 9 mesures par an, le poids affecté par segment d'influence à la mesure d'octobre, la dernière de l'année, est nettement plus élevé que les autres. En l'absence de mesures de concentrations élevées en novembre et décembre, on affecte donc un poids fort à une valeur faible. La conséquence est une sous estimation de la moyenne par segments d'influence. L'estimation par statistique classique est encore inférieure car elle ne prend en compte que 3 valeurs fortes.
- Avec 12 mesures par an, si les mesures sont régulièrement espacées, poids de krigeage et de segment d'influence sont similaires et à peu près tous égaux sauf aux extrémités (voir station 215500, FIG. 4-6, gauche). Les estimations sont alors similaires pour les deux pondérations. Si les mesures ne sont pas tout à fait régulières, par exemple FIG. 4-6, droite, station 216000, pour laquelle une mesure manque en janvier, entre avril et mai, et quatre mesures sont concentrées sur novembre-décembre, alors les poids diffèrent. Les estimations restent cependant proches quelle que soit la méthode considérée.
- Avec 15 mesures par an, la majorité des prélèvements correspondent à des valeurs fortes (janvier, avril, novembre, décembre) alors qu'une valeur faible manque en mai. Krigeage et segments d'influence corrigent le biais de l'échantillonnage et fournissent des estimations proches.
- Avec 18 mesures par an, les mesures sont resserrées en hiver. Le biais induit est corrigé à la fois par le krigeage et les segments d'influence.

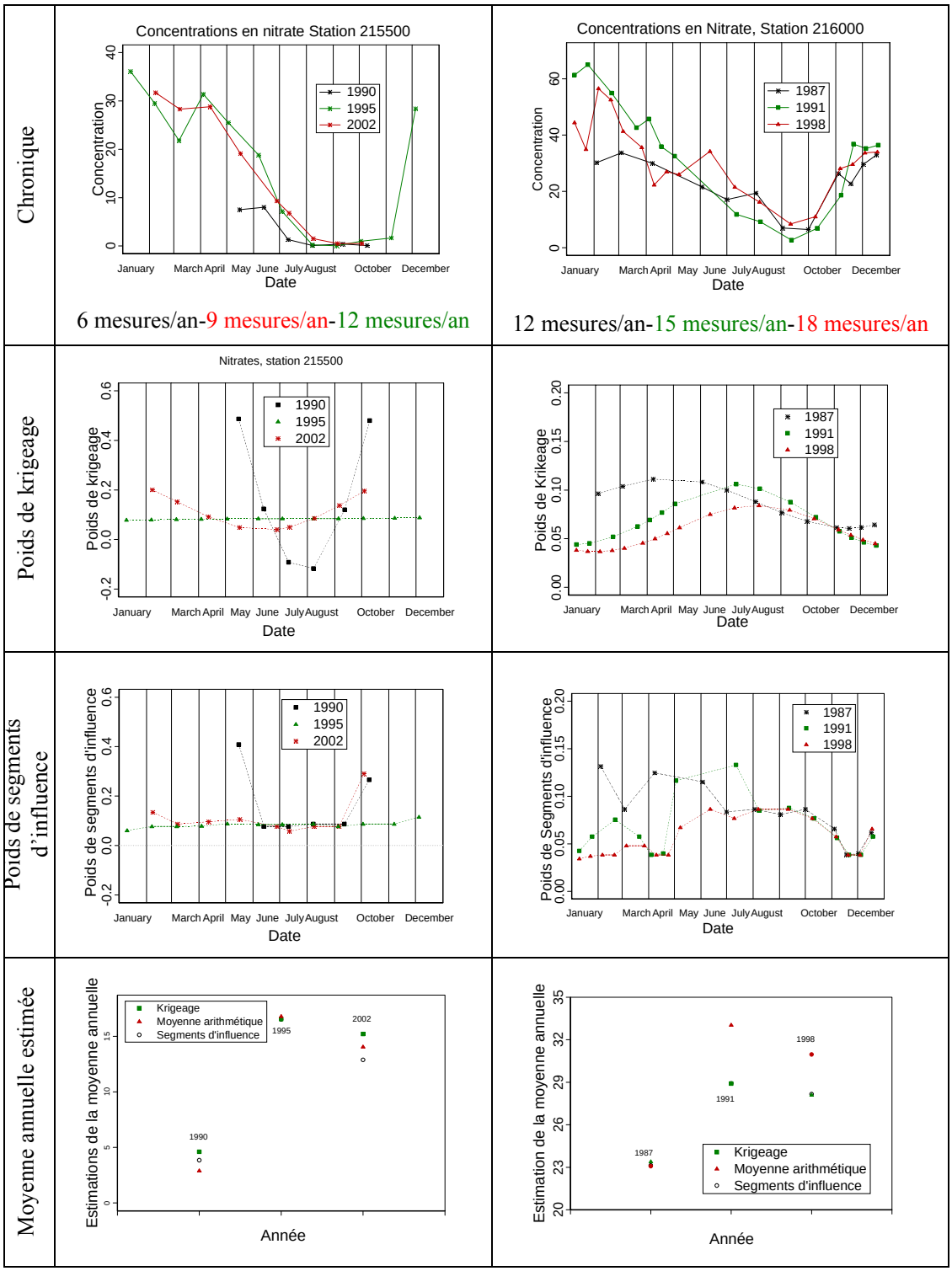


FIG. 4-6 : Comparaison des poids de krigeage, segments d'influence et des estimations de la moyenne annuelle pour différentes stratégies d'échantillonnage et différentes stations. A gauche, station 215500. A droite, station 216000.

Sur cet exemple, le krigeage et les segments d'influence donnent des estimations proches, même si les poids sont assez différents. La méthode des segments d'influence parait efficace lorsque les mesures sont resserrées sur une période mais moins bien adaptée en cas d'échantillons « incomplets » par exemple avec moins de 10 mesures par an dont une majorité en été et aucune certains mois d'hiver. Dans ce cas, le krigeage est préférable car la pondération prend en compte la périodicité annuelle, à travers la composante en cosinus. De plus l'intervalle de confiance annoncé par le krigeage est toujours plus faible. Enfin, en certaines stations, aucune corrélation n'est mise en évidence par le variogramme. Dans ce cas, l'hypothèse d'indépendance des statistiques classiques est valable, et il n'y a pas de raison d'affecter des poids différents aux données, ce que fait pourtant la technique géométrique des segments d'influence. Cela entraîne parfois des conséquences sur les estimations.

4.4 Modélisation de la série temporelle ?

Dans la majorité des cas, les chroniques en nitrate montrent une périodicité qui se retrouve dans les variogrammes, cette composante périodique étant moins marquée pour les nitrites et les orthophosphates. Un autre modèle est donc envisageable lorsque la concentration semble stationnaire durant plusieurs années. Il consiste à modéliser la concentration $Z(t)$ comme la somme de deux termes, une « moyenne périodique » et un résidu :

$$Z(t) = m(t) + R(t) \quad (4\text{-}3)$$

Un exemple de modélisation est donné au paragraphe 6.4 dans le cadre de la vérification des méthodes. Ce modèle avec les estimations de moyennes et quantiles associés est étudié dans (Bernard-Michel and de Fouquet, 2003).

Il n'est pas retenu dans la suite pour les raisons suivantes :

- Avec 12 mesures par an, la moyenne déterministe doit être calculée sur les mesures de plusieurs années. Expérimentalement, la variabilité peut être importante d'une année à l'autre, et dans ce cas la moyenne ainsi obtenue est très incertaine.
- Pour certaines substances, notamment pour les nitrites et les orthophosphates, les chroniques ne présentent pas de périodicité systématique et varient fortement d'une année à l'autre. Certaines présentent des valeurs fortes en hiver pour une année puis des valeurs faibles en hiver une autre année. La « dérive » périodique est donc difficile à définir. Dans le cas d'une procédure automatique, la modélisation de chroniques est plus compliquée que celle des variogrammes, car de nombreux modèles sont possibles, contrairement aux variogrammes pour lesquelles environ 6 types de modèles différents sont rencontrés sur les 269 stations pour les 3 paramètres.
- Enfin, si une modélisation de type « dérive + résidus » peut être utile en prédiction, son intérêt est moindre pour le calcul d'indicateurs annuels.

Sont présentés figure FIG. 4-7 différents types de chroniques. Un graphique représente le groupement de 10 années de chroniques (1996-2005) pour une station et un paramètre donnés. Ces graphiques rassemblent la majorité des types de chroniques rencontrées :

- Comme on le sait, dans la majorité des cas, les chroniques en nitrates présentent une périodicité avec des valeurs fortes en hiver, faibles en été (station 64000), mais l'inverse est aussi rencontré (station 175100). Un nombre non négligeable de stations ne montrent pas de périodicité mais présentent des valeurs fortes en été ou en hiver selon les années (station 80900). Si certaines chroniques sont facilement modélisables (station 6400), d'autres présentent une variabilité

interannuelle trop forte (station 175100), et d'autres sont difficilement modélisables (station 80900)

- Pour les nitrites et les orthophosphates, les types de chroniques sont identiques à ceux présentés pour les nitrates (FIG. 4-7). Les stations sans périodicité annuelle apparente sont plus nombreuses et l'amplitude des fluctuations est souvent importante et les courbes moins régulières.
- Enfin, si la périodicité n'est pas flagrante sur les chroniques, elle apparaît clairement sur le variogramme (FIG. 4-8).

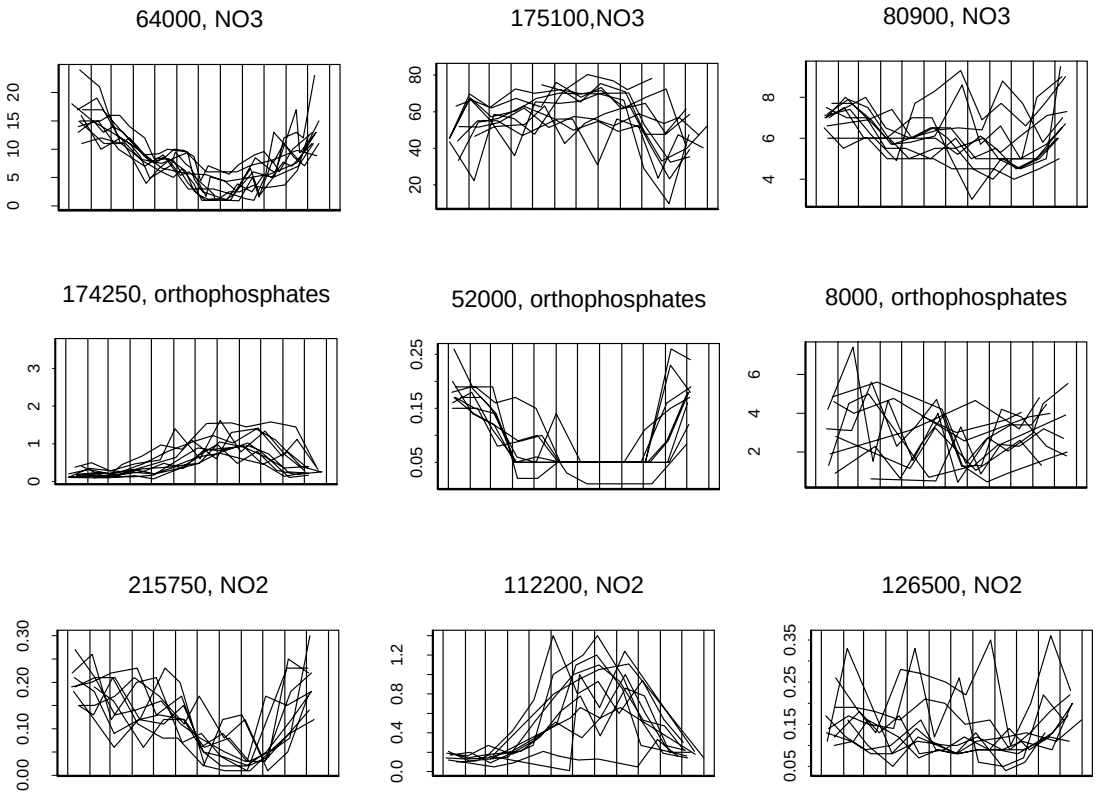


FIG. 4-7: Chroniques annuelles pour des concentrations en nitrates, orthophosphates et nitrites. Les chroniques annuelles sur 10 ans (1996-2002) sont présentées sur un même graphe. En abscisse, la date de mesure (jour + mois), en ordonnée, la mesure.

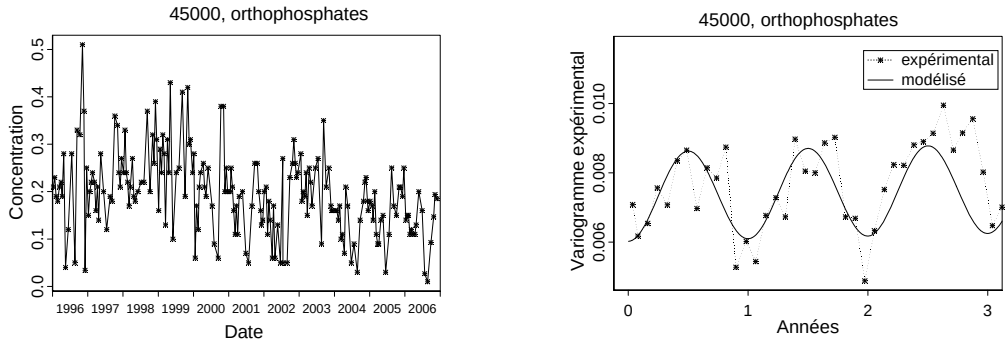


FIG. 4-8 : Chronique temporelle en orthophosphates, station 45000 (à gauche) et variogramme associé sur 3 ans (à droite)

4.5 Conclusion

La pondération par segments d'influence ou krigeage corrige les estimations de la moyenne assez nettement en particulier en cas d'échantillonnage préférentiel. Le krigeage fournit dans certains cas des poids négatifs qui pourraient être embarrassants pour l'estimation des quantiles. Mais les poids négatifs correspondent toujours à des configurations particulières des données qui sont toutes regroupées durant une même période, ce qui ne permet pas d'estimer correctement les indicateurs. L'apparition de poids négatifs peut alors être perçue comme une alerte aux mauvaises configurations. Les segments d'influence sont bien adaptés lorsque les mesures sont resserrées à une période mais ne sont pas toujours logiques lorsque les mesures sont irrégulières. De plus ils fournissent sous hypothèse d'indépendance des mesures des intervalles de confiance pour la moyenne annuelle bien supérieurs à ceux du krigeage. Toutefois, les estimations qu'ils fournissent sont proches de celles du krigeage et leur utilisation peut donc être envisagée dans le cadre d'une procédure automatique. Il faudra cependant traiter avec précaution les stations pour lesquelles les concentrations ne présentent pas de corrélation temporelle car les mesures se voient alors affectées de poids différents selon l'échantillonnage, ce qui n'est pas logique. Le krigeage, au contraire s'adapte à toutes les configurations.

Chapitre 5

Estimation du quantile

Assez développée dans la littérature, la théorie des valeurs extrêmes n'est ici pas adaptée. En effet, le quantile 90 est fort, mais non "extrême". Par ailleurs, les résultats sont toujours asymptotiques et supposent donc un grand nombre de données, alors que la fréquence d'échantillonnage varie généralement de 4 à 18 mesures annuelles par station.

Nous proposons de corriger le biais du quantile empirique par une interpolation linéaire de la fonction de quantile empirique.

Une pondération des mesures par les poids de krigeage de la moyenne annuelle est ensuite présentée dans le cas de mesures corrélées et d'échantillonnage préférentiels.

5.1 Bibliographie

Pour corriger le biais du quantile empirique, plusieurs méthodes de lissage de la fonction de quantile, rencontrés dans la littérature, sont envisageables.

Le principe des L-estimateurs est d'estimer la densité, fonction de répartition ou fonction de quantile d'une variable aléatoire par une combinaison linéaire pondérée des statistiques d'ordre. Une publication très complète (Zielinski, 2004) commente ces différentes méthodes pour des échantillons de petite taille. De nombreux estimateurs sont comparés : le quantile empirique, les estimateurs à noyaux (Harell et Davis, Parzen), les estimateurs pondérés (Huang et Brill). Il en ressort aux travers d'exemples (Zielinsky, 2005) que pour des échantillons de petite taille, le quantile empirique reste finalement le plus approprié. Cet auteur indique qu'il est difficile de démontrer la pertinence d'un estimateur par rapport à un autre pour une loi quelconque. Si de nombreuses propriétés peuvent être déduites pour chacun de ces estimateurs, le choix de l'un par rapport à l'autre ne se justifie souvent que par des exemples. Nous n'avons pas retenu ces méthodes pour plusieurs raisons : d'une part leurs propriétés sont établies pour des données indépendantes. Ensuite plusieurs d'entre elles reposent sur le choix d'une fonction à noyaux qui est utilisée pour pondérer les données dans le but de lisser la fonction de quantile. C'est une hypothèse forte et difficile à justifier sur un nombre de mesures à peine égal à 10. Nous avons préféré la mise en œuvre pratique d'estimations simples sur des cas concrets. De plus, nous ne cherchons pas ici le quantile 90 d'une loi statistique déterminée. En ce qui concerne les ordres proches de 1, la plupart des méthodes proposent des poids dont la somme est égale à 1, le quantile estimé est donc borné par la valeur maximale observée, et le problème de sous-estimation avec moins de 10 mesures n'est pas résolu. Enfin, estimer un quantile 10 mesures ou moins n'est pas suffisant et conduit forcément à une estimation éloignée de la réalité.

Globalement, on trouve rarement des ouvrages sur l'estimation des quantiles en présence de corrélation temporelle ou spatiale. Seule la thèse de D. Florea Draghicescu (Florea Draghicescu, 2002) traite ce sujet, mais beaucoup de données sont nécessaires, la fonction de quantile étant déterminée en fonction du temps.

Dans ce manuscrit, nous nous retenons la méthode de lissage la plus simple qui consiste à interpoler linéairement la fonction de quantile empirique. Elle a été retenue après comparaison à des méthodes de

lissage plus compliquées faisant appel à une fonction d'anamorphose (Lajaunie, 1993 ; Rivoirard, 1994).

5.2 Interpolation linéaire du quantile empirique

On peut remplacer la fonction de quantile empirique en escalier par une fonction de quantile linéaire par morceaux (FIG. 5-1). Fixant a priori que celle-ci passe par « le milieu des marches de l'escalier », c'est-à-dire les points $\left(p_i,(Z_{(i)}+Z_{(i-1)})/2\right)$ avec $p_i=\dfrac{i-1}{n}$, cela revient à remplacer la fonction de répartition empirique en escalier par une fonction de répartition linéaire par morceaux, passant par les points $(Z_i,(p_{i+1}+p_i)/2)$.

Le quantile est alors défini par :

$$Q(x)=\left[\sum_i\left(\frac{Z_{(i+1)}-Z_{(i-1)}}{2*(p_{i+1}-p_i)}\right)(x-p_i)+\frac{(Z_{(i)}+Z_{(i-1)})}{2}\right]I_{x\in]p_i,p_{i+1}]} \quad (5\text{-}1)$$

$$\text{avec } Z_{(n+1)}=Z_{(n)},\ Z_{(0)}=\frac{Z_{(1)}}{2} \text{ et } p_i=F(Z_{(i)})=\frac{i-1}{n}$$

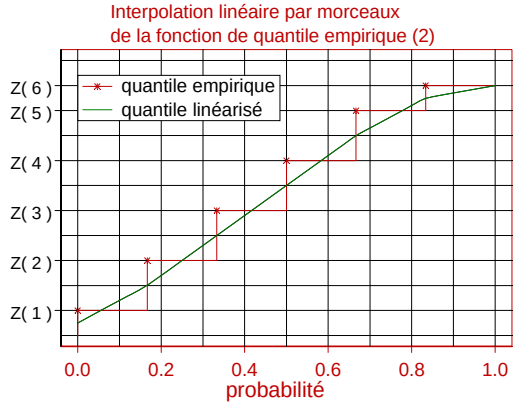


FIG. 5-1: interpolation linéaire du quantile empirique

Cette méthode revient à approcher la loi de Z par une combinaison de lois uniformes.

Dans la suite du manuscrit, nous désignerons cette méthode comme le « quantile linéarisé » ou « quantile interpolé linéairement ».

5.3 Tests de la méthode pour des variables aléatoires indépendantes

Comme pour le quantile empirique, le quantile linéarisé est biaisé. Un calcul théorique de ce biais serait préférable, mais d'une part ce calcul, présenté pour la loi uniforme dans le cas de variables aléatoires indépendantes, est généralement compliqué mais surtout il nécessite la connaissance de la loi des mesures, ce dont on ne dispose jamais. Une approche par simulation a donc été effectuée. Elle

est validée par la comparaison des biais et variances théoriques et expérimentales des deux estimations pour une loi uniforme.

Ainsi, les biais de chaque méthode, déduits par simulation, seront présentés pour des mesures indépendantes issues de lois usuelles, puis pour des mesures réelles pour lesquelles la corrélation temporelle doit être prise en compte (Chapitre 6)

5.3.1 Loi uniforme : calcul théorique du biais

La comparaison du biais est présentée ici pour des variables aléatoires indépendantes de loi uniforme entre 0 et 1.

Rappelons l'équation du quantile d'ordre *x* (*x* appartenant à [0,1]) interpolé linéairement, pour n mesures $Z(i)$, $i \in \{1,...,n\}$ vérifiant une loi uniforme *U(0,1)* :

$$Si\, x \geq 1-\frac{1}{n} \Leftrightarrow n \leq \frac{1}{1-x}, \hat{Q}(x) = \left[\frac{n(x-1)}{2}+1\right]Z(n) + \frac{n(x-1)}{2}Z(n-1)$$

$$Si\, x \leq 1-\frac{1}{n} \Leftrightarrow n \geq \frac{1}{1-x}, \hat{Q}(x) = \frac{1}{2}[(nx-k+1)Z(k+1) + Z(k) + (k-nx)Z(k-1)]$$

$$avec\, k\, tel\, que\, x \in \left]\frac{k-1}{n}, \frac{k}{n}\right]$$

Pour une loi uniforme, les moments des statistiques d'ordre sont connus :

- Espérance et variance de la statistique d'ordre k notée $Z_{(k)}$:

$$E(Z_{(k)}) = \frac{k}{n+1},\ Var(Z_{(k)}) = \frac{k(n-k+1)}{(n+1)^2(n+2)} \quad (5\text{-}2)$$

- Covariance des statistique d'ordre i et j :

$$Cov(Z_{(i)}, Z_{(j)}) = \frac{i(n-j+1)}{(n+1)^2(n+2)} avec\, i < j \quad (5\text{-}3)$$

On en déduit l'espérance et la variance du quantile interpolé linéairement, d'ordre x :

$$Si\, n \leq \frac{1}{1-x}, E(\hat{Q}(x)) = \frac{n}{2(n+1)}(x+1)\ et\ Var(\hat{Q}(x)) = \frac{\frac{n^3}{4}(x-1)^2 + nx}{(n+1)^2(n+2)} \quad (5\text{-}4)$$

$$Si\ n \geq \frac{1}{1-x},\ E(\hat{Q}(x)) = \frac{nx+0,5}{(n+1)} \quad (5\text{-}5)$$

$$Var(\hat{Q}(x)) = \frac{2k^2(n+1) - 2k(2nx+1)(n+1) + 2(x^2n(n-1) + 2x(3n+1)+1)n}{4(n+1)^2(n+2)} \quad (5\text{-}6)$$

Pour le cas particulier du quantile 90, on obtient donc :

$$Si\, n \leq 10, E(\hat{q}_{90}) = \frac{0,95n}{(n+1)} et\ Var(\hat{q}_{90}) = \frac{0,0025n^3 + 0,9n}{(n+1)^2(n+2)}$$

$$Si\, n \geq 10, E(\hat{q}_{90}) = \frac{0,9n+0,5}{(n+1)} et$$

$$Var(\hat{q}_{90}) = \frac{2k^2(n+1) - 2k(1,8n+1)(n+1) + 2(0,81n(n-1) + 1,8(3n+1)+1)n}{4(n+1)^2(n+2)}$$

5.3.2 Evaluation du biais par simulation

La connaissance du biais par un calcul théorique permettrait de la corriger, mais l'expression des moments des statistiques d'ordre est rarement explicite et surtout la loi des mesures est inconnue en pratique. Pour évaluer le biais du quantile empirique, nous procédons donc par simulations. Afin de valider cette démarche, nous allons dans un premier temps construire des exemples pour des simulations de lois statistiques usuelles (normale, lognormale, gamma, exponentielle et uniforme) pour lesquelles le quantile 90 est connu. Par exemple le quantile 90 de la loi normale N(0,1) est de 1.27. Pour examiner les moyennes des écarts, un millier de simulations est en général suffisant. La fréquence des mesures variant différemment selon les stations, nous estimerons le quantile 90 pour des échantillons de taille allant de 4 à 36 mesures. Comme nous travaillons ici sur des mesures indépendantes, la notion de date de prélèvement, et donc par exemple de régularité de l'échantillonnage, n'intervient pas.

5.3.2.1 Validation de la distribution simulée

Les fonctions de simulations du logiciel Splus ont été retenues pour la programmation après comparaison avec nos propres simulations (Box & Jenkins pour la loi normale et méthode congruentielle pour la loi uniforme). Les simulations ont été validées par comparaison de la fonction de répartition empirique sur 1000 mesures avec la courbe théorique (FIG. 5-2). Bien qu'imparfaites pour la loi uniforme et la loi gamma notamment, les courbes ont été jugées suffisamment satisfaisantes avec 1000 simulations. Pour plus de détails sur les méthodes de simulation, se reporter à (Lantuejoul, 2001).

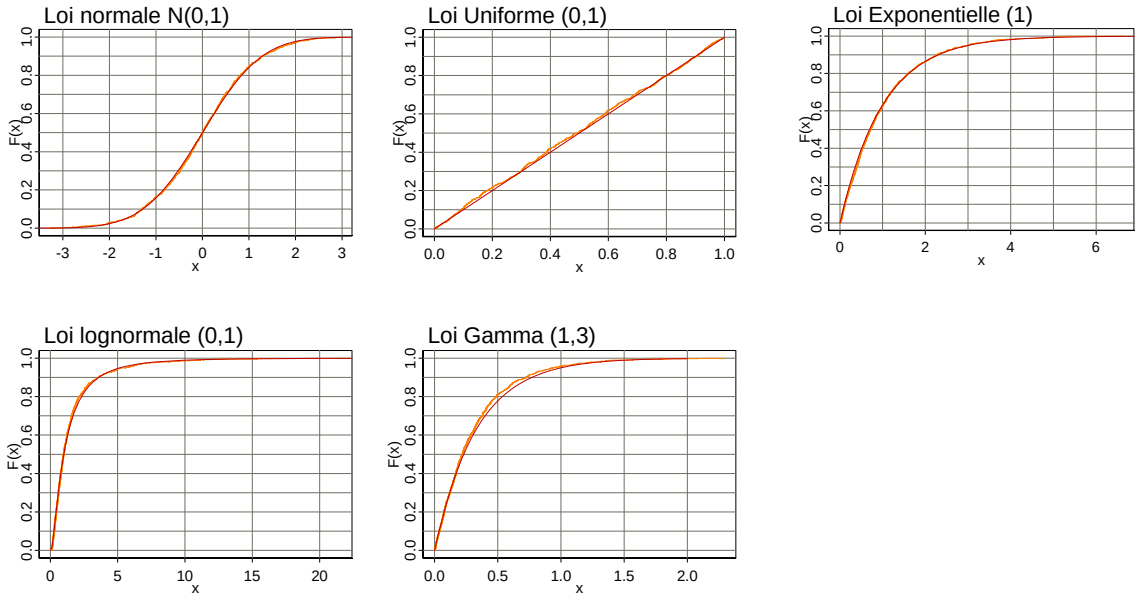


FIG. 5-2: Fonctions de répartition théoriques et expérimentales sur 1000 simulations pour différentes lois : de gauche à droite, normale (0,1), uniforme(0,1), exponentielle(1), lognormale(0,1) et gamma(1,3). En abscisse,

5.3.2.2 Comparaison du biais théorique au biais expérimental pour la loi uniforme

Nous avons simulé N=1000 fois n mesures d'une même loi uniforme entre 0 et 1 pour des échantillons de taille n variant de 4 à 36 mesures. Le quantile 90 de la loi est noté q_{th}. On note $x_1^j, x_2^j, ...x_n^j$ le *jème* échantillon de taille *n. :*

- Sur l'échantillon *j*, q_{th} est estimé par $\hat{q}_j$ par interpolation linéaire du quantile empirique
- Les erreurs d'estimations sur l'ensemble des échantillons sont alors $\hat{q}_j - q_{th}$ avec $j \in \{1,..n\}$.

Alors l'espérance de l'estimateur du quantile est approchée par $\bar{Q}$ où $\bar{Q} = \frac{1}{N}\sum_{j=1}^{N} \hat{q}_j$ et la variance de cet estimateur par $\frac{1}{N}\sum_{j=1}^{N}(\hat{q}_j - \bar{Q})^2$.

Ces résumés statistiques sont comparés aux calculs théoriques du paragraphe 5.3.1 sur la figure FIG. 5-3 pour l'estimation par le quantile empirique, et sur la figure FIG. 5-4 pour le quantile linéarisé.

La démarche est cohérente avec la théorie, malgré quelques écarts non négligeables (pour n=10, 11, 20 et 21) entre l'écart-type expérimental et l'écart-type « théorique » approché. Nous n'avons pas cherché à vérifier si cet écart provenait d'un nombre trop réduit de simulations, jugeant les résultats suffisants pour évaluer le biais d'une méthode.

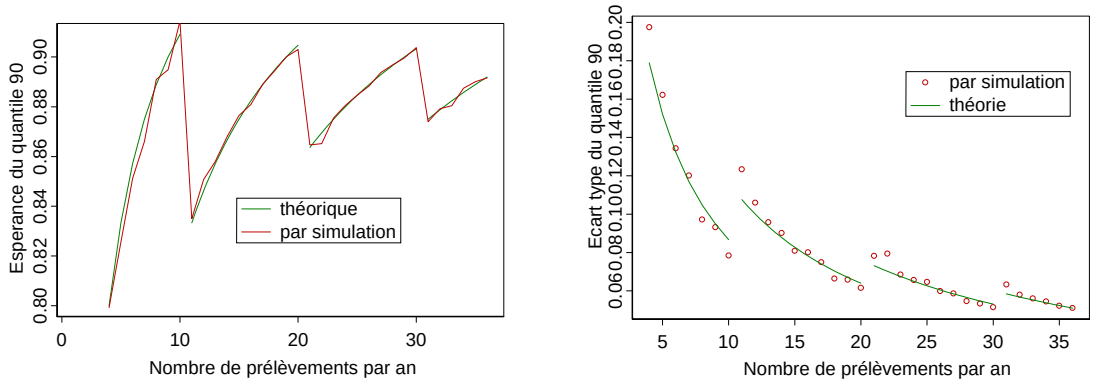


FIG. 5-3 : Courbes théoriques et simulées de l'espérance et de l'écart-type d'un quantile 90 en fonction du nombre de prélèvements. Loi uniforme.

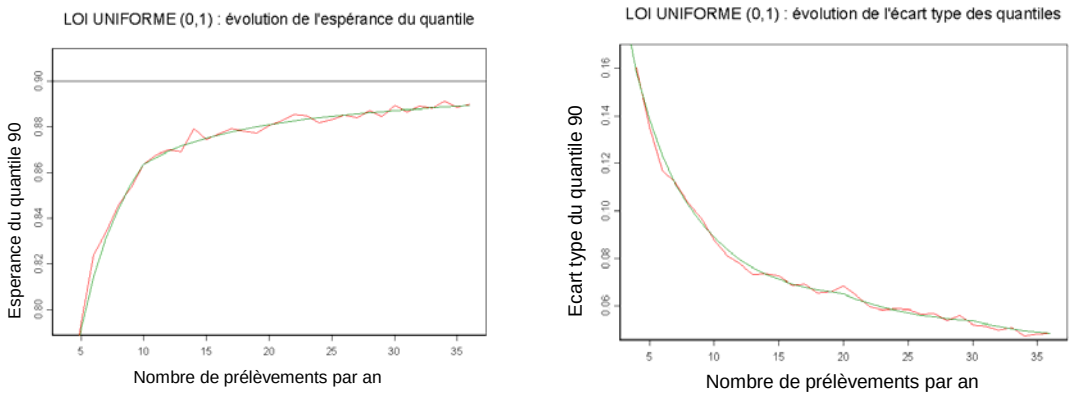


FIG. 5-4 : Courbes théoriques et simulées de l'espérance et de l'écart-type d'un quantile 90 interpolé linéairement, en fonction du nombre de prélèvements, pour une loi uniforme

Les calculs sont relativement simples pour une loi uniforme, mais ils s'avèrent plus compliqués pour les autres lois. L'évaluation du biais par simulation étant cohérente avec la théorie, nous procédons désormais par simulation pour évaluer le biais et la variance des estimateurs des quantiles.

5.3.2.3 Calcul expérimental du biais pour d'autres lois

Quatre types de distribution ont été examinés :

- Loi normale réduite (moyenne nulle, écart type 1)
- Loi lognormale, de transformée logarithmique réduite (moyenne et variance de la loi lognormale : $M = e^{\frac{1}{2}},\ \sigma^2 = M^2(e-1)$, avec e=2.718, soit $M = 1.65$ et $\sigma^2 = 4.67$)
- Loi exponentielle de moyenne 1 et de variance 1
- Loi gamma de moyenne 1/3 et de variance 1/9

Les résultats sont présentés FIG. 5-5.
Pour la loi normale, le biais est très faible à partir de 10 prélèvements par an. Pour les autres lois, il reste important mais plus faible et plus « régulier » que celui induit par la règle des 90%. Aucune de ces lois n'est vraiment représentative de la loi suivie par les concentrations car non seulement elle est généralement bimodale mais de plus les concentrations sont toujours positives. Le calcul théorique de la variance d'estimation n'est pas présentée car difficilement calculable. Il peut cependant être évalué à l'aide du carré moyen résiduel.

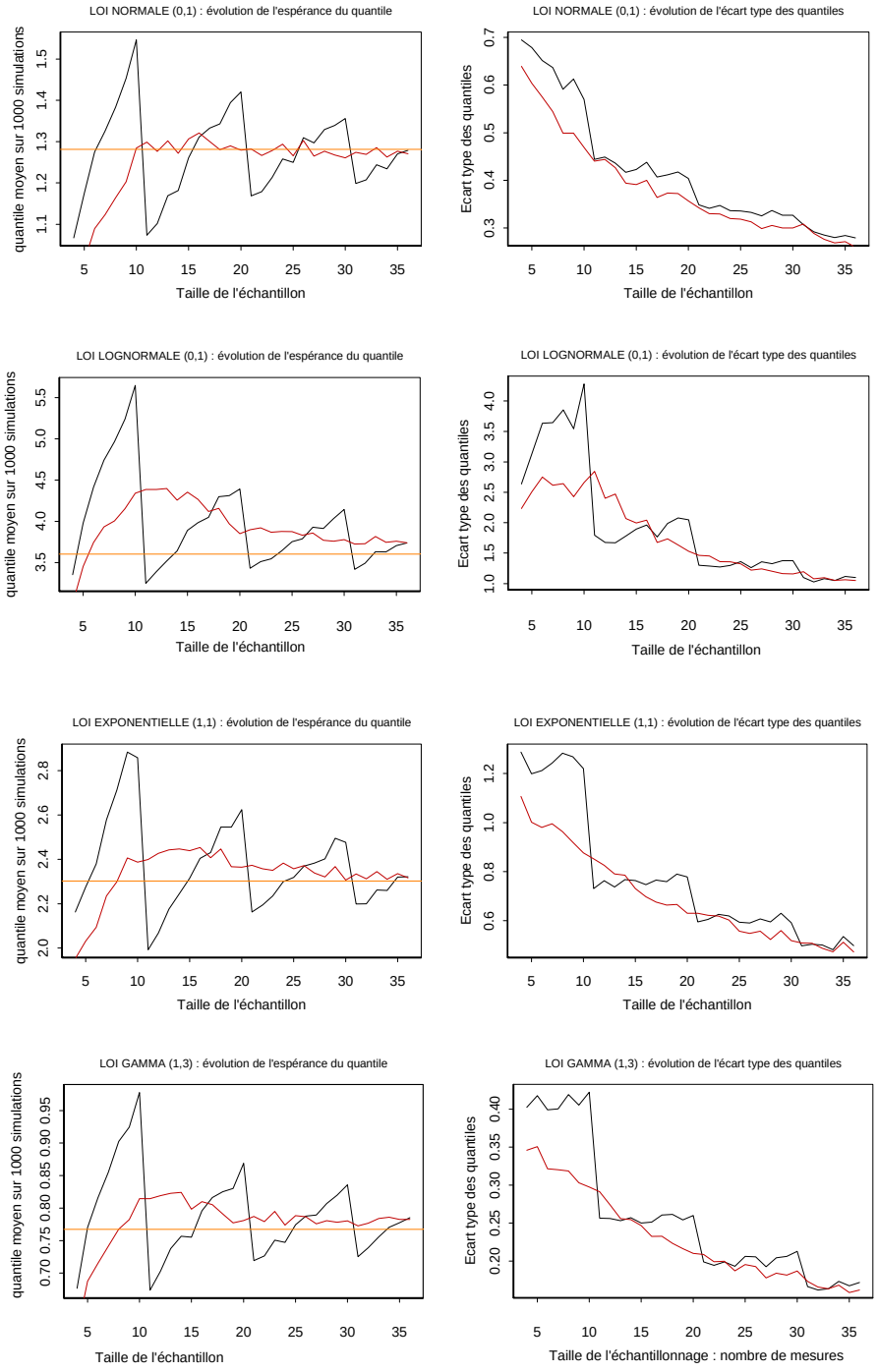


FIG. 5-5 : Evolution de l'espérance et de l'écart type du quantile 90 en fonction de la taille de l'échantillonnage et de la méthode d'estimation (quantile empirique escalier en noir ou linéarisé en rouge) pour différentes lois (normale, uniforme, lognormale, exponentielle, gamma)

5.3.3 Généralisation aux quantiles de tout ordre

Le quantile 90 est préconisé par le SEQ-Eau, mais il est intéressant d'examiner le biais des quantiles d'autres ordres. Le biais du quantile linéarisé (FIG. 5-6) et le biais relatif (FIG. 5-7) (l'écart est rapporté à la valeur théorique du quantile) sont présentés en fonction du nombre de prélèvements pour des quantiles d'ordre allant de 10% à 90%. On sépare dans les figures les résultats théoriques (à droite) et moyennés sur 1000 simulations (à gauche).

La méthode proposée est d'autant meilleure que l'ordre du quantile s'approche de 50% (la médiane) pour lequel le biais est nul. Le biais relatif est plus faible pour les quantiles d'ordre élevé que pour ceux d'ordre faibles, le dénominateur étant plus élevé. La figure FIG. 5-8, qui présente l'espérance du quantile empirique en fonction du nombre de prélèvement, montre que la médiane est très mal estimée par le quantile empirique sans linéarisation.

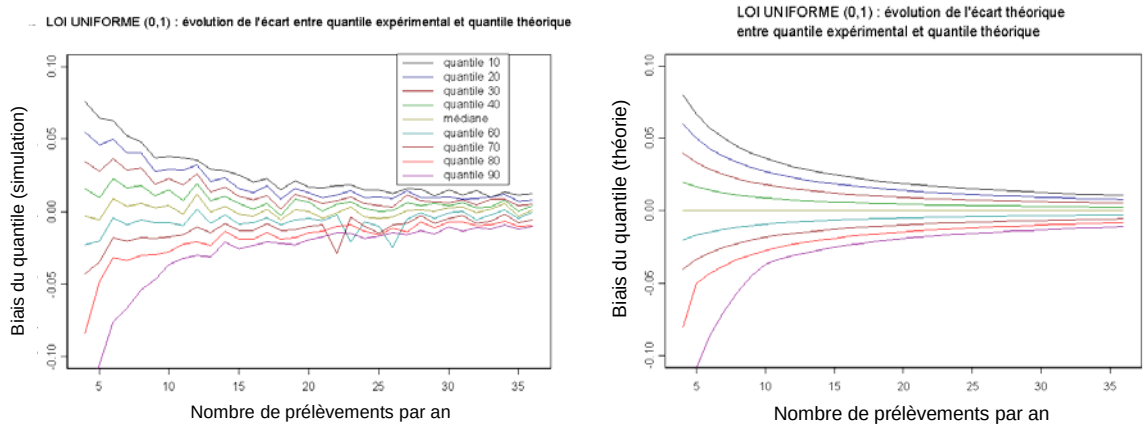


FIG. 5-6 : Loi uniforme : écart entre quantile empirique interpolé linéairement et quantile théorique, obtenu par simulation (à gauche) ou calculé théoriquement (à droite).

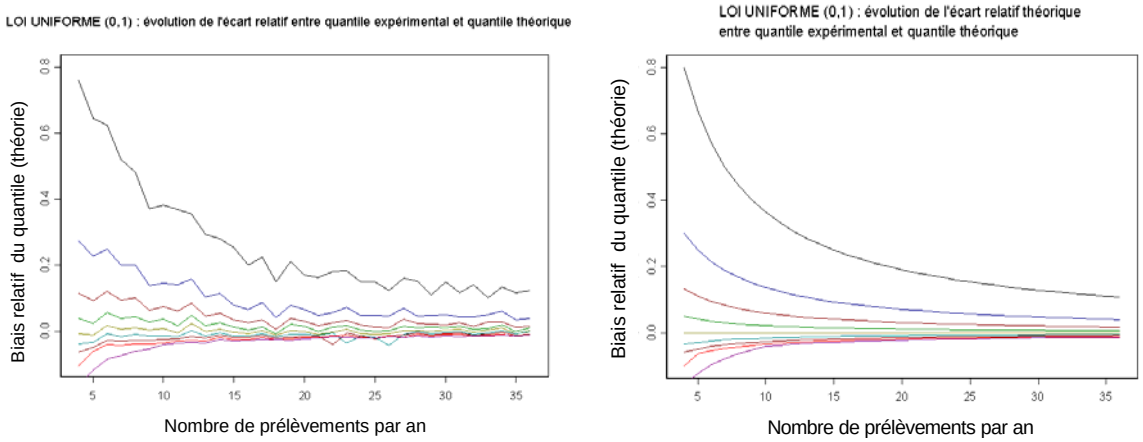


FIG. 5-7 : Loi uniforme : écart relatif entre quantile empirique interpolé linéairement et quantile théorique, obtenu par simulation (à gauche) ou calculé théoriquement (à droite).

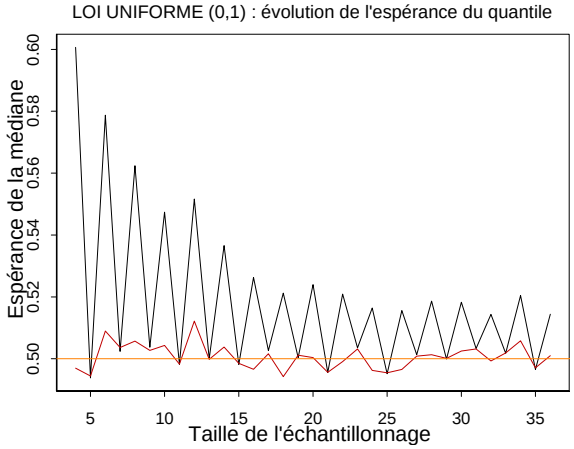


FIG. 5-8: Loi uniforme. Evolution de l'espérance du quantile 50 (médiane) estimé par le quantile empirique et le quantile empirique linéarisé, en fonction de la taille d'échantillonnage.

5.3.4 Problème de la première et de la dernière classe

Pour la classe la plus élevée, l'interpolation est comprise entre la valeur maximale observée et la valeur maximale fixée par la modélisation, pour partie arbitraire. Améliorer le quantile linéarisé serait nécessaire lorsque le quantile recherché est d'ordre supérieur à $\frac{n-1}{n}$, n désignant le nombre de données, ce qui est le cas du quantile 90 estimé avec moins de 10 données. Le quantile recherché peut alors être supérieur au maximum observé, ce qui n'est pas pris en compte si l'on borne l'estimateur par la valeur maximale des mesures. Le problème se pose également pour les quantiles d'ordre faible, inférieur à $\frac{1}{n}$. Cependant, le choix d'un maximum paraît trop subjectif et nous conduit plus à déconseiller des calculs de quantile avec moins de 10 mesures par an ou de grouper les mesures sur trois ans. D'autres méthodes non examinées sont cependant envisageables, comme une extrapolation de la pente. Ces questions se posent de la même manière pour les quantiles d'ordre faible. Dans ce cas, la question d'une extrapolation se pose pour la première classe. Le choix a d'ailleurs été fait ici de ne pas fixer à 0 la valeur minimale des concentrations. Ce choix n'a aucune conséquence par la suite car nous nous intéressons généralement au quantile 90.

5.3.5 Estimateurs des quantiles par anamorphose

Nous examinons maintenant l'utilisation d'une fonction d'anamorphose en la comparant aux estimateurs précédemment présentés. Le choix d'une simple interpolation linéaire résulte de cette étude.

Les biais et écarts types d'estimation de ces méthodes ont été estimés sur des simulations d'échantillons indépendants et de lois connues.

Passage par une fonction d'anamorphose

Estimer un quantile, c'est de façon équivalente estimer la fonction de répartition, autrement dit la loi de probabilité. Dans la plupart des cas, l'histogramme des données ne rappelle aucune loi connue.

Utiliser une anamorphose, c'est écrire la variable aléatoire comme la transformée par une fonction (ou anamorphose) d'une variable aléatoire fixée, de loi gaussienne par exemple (Rivoirard, 1994 ; Lajaunie, 1993 ; Maréchal, 1978). Connaissant le quantile d'une gaussienne, on en déduira par anamorphose le quantile recherché.

Soit donc à chercher Φ strictement croissante (donc bijective) telle que $Z = \phi(Y)$ avec *Y* gaussienne. Le quantile d'ordre p de Z et celui de Y vérifient respectivement $F(z) = p$ et $G(y) = p$ avec G fonction de répartition d'une loi gaussienne centrée réduite.

Le quantile d'ordre p de Z s'obtient donc par le calcul suivant :

$$q(p) = \phi\left(G^{-1}(p)\right)$$

En réalité, on a seulement accès à des réalisations de Z. On peut estimer Φ par l'estimateur suivant, appelé anamorphose empirique :

$$\hat{\phi}(y) = \sum_i Z_{(i)} I_{y \in D_i} \quad \text{avec} \quad D_i = \left[G^{-1}\left(\frac{i-1}{n}\right), G^{-1}\left(\frac{i}{n}\right) \right[$$

Le lissage de l'anamorphose empirique par une fonction continue permet alors d'estimer les quantiles. On propose deux types d'interpolation :

- Une interpolation linéaire de l'anamorphose par une fonction affine par morceaux passant par les points $\left(G^{-1}(p_i), (Z_{(i)} + Z_{(i-1)})/2\right)$ où G est la fonction de répartition d'une loi gaussienne centrée réduite et $p_i = \frac{i-1}{n}$ (voir FIG. 5-9 à gauche).
- Une interpolation à partir du développement hermitien de la fonction d'anamorphose. (voir FIG. 5-9 à droite).

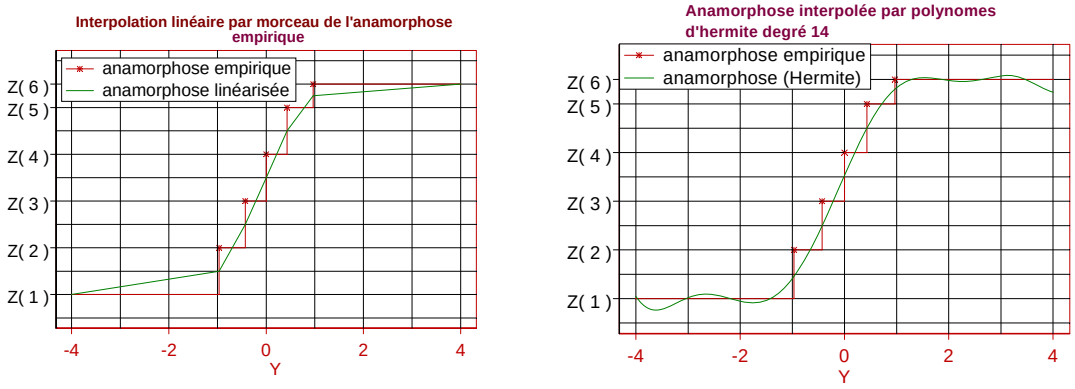


FIG. 5-9 : Exemple d'interpolation de l'anamorphose empirique. A gauche : par interpolation linéaire. A droite : par développement hermitien.

Comparaison des estimateurs pour des variables indépendantes

Les critères de comparaison retenus pour les estimateurs de quantile sont les suivants :

- comparaison du quantile théorique au quantile calculé sur échantillon ;
- évolution de l'erreur quadratique moyenne ;
- évolution de l'intervalle contenant 95% des estimations du quantile ;
- histogramme des estimations du quantile, pour des échantillons de 12 mesures.

Les courbes sont tracées pour différentes lois de probabilité en fonction de la taille de l'échantillon.

Les lois de probabilité considérées sont les suivantes:

- loi normale réduite ;
- loi lognormale, construite comme l'exponentielle de la précédente ;
- loi exponentielle de paramètre 1 (espérance et variance égales à 1);
- loi uniforme entre 0 et 1 ;
- loi gamma de moyenne 1/3 et variance 1/9.

Aucune des méthodes étudiées n'est sans biais. La règle des 90%, comme on l'a vu précédemment, actuellement utilisée par les agences de l'eau montre dans tous les cas un biais important. Pour le quantile 90, cette règle introduit de fortes discontinuités de la valeur estimée, à chaque changement de dizaine de la taille de l'échantillon, rendant discutable une comparaison inter stations ou interannuelle à partir de deux échantillons comportant par exemple 9 et 11 mesures respectivement. De plus cette méthode présente des écarts types d'estimation légèrement supérieurs à ceux des trois autres méthodes. Ces trois nouvelles méthodes sont à peu près équivalentes et, bien que biaisées, permettent une meilleure comparaison des stations entre elles, au moins lorsque le nombre de mesures reste voisin et supérieur à une dizaine. Les méthodes faisant appel à une anamorphose, sont plus compliquées à implémenter : elles nécessitent le calcul de l'anamorphose empirique et sa modélisation, avec par exemple le choix du degré du polynôme.

Lorsque les données sont peu nombreuses, 5 ou 6 mesures par an par exemple, la règle des 90 donne en moyenne une meilleure estimation que les autres méthodes.

L'interpolation linéaire de la fonction de quantile empirique apparaît finalement comme un bon compromis entre l'amélioration de l'estimation et sa faisabilité pratique dans le contexte du SEQ-Eau. Lorsque les histogrammes en deux stations sont très différents, ou pour vérifier le caractère « significatif » d'un écart entre deux stations, il faudrait tenir compte de la précision de l'estimation du quantile. Ce calcul, beaucoup plus complexe que pour la moyenne, nécessite d'introduire des hypothèses sans doute peu réalistes. Cette difficulté pourrait être levée par une méthode de rééchantillonnage, (voir par exemple (Saporta, 1990)) validée en présence de corrélation temporelle.

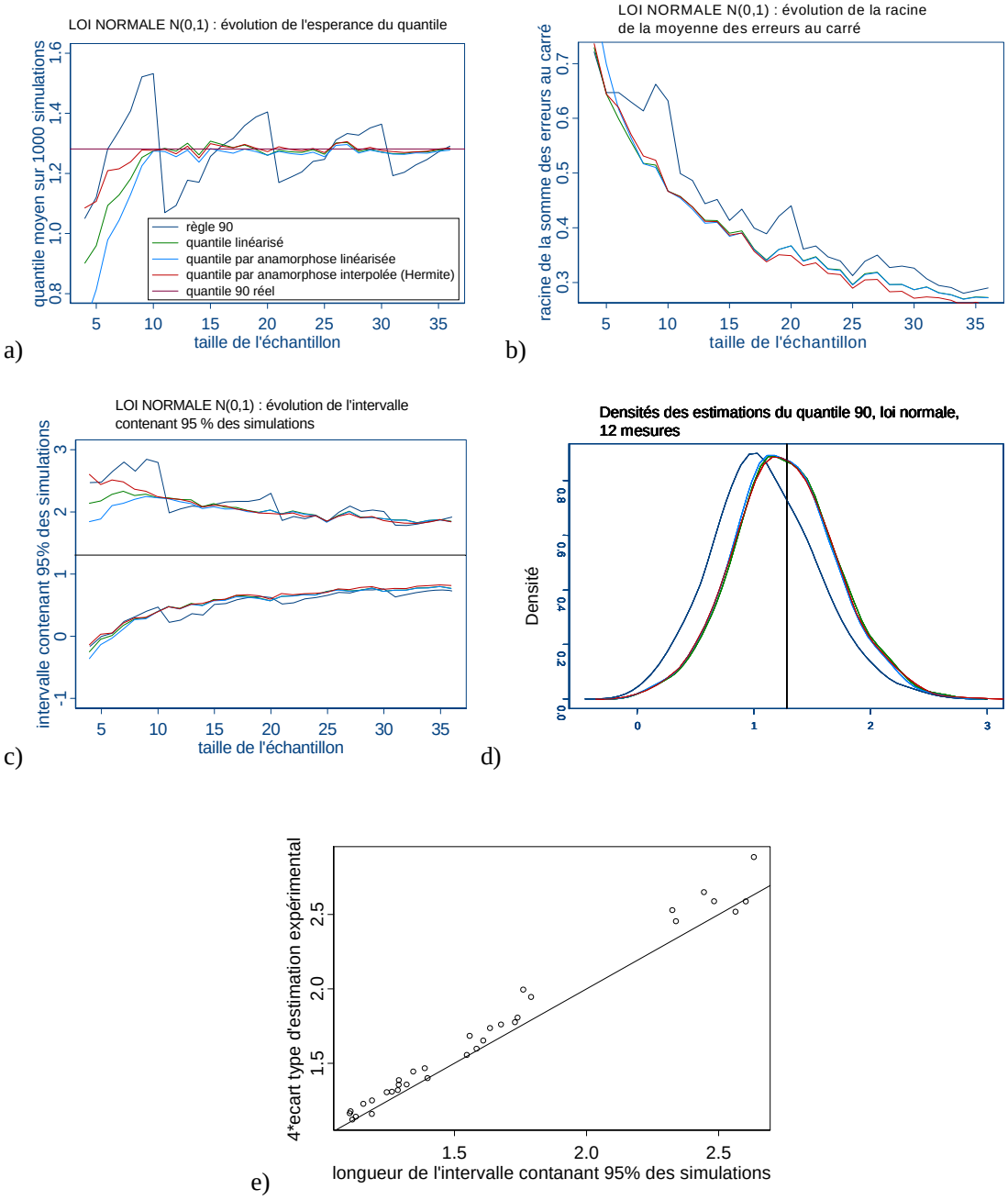


FIG. 5-10 : Evaluation du quantile 90 pour une loi normale ; calcul sur 1000 simulations. a) quatre estimateurs du quantile 90, en fonction du nombre de mesures. b) écart-type d'estimation. c) intervalle de confiance à 95%. d) histogramme des estimations pour un échantillon de 12 mesures. e) nuage de corrélation entre la longueur de l'intervalle de confiance à 95 % et 4 fois l'écart-type d'estimation.

5.3.6 Probabilité de dépassement de seuil

Quand il existe une densité, la fonction de quantile détermine la probabilité de dépassement de seuil. En effet, la fonction de quantile correspond à l'inverse de la fonction de répartition, et la probabilité de dépassement de seuil à son complémentaire.

La probabilité de dépassement du seuil « s » peut être considérée globalement ou localement :

- globalement : pour l'ensemble de la population : tirant une date au hasard et uniformément parmi 365 (ou 366), c'est la probabilité que la valeur observée dépasse « *s* ». C'est la probabilité « a priori », sur l'ensemble de la population,
- localement : à une date « *t* », en tenant compte des mesures effectuées vers cette date (donc conditionnellement aux données situées au voisinage de cette date). Mais cela nécessite un minimum de mesures.

Le second calcul est utile pour la surveillance des concentrations, par exemple lors des périodes de fortes concentrations.

Le premier calcul correspond mieux à la démarche du SEQ-Eau. Une fois la fonction de répartition empirique modélisée, ou ce qui est équivalent, la fonction de quantile empirique, ce calcul est immédiat : il suffit de lire sur la courbe la valeur associée au seuil « *s* ».

La méthode consiste alors :

- à pondérer les données, par les poids de krigeage de la moyenne annuelle, ou par segment d'influence
- à interpoler linéairement par segment la fonction de quantile empirique ;
- à lire sur cette fonction une évaluation de la probabilité au seuil « *s* ».

5.4 Pondération des données pour des variables corrélées

En présence de corrélation temporelle, et en particulier en cas d'échantillonnage préférentiel, les quantiles sont biaisés. Comme pour la moyenne, une pondération est nécessaire pour corriger ce biais.

Dans le cas des nitrates, ajouter des valeurs hivernales doit améliorer la précision du quantile 90 sans pour autant l'augmenter artificiellement.

Nous proposons de pondérer les mesures par les poids de krigeage de la moyenne annuelle.

Une fois les poids calculés, comment les prendre en compte dans la fonction de répartition ou de quantile ? En l'absence de pondération, la fonction de répartition empirique sur un échantillon de taille *n* se calculait ainsi :

$$F_n(Z) = \frac{1}{n}\sum_i I_{(Z(x_i)<z)} \quad \text{où les mesures sont les } Z(x_i) \quad (5\text{-}7)$$

Si on prend en compte les *n* poids $p_1, p_2, ..., p_n$, on a alors :

$$F_n(Z) = \sum_i p_{(i)} I_{(Z(x_i)<z)} \quad (5\text{-}8)$$

où les mesures sont les $Z(x_i)$

où $p_{(i)}$ est la somme des poids de krigeage des (i) premières mesures ordonnées dans l'ordre croissant.

Il faut cependant vérifier la positivité des poids de krigeage, sinon on se trouve dans un cas de fonction de répartition non monotone ce qui est incohérent et n'a aucun sens physique. Nous reviendrons par la suite sur les rares cas de poids négatifs.

Le quantile empirique est alors défini par :

$$q_p = Z_{(i)} \text{ pour } p_{(i-1)} \leq p < p_{(i)}$$

Ce quantile correspond à une fonction en escalier, mais la pondération peut aussi être associée à l'interpolation linéaire du quantile On a alors :

$$Q(x) = \left[\sum_i \left(\frac{Z_{(i+1)} - Z_{(i-1)}}{2 * (p_{(i+1)} - p_{(i)})} \right) (x - p_{(i)}) + \frac{(Z_{(i)} + Z_{(i-1)})}{2} \right] I_{x \in]p_{(i)}, p_{(i+1)}]}$$

L'influence de la pondération sur la fonction de quantile ainsi pondérée est présentée figure FIG. 5-11. Cet exemple est construit de la façon suivante : la chronique des concentrations en nitrates pour la station « quai du roi » a été complétée par simulation. Les mesures sont donc journalières et la fonction de quantile connue. De cette « chronique » est extrait un échantillon avec des mesures hivernales plus nombreuses (18 mesures hivernales, 6 mesures estivales). On compare alors la fonction de quantile empirique suivante à celle de référence dans les trois cas suivants:

- Absence de pondération
- Pondération par segment d'influence
- Pondération par krigeage de la moyenne annuelle

Les poids de krigeage ou par segment d'influence sont sensiblement égaux sur cet exemple.

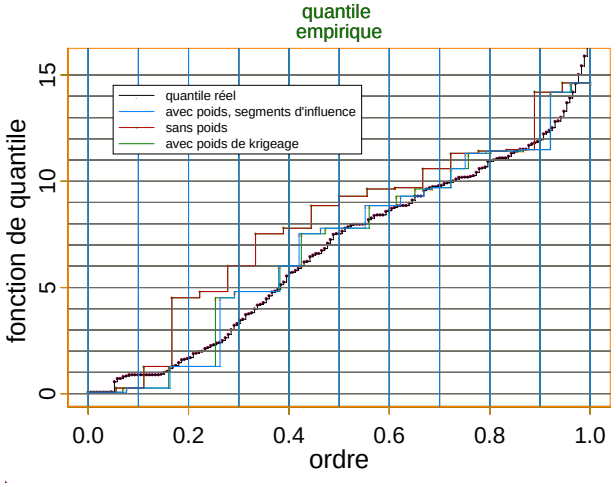


FIG. 5-11 : Prise en compte des poids de krigeage sur l'estimation des quantiles

Les courbes des quantiles obtenues avec une pondération approchent mieux la courbe réelle en particulier pour des quantiles proches de la médiane. Notons que la pondération des mesures et la linéarisation du quantile peuvent avoir des effets opposés sur l'estimation du quantile.

5.5 Exemples

Deux effets sont à étudier sur les estimations du quantile : la pondération et la linéarisation. En pratique, une très grande diversité de situations se présente sur le réseau Loire Bretagne. D'une part, l'échantillonnage est préférentiel en été ou en hiver. Ensuite, les variations saisonnières diffèrent selon la substance. L'influence de la pondération est donc différente selon le polluant et sa fréquence d'échantillonnage. De plus, si les données sont peu nombreuses (moins de 10 mesures) alors pour le quantile 90, la pondération n'intervient quasiment plus, car on reste systématiquement dans la dernière classe de la fonction escalier. C'est d'ailleurs pour cette raison que la règle des 90% est parfois appliquée par les agences sur trois années de mesures.

Ces différents cas sont examinés ici :

Cas d'un échantillonnage préférentiel en été : 6 mesures annuelles

Quatre estimateurs du quantile 90 sont présentés FIG. 5-12 sur des données réelles de stations à six mesures annuelles préférentielles en été :

- Le quantile empirique
- Le quantile empirique interpolé linéairement
- Le quantile empirique pondéré par les poids de krigeage de la moyenne
- Le quantile empirique pondéré par les poids de krigeage de la moyenne et interpolé linéairement

Deux paramètres, nitrates et orthophosphates, sont examinés. Les indicateurs en nitrates sont étudiés sur la station 200595 (FIG. 5-12, haut), qui présente des variations saisonnières avec des valeurs fortes en hiver et faibles en été. Les indicateurs en orthophosphates sont étudiés sur la station 171550 (FIG. 5-12, bas), qui à l'inverse présente des valeurs fortes en été et faibles en hiver. Les quantiles sont calculés sur une année (FIG. 5-12, à gauche), puis sur trois années (FIG. 5-12, à droite).

Sur le bassin Loire Bretagne, la fréquence de 6 mesures par an correspond à un surréchantillonnage en été, donc à un surplus de valeurs faibles pour les nitrates et un surplus de valeurs fortes pour les orthophosphates. L'influence des poids est donc inverse dans les deux cas.

Avec 6 mesures par an, avec ou sans pondération, le quantile 90 revient à travailler sur la dernière classe de l'histogramme expérimental et la pondération n'y change presque rien alors que l'interpolation linéaire modifie la modélisation de l'histogramme pour toutes les classes y compris la dernière. Pour le quantile 90, c'est donc l'interpolation linéaire qui en pratique a le plus d'influence. Attention cependant, l'interpolation linéaire associée à la pondération ne coïncide pas avec l'interpolation linéaire seule, les poids lorsqu'ils sont associés à l'interpolation ont donc une influence et ne doivent pas être écartés des estimations.

Considérons l'exemple de 6 mesures pour lequel on dispose de 4 mesures faibles en été, et deux mesures fortes en hiver (toutes supérieures à celle d'été). Si on affecte un poids de 1/8 aux valeurs d'été qui sont plus nombreuses, et un poids de 1/4 aux valeurs d'hiver, un rapide calcul montre que le quantile 90 reste estimé par la valeur maximale. Sans pondération la dernière classe correspond à l'intervalle [0.83, 1]. Avec pondération, il correspond à l'intervalle [0.75, 1].

Un calcul étendu à trois années n'améliore pas beaucoup les calculs, les mesures étant concentrées en été.

On peut voir sur les graphiques que la pondération a un effet important lorsqu'on ne s'intéresse pas à des quantiles « extrêmes » correspondant à la première ou dernière classe de la fonction de répartition.

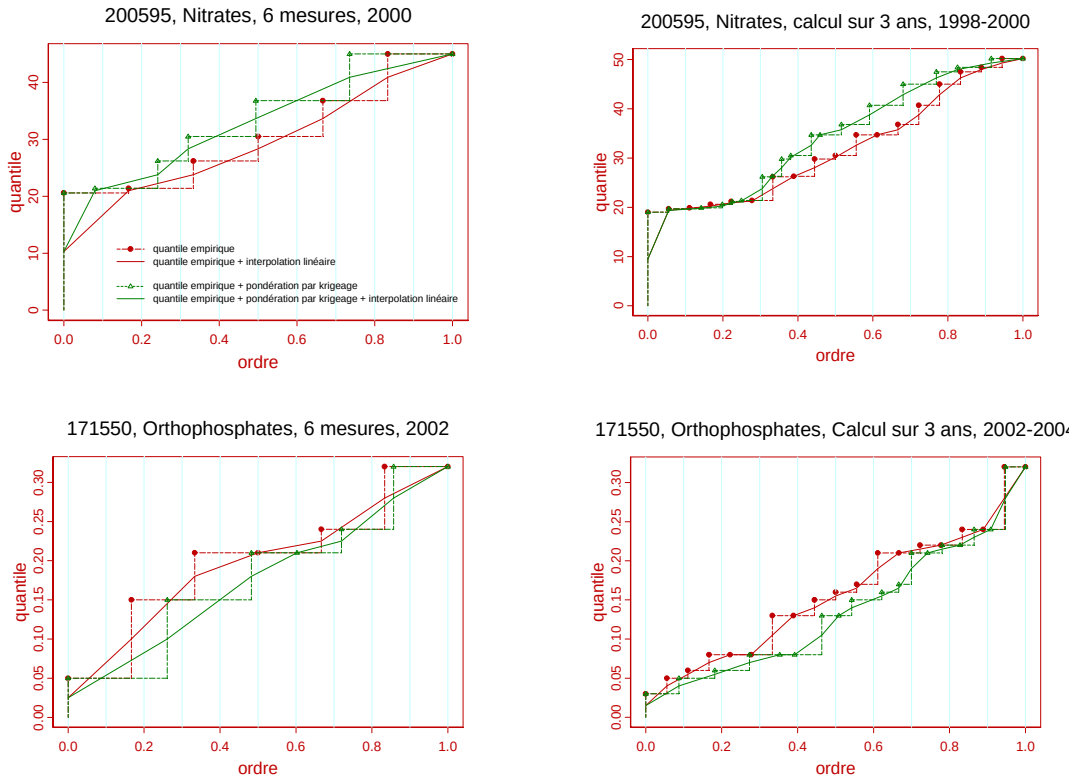


FIG. 5-12 : Quantile empirique avec et sans pondération par krigeage, avec et sans interpolation linéaire. Applications à des stations avec 6 mesures annuelles préférentielles en été. En haut : nitrates. En bas : orthophosphates. A gauche : quantile estimé sur une année. A droite : quantile estimé sur trois années.

Cas d'un échantillonnage régulier : 12 mesures annuelles

Avec 12 mesures annuelles régulièrement réparties, alors les poids de krigeage ne modifient pas la fonction de quantile (voir FIG. 5-13). Seule l'interpolation linéaire produit un changement dans les estimations. En revanche, pour un échantillonnage irrégulier, la fonction de quantile est légèrement modifiée par la pondération.

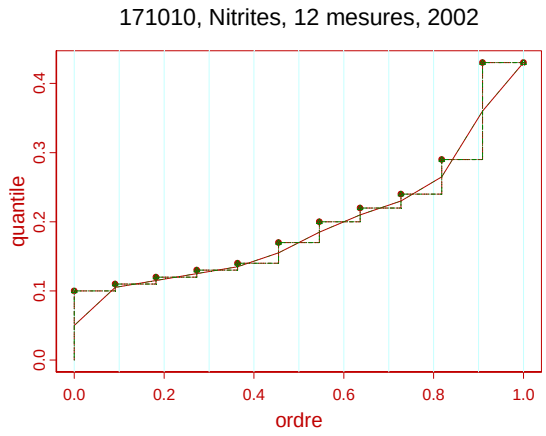


FIG. 5-13 : Fonction de quantile pour la station 171010 en nitrites. 12 mesures régulières annuelles.

Cas d'un échantillonnage préférentiel : 18 mesures annuelles

Dans le cas d'un échantillonnage de 18 mesures annuelles, on a comme précédemment deux possibilités :

- Soit les mesures sont fortes en hiver et les quantiles empiriques ou linéarisés sont surestimés. La pondération corrige alors ce biais. C'est le cas pour les nitrates de la station 56000 en 1995 (FIG. 5-14 à gauche). La pondération modifie l'estimation du quantile 90, même en l'absence de linéarisation.
- Soit les mesures sont faibles en hiver et les quantiles empiriques ou linéarisés sont sous estimés. C'est le cas pour les nitrites en 1995, sur cette même station 56000.

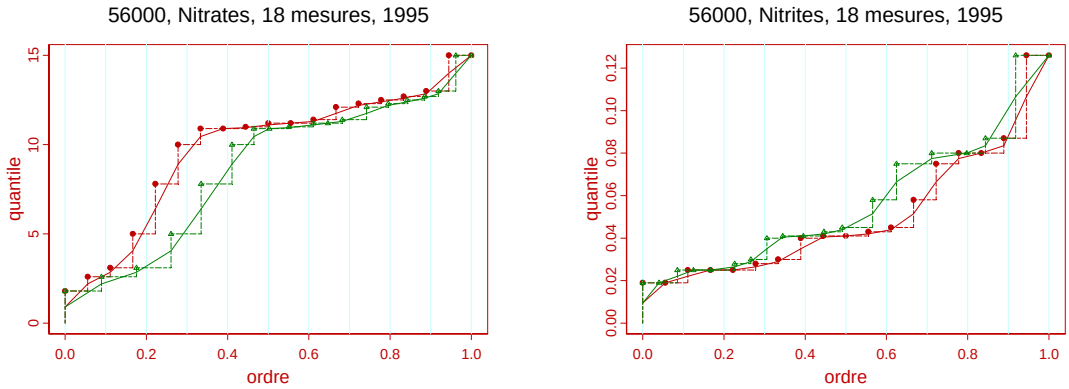


FIG. 5-14 : Estimation de la fonction de quantile par le quantile empirique linéarisé ou non, pondéré ou non. Station 56000 en 1995. A gauche : nitrates. A droite : nitrites.

5.6 Conclusion

Pour l'estimation des quantiles, des exemples de calcul sur des données indépendantes de loi connue montrent que l'interpolation du quantile empirique améliore en moyenne les estimations pour une fréquence supérieure à 10 mesures par an. La pondération corrige les biais d'un échantillonnage préférentiel seulement si les mesures à cette période sont fortes. Elle a parfois un effet opposé à celui de l'interpolation linéaire. Ajouter des valeurs faibles modifie en revanche rarement les estimations du quantile 90. Le calcul théorique d'incertitude sur l'estimation des quantiles reste compliqué.

Chapitre 6

Validation des méthodes

Le krigeage fournit en théorie le meilleur estimateur linéaire de la moyenne annuelle au sens de la minimisation de la variance d'estimation, et on a vu que l'interpolation linéaire corrige le biais du quantile empirique pour un nombre de prélèvements supérieur à 10 par an. Cependant, il reste à valider les méthodes pour des concentrations dont la loi est inconnue et pour lesquelles les variogrammes sont plus ou moins bien adaptés. La simulation est un moyen de valider les méthodes. Sur les simulations pour lesquelles les indicateurs sont supposés connus, nous testerons la qualité des estimateurs pour des échantillonnages réguliers, irréguliers et préférentiels, conformément à ceux utilisés par les agences.

6.1 Comment valider les méthodes ?

Le krigeage est le meilleur estimateur linéaire, puisqu'il fournit une estimation sans biais et de variance d'estimation minimale. Mais le résultat fourni par le krigeage est optimal lorsque le modèle géostatistique représente correctement la réalité. Qu'en est-il en pratique ? De plus, la pondération par krigeage, optimale pour la moyenne, le reste-t-elle pour le quantile ?

La pertinence d'un modèle géostatistique est souvent analysée par « validation croisée ». Cela consiste à effacer certaines données pour en effectuer l'estimation à partir des données conservées. L'erreur d'estimation et l'erreur quadratique « expérimentale » sont alors comparées aux prévisions du modèle géostatistique suivant différents critères. Mais si l'on revient à la question pratique des mesures par station, le problème se pose un peu différemment, puisque la valeur recherchée est la moyenne ou le quantile annuel, jamais mesurés.

En présence de corrélation temporelle, il n'existe pas d'étude théorique sur le quantile. Nous allons donc examiner l'influence de la pondération à l'aide de simulations de Fonctions Aléatoires de loi (spatiale) normales, présentant une corrélation temporelle. Différents types de simulations peuvent être étudiés:

- Des simulations non conditionnelles de fonctions aléatoires de loi de probabilité et de variogramme fixés
- Des simulations conditionnées par des mesures réelles de concentrations, les chroniques simulées étant ainsi proches de la réalité.

Plusieurs échantillonnages peuvent ensuite être testés :

- Des échantillons réguliers. Dans ce cas pour la moyenne comme pour le quantile, les estimations vont être très proches quelle que soit la méthode choisie. Seuls les intervalles de confiance associés vont différer.
- Des échantillons irréguliers, qui vont permettre de valider l'intérêt de la pondération par krigeage ou par segments d'influence. Les échantillons peuvent être tirés aléatoirement ou avec des mesures resserrées sur une période, en se référant aux différents types d'échantillonnages du paragraphe 2.3.4 observés sur les stations.

Les stratégies d'échantillonnage testées ne sont pas toujours les mêmes selon le type de simulation. Nous rappellerons donc à chaque début de paragraphe les échantillons considérés.

Ces tests de validation sont présentés pour différentes stations et lois dans les rapports (Bernard-Michel and de Fouquet, 2003 ; Bernard-Michel and de Fouquet, 2004). Nous récapitulons ici les principaux exemples.

6.2 Simulations non conditionnelles

Les simulations non conditionnelles permettent de tester les méthodes sur des modèles généraux, l'objectif étant principalement d'évaluer la réduction de biais.

Différents types d'échantillonnage sont envisageables :

- Des échantillons préférentiels avec renforcement en hiver ou en été : dans le cas de simulations non conditionnelles, cette démarche paraît peu adaptée car d'une part, les fonctions aléatoires simulées ne seront pas nécessairement choisies périodiques, et si elles le sont, les schémas en cosinus ont une phase aléatoire. L'échantillonnage préférentiel sera donc étudié sur des simulations conditionnelles.
- Des échantillons réguliers
- Des échantillons irréguliers. Soit on extrait des échantillons de taille n aléatoirement entre 1 et 365 jours. Soit de façon plus réaliste, on extrait des échantillons par un tirage aléatoire stratifié, c'est-à-dire qu'on divise l'année en n intervalle dans lesquels on tire aléatoirement une valeur.

6.2.1 Les simulations

1000 chroniques annuelles (365 valeurs) sont simulées, de loi normale, avec pour covariance temporelle :

- Soit un schéma cosinus de période 365.25 et d'amplitude 20 (voir (Bernard-Michel and de Fouquet, 2004) pour les amplitudes 2, 8 et 16)
- Soit un schéma sphérique de portées 25 et de palier 1 (voir (Bernard-Michel and de Fouquet, 2004) pour les paliers 50, 75 et 100)

D'autres paramètres étudiés pour ces covariances sont présentés dans (Bernard-Michel and de Fouquet, 2004).

Pour les variogrammes en cosinus (mg^2/L^2), les simulations sont construites par décomposition spectrale.

$Z(t) = A\cos\left(\frac{2\pi t}{T}\right) + B\sin\left(\frac{2\pi t}{T}\right)$, où A et B sont deux variables aléatoires indépendantes de loi normale centrée et de variance σ^2, $Z(t)$ suit alors une loi normale et admet comme covariance un schéma cosinus de période T et d'amplitude $2\sigma^2$ non ergodique : $C(Z(t), Z(t+h)) = \sigma^2 \cos\frac{2\pi h}{T}$.

Pour le schéma sphérique, les simulations sont effectuées par la décomposition matricielle de Cholesky.

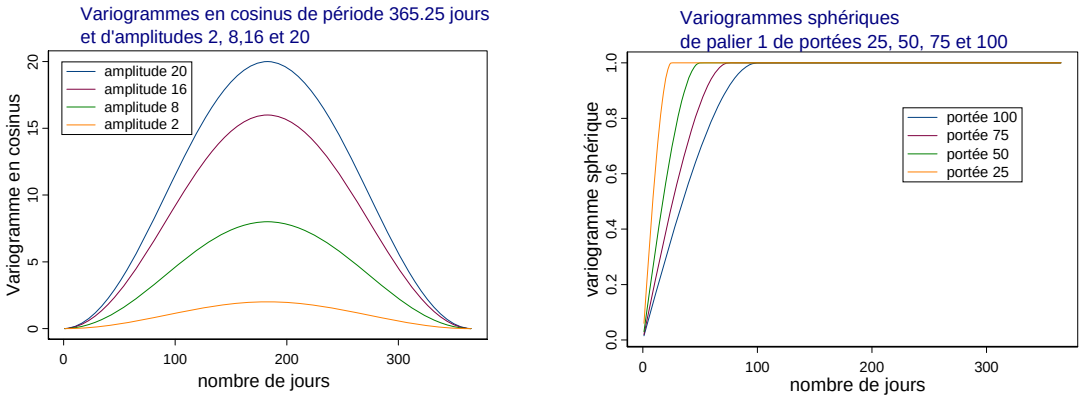


FIG. 6-1 : Variogrammes théoriques. A gauche, schéma cosinus d'amplitude 2, 8, 16, 20 et de période 365 jours. A droite, schéma sphérique de portée 25, 50, 75, 100 et de palier 1.

Les simulations ont été validées par plusieurs calculs :

- le variogramme moyen sur 1000 simulations coïncide avec le variogramme théorique simulé
- La loi suivie par les mesures est approximativement normale pour chaque date de prélèvement (exemple d'histogramme et de fonction de répartition pour le jour 180

On a ainsi simulé une fonction aléatoire. Pour une date fixée, les 1000 valeurs simulées sont les réalisations d'une même variable gaussienne. Ce modèle de covariance présente des propriétés particulières, en particulier il n'est pas ergodique : pour certaines grandeurs, la moyenne spatiale sur une réalisation ne converge pas vers l'espérance mathématique correspondante dans le modèle. Par ailleurs, le champ (une année) n'est pas toujours très grand vis-à-vis de la portée, pour le schéma sphérique. Enfin, l'histogramme d'une chronique annuelle (donc sur une réalisation et à champ limité) n'est pas gaussien (exemple FIG. 6-2).

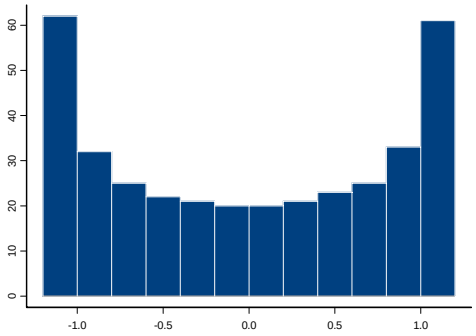


FIG. 6-2 : Histogramme pour une réalisation sur un an : loi normale et covariance en cosinus.

Pour l'un des calculs, nous avons effectué dix fois les 1000 simulations afin de vérifier si les conclusions restaient identiques, ce qui était effectivement le cas.

6.2.2 Echantillonnage, validation

De chaque simulation, on extrait un échantillon sur lesquels on va estimer les indicateurs afin d'évaluer le biais et la précision des méthodes. Deux types d'échantillonnage sont examinés :

- Echantillons réguliers de taille *n* : l'année de 365 jours est divisée en *n* segments, les *n* dates se situant au milieu de chacun de ses segments.
- Echantillons irréguliers de taille *n* aléatoires stratifiés: l'année est divisée en *n* segments et on tire aléatoirement, selon une loi uniforme, une date sur chacun des segments. On retient alors les mesures associées à ces dates.

Validation

Nous étudions alors le biais et la précision des estimations du quantile 90. Les méthodes d'estimation comparées sont :

- La règle des 90%
- La règle des 90% pondérée par les poids de segments d'influence
- La règle des 90 pondérée par les poids de krigeage de la moyenne annuelle
- La règle des 90% linéarisée
- La règle des 90% linéarisée, pondérée par les poids de segments d'influence
- La règle des 90% linéarisée, pondérée par les poids de krigeage

Nous avons estimé le quantile 90 par chacune de ces méthodes pour des échantillons réguliers ou irréguliers de taille variant de 6 mesures par an à 36 mesures par an. Les critères retenus sont les suivants :

- ***Quantile moyen sur 1000 simulations*** :

  C'est le quantile moyen $q_{moyen,n}$ sur 1000 simulations. Nous pouvons alors le comparer au quantile réel q_{moyen} qui correspond à la moyenne des 1000 quantiles 90 calculés sur chacune des simulations de 365 jours.

  Si on note $q^i_{90,n}$ le quantile d'ordre alpha, calculé sur la simulation i avec un échantillon de taille n. Alors :

$$q_{moyen} = \frac{1}{1000}\sum_{i=1}^{1000} q^i_{90,365} \text{ et } q_{moyen,n} = \frac{1}{1000}\sum_{i=1}^{1000} q^i_{90,n}$$

- ***Ecart type résiduel moyen*** **:**

  Il est calculé comme la racine de la moyenne des erreurs quadratiques sur 1000 simulations. Il permet d'approcher l'écart type d'estimation du quantile.

$$\sigma_n = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(q^i_{90,n} - q^i_{90,365}\right)^2}$$

- ***L'intervalle de confiance à 95% expérimental*** :

C'est l'intervalle contenant 95% des estimations. Il est déterminé par le quantile d'ordre 0.025 et celui d'ordre 0.975 de ces estimations $q^i_{90,n}, i \in \{1,...1000\}$. Il permet d'évaluer l'intervalle de confiance non accessible par un calcul théorique.

6.2.3 Les résultats

Covariance périodique

Les résultats présentés FIG. 6-3 montrent que les poids de segments d'influence et de krigeage ne changent pas les estimations pour des échantillons réguliers, sauf à proximité des points de discontinuités de la fonction de quantile. Pour les échantillons irréguliers, la pondération par segment d'influence a un effet plus marqué que celle par krigeage et réduit mieux le biais. Pour le quantile empirique interpolé linéairement, avec ou sans pondération et quelle que soit la méthode de pondération choisie, les estimations sont similaires en moyenne.

Avec moins de 10 mesures par an, l'interpolation linéaire n'améliore pas les estimations, elle peut même les détériorer. Mais au delà de 10 mesures par an, le biais se réduit très nettement. Avec 10 mesures, l'écart entre le quantile réel et le quantile moyen est inférieur à 2%. L'interpolation linéaire a aussi pour avantage de « lisser » le biais, ce qui rend possible la comparaison entre stations d'effectifs différents. Enfin, l'écart type d'estimation expérimental est plus faible pour l'ensemble des quantiles interpolés linéairement. L'intervalle de confiance expérimental à 95%, reste toutefois très important pour toutes les méthodes allant jusqu'à $\pm 50\%$.

Covariance sphérique

Les résultats présentés FIG. 6-4 conduisent à des résultats similaires au cosinus.

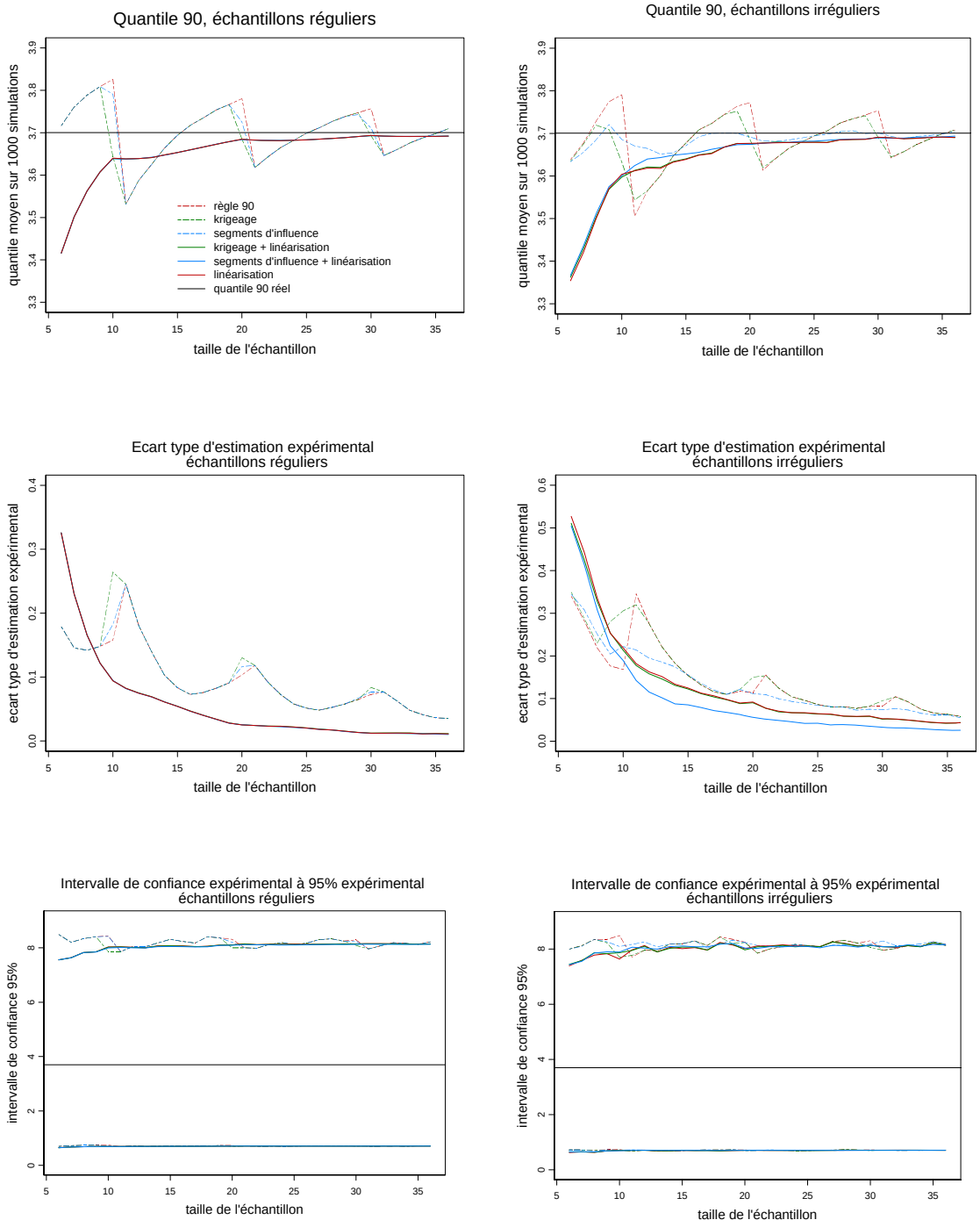


FIG. 6-3 : Variogramme périodique d'amplitude 20 et de période 365.25. A gauche, échantillons réguliers. A droite, échantillons irréguliers. De haut en bas : quantile 90 moyen sur 1000 simulations, Ecart type d'estimation expérimental, Intervalle de confiance expérimental à 95%.

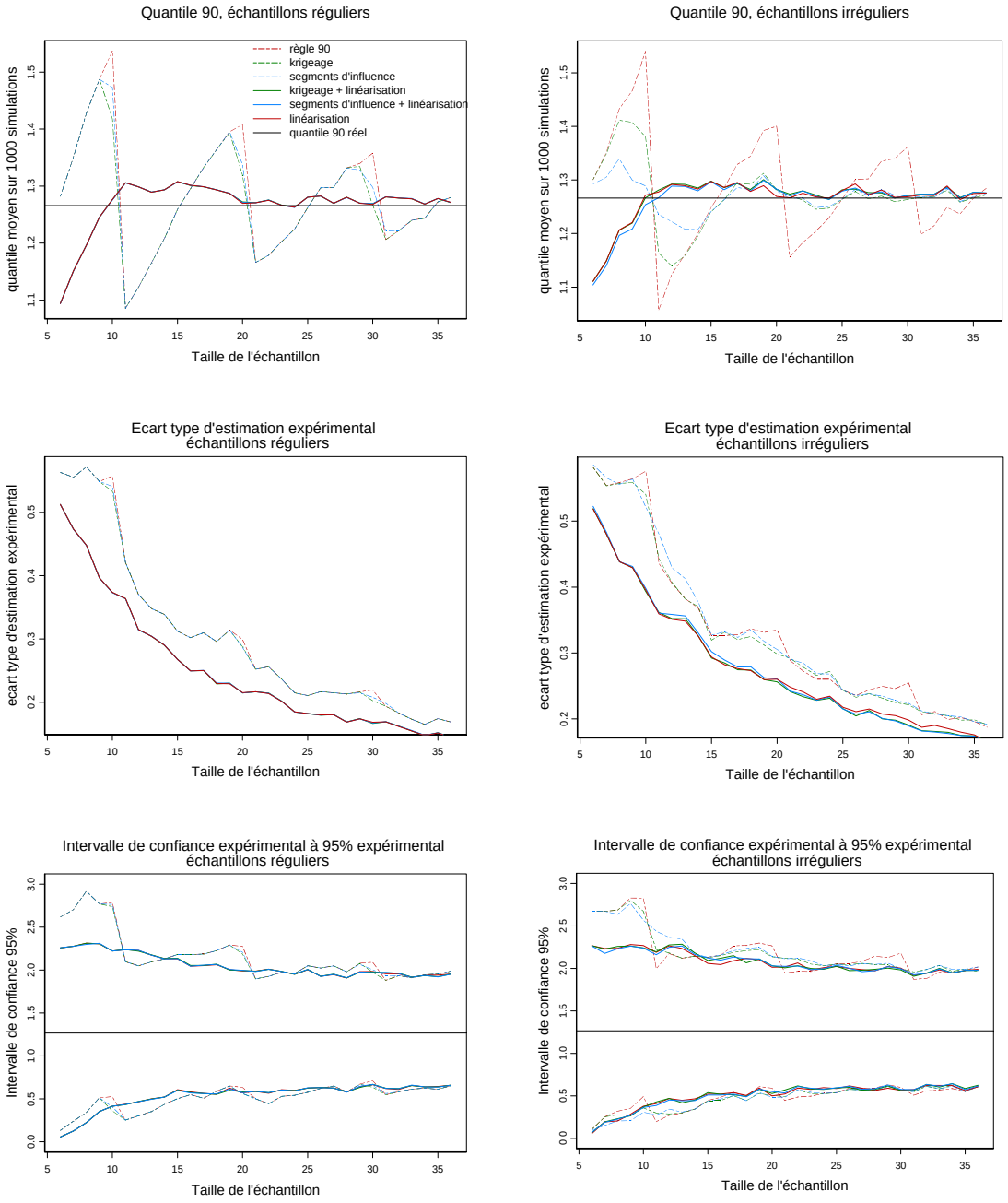


FIG. 6-4 : Variogramme sphérique de portée 25 et de palier 1. A gauche, échantillons réguliers. A droite, échantillons irréguliers. De haut en bas : quantile 90 moyen sur 1000 simulations, Ecart type d'estimation expérimental, Intervalle de confiance expérimental à 95%.

6.3 Comparaison de différents échantillonnages sur des chroniques reconstituées

Les tests précédents effectués sur des simulations de lois connues avec des corrélations choisies permettent de vérifier la pertinence des méthodes proposées sur des cas généraux. Mais ces cas étudiés ne sont pas toujours très représentatifs du contexte « qualité de l'eau ». Par exemple, les concentrations en nitrates, nitrites et orthophosphates présentent des histogrammes très différents des lois usuelles. Afin de tester les différentes méthodes sur des simulations proches de la réalité, on utilise des simulations calées sur des stations bien informées. L'idée est la suivante : on choisit pour chacun des paramètres les stations ayant fait l'objet d'un suivi très précis avec pratiquement une mesure par jour ou tous les deux jours. On complète alors la chronique par simulation séquentielle gaussienne (Lantuejoul, 2001) et on considère cette série complète de 365 jours comme la réalité. Le quantile 90 et la moyenne annuelle sur la chronique ainsi reconstituée sont les indicateurs que l'on cherche à estimer.

N échantillons de même taille *n* fixée sont extraits afin d'étudier les biais expérimental et les écarts types résiduels.

Les échantillonnages considérés ici sont ceux classiquement rencontrés dans le domaine de l'eau:

- 6 mesures par an, avec 2 mesures en hiver (mars, décembre) et 4 en été (juin, juillet, aout, septembre)
- 9 mesures par an, avec 6 en été (mai, juin, juillet, aout, septembre, octobre) et 3 en hiver (janvier, mars, décembre)
- 12 mesures par an (une mesure par mois)
- 15 mesures par an avec 9 mesures en hiver et 6 mesures en été (une mesure par mois sauf en janvier mars et mai pour lesquels on dispose de deux mesures par mois)
- 18 mesures par an préférentielles en hiver, avec 12 mesures en hiver (2 mesures par mois en janvier, février, mars, avril, novembre et décembre) et 6 en été (une mesure par mois en mai, juin juillet, aout, septembre, octobre)
- 18 mesures par an préférentielles en été, avec 12 mesures en été (3 mesures par mois en juin, juillet, aout et septembre) et 6 en hiver (3 mesures par mois en mars et décembre). Ce type d'échantillonnage ne se rencontre pas dans les données, mais il permet de tester l'efficacité des méthodes lorsque les indicateurs sont calculés sur 3 ans par manque d'effectif (6 mesures par an, sachant que dans ce cas la majorité des mesures sont estivales)

Ces échantillons sont obtenus par un tirage aléatoire stratifié. Par exemple pour 15 mesures par an, on tire uniformément une valeur par quinzaine en janvier, mars et mai, et une pour les autres mois.

Pour les différents échantillonnages, on calcule :

- Les moyennes annuelles estimées par statistique, krigeage et pondération par segments d'influence.
- Les écarts types d'estimation théoriques associés
- Les quantiles 90 annuels estimés par le quantile empirique linéarisé ou non pour différentes pondérations : poids de krigeage et segments d'influence.

On en déduit les biais expérimentaux et les écarts types résiduels. Dans le cas de l'estimation de la moyenne, ces écarts types sont comparés aux écarts types théoriques moyens annoncés par les méthodes. Ces résultats sont accompagnés de l'intervalle de confiance expérimental contenant 95% des estimations.

Sur le bassin Loire Bretagne, peu de stations sont bien informées, il est donc impossible d'appliquer cette démarche à l'ensemble des stations du réseau et d'en déduire des statistiques générales sur les biais et incertitudes des estimations. Deux exemples sont présentés ici : les concentrations en nitrates à la station 50500 en 1985 pour laquelle on dispose de 172 mesures et celles en nitrites à la station 57800 en 1996 pour laquelle on dispose de 64 mesures. Ces deux chroniques annuelles ont été complétées par simulation séquentielles après avoir écartées les valeurs aberrantes éventuelles (voir FIG. 6-5). Ces deux chroniques présentent une corrélation temporelle et des variations saisonnières opposées. Les variogrammes (FIG. 6-5 en bas) pour ces chroniques ont été modélisés de façon semi-automatique à partir de 10 années de mesures, dans la perspective d'une application systématique des méthodes. Une étude plus approfondie de la station permettrait une meilleure modélisation du variogramme et probablement une amélioration des estimations et écarts type d'estimation.

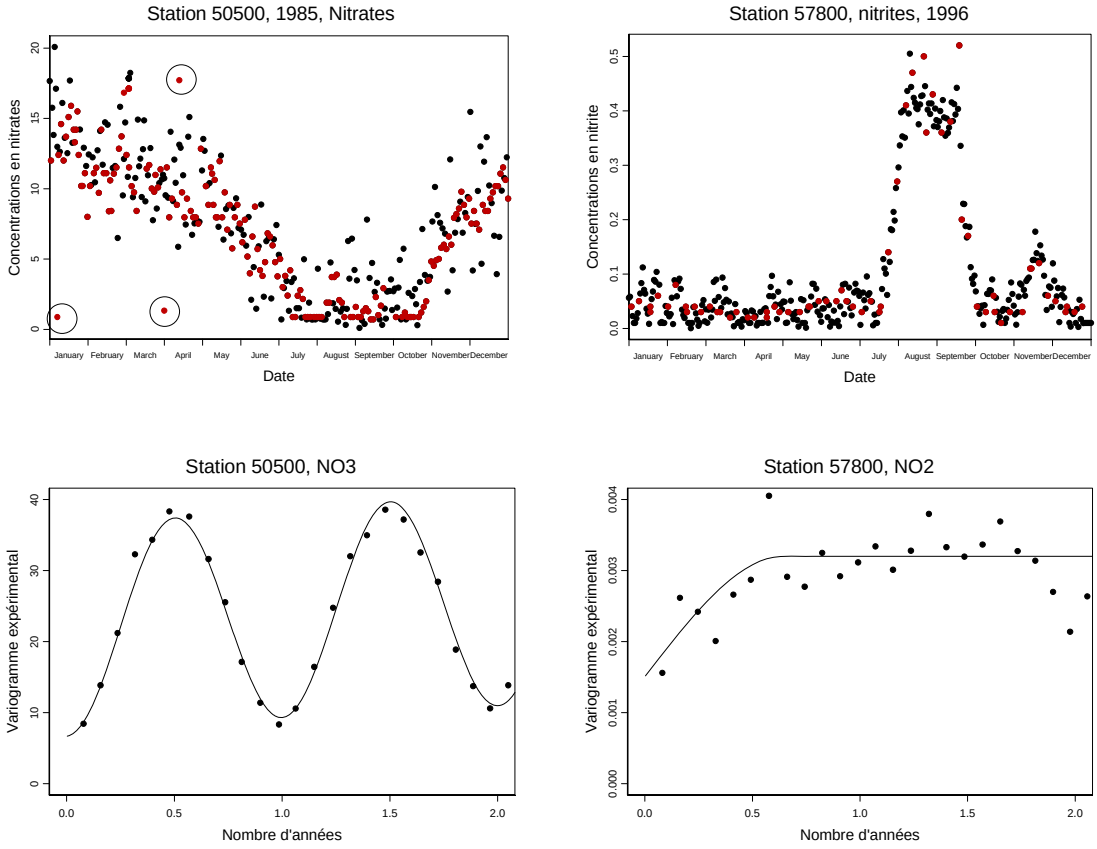


FIG. 6-5 : En haut : à gauche : station 50500, concentration en nitrates en 1985 ; à droite : station 57800, concentrations en nitrites en 1996. En rouge: valeurs mesurées. En noir : valeurs simulées. Les valeurs entourées, considérées comme aberrantes ont été ôtées. En bas : les variogrammes expérimentaux et modélisés, calculés avec des mesures entre 1985 et 2005.

Choix du nombre d'échantillons à simuler

Il y a approximativement 30^{12} possibilités de choisir un échantillon de 12 mesures par an avec une mesure par mois. Nous n'allons pas simuler tous les échantillons possibles pour chaque stratégie

d'échantillonnage. La question est plutôt de savoir à partir de combien de tirages, les moyennes et les écarts types d'estimations ainsi que les écarts types de dispersion commencent à converger. Nous avons étudié cette convergence sur la station 50500 pour les nitrates en procédant à 10000 tirages aléatoires d'échantillons. A partir de 3000 simulations, les moyennes des estimations sont stabilisées et permettent d'étudier efficacement les biais des méthodes et les intervalles de confiance. Pour les stations suivantes, nous effectuons 3000 tirages.

Estimation de la moyenne annuelle

Pour une station étudiée et un type d'échantillonnage considéré, les estimations sont résumées par un tableau. Des exemples sont présentés TAB. 6-1, TAB. 6-2 et TAB. 6-3.
L'ensemble de ces résultats est synthétisé par les graphiques de la figure FIG. 6-7 qui présentent pour chaque type d'échantillonnage les moyennes sur 3000 échantillons aléatoires des estimations par trois méthodes (statistique classique, krigeage et segments d'influence) ainsi que l'intervalle de confiance à 95%, égal à $\pm$ deux fois l'écart type d'estimation sous hypothèse de normalité, qui est effectivement vérifiée (FIG. 6-6)

Le biais est calculé expérimentalement comme la moyenne des erreurs d'estimation des 3000 échantillons. Il est présenté TAB. 6-4 en pourcentage relatif à la moyenne de référence pour chaque type d'échantillonnage.
Les estimations des moyennes annuelles en nitrates et nitrites présentent des biais inverses plus ou moins importants en fonction de la méthode choisie et de l'échantillonnage.

Dans tous les cas, la pondération par krigeage et segments d'influence corrige le biais des estimations pour un échantillonnage préférentiel en été ou en hiver. Dans le cas d'un échantillonnage préférentiel en été, les estimations moyennes restent biaisées, ce qui est une conséquence d'un échantillonnage mal réparti pour lequel 6 mois de l'année ne sont jamais échantillonnés. Passer d'un calcul sur un an (6 mesures) à un calcul sur 3 ans (on estime la moyenne tri-annuelle avec 18 mesures) réduit l'intervalle de confiance prédit mais l'estimation reste biaisée, en particulier certains mois apportant de l'information sur la distribution ne sont jamais échantillonnés. Dans ce cas de configuration, les segments d'influence apparaissent moins biaisés que le krigeage.

L'écart type d'estimation permet de juger de la qualité de l'estimation et d'en évaluer la précision. Multiplié par 4 sous hypothèse de normalité, il permet de déduire l'amplitude des variations de l'intervalle de confiance à 95%.
Deux approches du calcul de l'écart type d'estimation sont abordées ici :

- L'approche expérimentale : on calcule l'écart type résiduel.
- L'approche théorique : l'écart type est prédit par le modèle. Les écarts type d'estimation prédits pour chaque échantillon sont moyennés.

L'ensemble de ces résultats est présenté TAB. 6-4 en pourcentage relatif. Les résultats donnés par les méthodes actuelles sont grisés. La comparaison des deux approches permet de vérifier la pertinence du modèle et si l'intervalle de confiance qu'il prédit est réaliste.

Dans tous les cas, l'intervalle de confiance prédit par le krigeage est inférieur à celui annoncé par les statistiques. Ces intervalles de confiance s'avèrent ici toujours fiables : d'une part les écarts types d'estimation dont ils sont déduit sont toujours proches des écarts types d'estimation expérimentaux et ils encadrent toujours la valeur réelle de l'indicateur (TAB. 6-1, TAB. 6-2, TAB. 6-3 et TAB. 6-4) dans tous les cas de configuration. En revanche, l'écart type d'estimation annoncé par les statistiques ou les segments d'influence est dans la plupart des cas très éloigné de l'écart type d'estimation expérimental. Par exemple, pour 12 mesures, l'écart type expérimental est de 8% pour les nitrates

contre 19% pour la prédiction. Ces différences sont importantes, surtout si l'écart type d'estimation est ramené à un intervalle de confiance à 95%, donc multiplié par 4.

Dans la majorité des cas, l'écart type d'estimation prédit par statistique surestime très nettement l'écart type d'estimation réel. On constate toutefois que le krigeage dans certains cas prédit à l'inverse un écart type d'estimation plus faible que l'expérimental, par exemple pour les nitrites lorsque l'échantillonnage préférentiel est en été.

En conclusion, cette étude expérimentale montre que dans tous les cas, il est préférable d'utiliser le krigeage ou les segments d'influence pour l'estimation de la moyenne annuelle car ils réduisent les biais causés par l'échantillonnage. Le krigeage permet de plus le calcul d'un intervalle de confiance proche de celui observé expérimentalement. Dans les exemples, le calcul de la variance d'estimation par segments d'influence a été réalisé en supposant que les données sont indépendantes alors que le krigeage suppose la structure temporelle connue et utilise cette information. On pourrait aussi calculer une variance d'estimation avec les segments d'influence qui tiendrait compte de la corrélation temporelle. Les variances d'estimation seraient alors sans doute proches de celles obtenues par krigeage mais leur calcul nécessite comme pour le krigeage la modélisation des variogrammes expérimentaux.

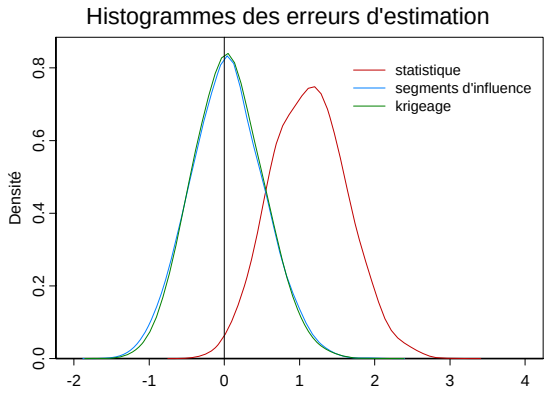


FIG. 6-6 : Nitrates, station 50500, 1985. Histogrammes des erreurs d'estimation expérimentales par statistique, segments d'influence et krigeage. Cas d'un échantillonnage de 18 mesures par an.

Moyenne annuelle de référence : 7.20	Estimation moyenne de la moyenne annuelle	Ecart type d'estimation moyen	Ecart type résiduel	Intervalle de confiance expérimental moyen contenant 95 % des estimations
Statistique	8.33	1.04	1.22	[7.39 ;9.31]
Krigeage	7.25	0.64	0.46	[6.39 ;8.18]
Segments d'influence	7.24	1.18	0.48	[6.32 ;8.21]

TAB. 6-1 Synthèse statistique sur la moyenne annuelle sur 3000 échantillons. Exemple pour un échantillonnage de 18 mesures par an préférentielles en hiver. Station 50500, nitrates, année 1985.

Moyenne annuelle de référence : 7.20	Estimation moyenne de la moyenne annuelle	Ecart type d'estimation moyen	Ecart type résiduel	Intervalle de confiance expérimental moyen contenant 95 % des estimations
Statistique	7.20	1.33	0.57	[6.14 ;8.37]
Krigeage	7.17	0.75	0.55	[6.12 ;8.30]
Segments d'influence	7.15	1.35	0.56	[6.11 ;8.30]

TAB. 6-2 Synthèse statistique sur la moyenne annuelle sur 3000 échantillons. Exemple pour un échantillonnage de 12 mesures par an. Station 50500, nitrates, année 1985.

Moyenne annuelle de référence : 0.0994	Estimation moyenne de la moyenne annuelle	Ecart type d'estimation moyen	Ecart type résiduel	Intervalle de confiance expérimental contenant 95 % des estimations
Statistique	0.10	0.03	0.01	[0.07 ;0.12]
Krigeage	0.10	0.01	0.01	[0.07 ;0.12]
Segments d'influence	0.10	0.03	0.01	[0.07 ;0.12]

TAB. 6-3 Synthèse statistique sur la moyenne annuelle sur 3000 échantillons. Exemple pour un échantillonnage de 12 mesures par an. Station 57800, nitrites, année 1996.

Nombre de prélèvements		6			9			12			15			18 été			18 hiver		
B= biais en % ET= écart type **résiduel** relatif en % ETP= écart type d'estimation **prédit** relatif en %		**B**	**ET**	**ETP**	**B**	**ET**	**ETP**	**B**	**ET**	**ETP**	**B**	**ET**	**ETP**	**B**	**ET**	**ETP**	**B**	**ET**	**ETP**
Nitrates	Statistique	-25	27	24	-10	13	22	0	8	19	12	14	17	-25	25	14	15	17	14
	Krigeage	-7	14	16	1	10	12	0	8	10	2	7	9	-5	9	10	1	6	9
	Segments d'influence	65	13	29	0	9	24	0	8	19	0	7	18	-4	10	20	1	7	16
Nitrites	Statistique	54	59	71	17	23	50	0	13	38	-11	16	31	53	55	39	-17	20	26
	Krigeage	30	35	27	1	12	16	1	12	11	1	11	10	22	25	20	0	11	10
	Segments d'influence	21	28	74	-1	12	47	1	12	39	1	11	36	20	24	48	1	11	34

TAB. 6-4: Biais et écarts type d'estimation expérimentaux et prédits par la méthode pour l'estimation de la moyenne annuelle des concentrations en nitrates, station 50500 en 1985 et pour la moyenne annuelle des concentrations en nitrites, station 57800 en 1996.

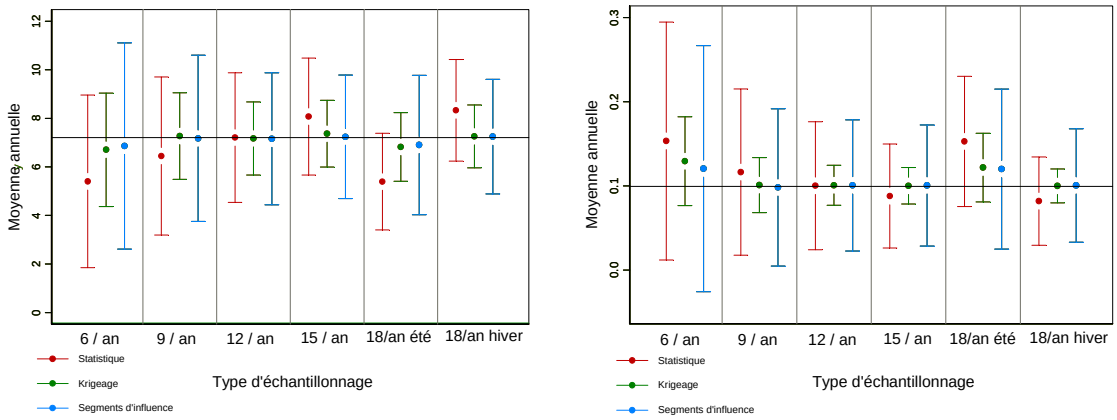


FIG. 6-7 : Estimation de la moyenne annuelle par 3 méthodes et intervalle de confiance à 95% prédit par chaque méthode. A gauche : Station 50500, nitrates, année 1985. A droite : Station 57800, nitrites, année 1996. Echantillonnage préférentiel et régulier. Les résultats par statistique classique sont présentés en rouge, par krigeage en vert et par segments d'influence en bleu, les indicateurs de référence en noir.

Estimation du quantile 90

Des résultats expérimentaux sont reportés au tableau TAB. 6-5 pour l'estimation du quantile 90. La figure FIG. 6-8 présente les estimations moyennes du quantile 90 et les intervalles de confiance expérimentaux à 95% pour différents types d'échantillonnage préférentiels ou réguliers. Ces intervalles ont pour bornes les quantiles d'ordre 0.975 et 0.025 des estimations. Pour le quantile, ces intervalles n'étant pas symétriques, il est intéressant de les présenter car ils indiquent la dispersion des estimations. Les biais expérimentaux et les écarts types résiduels relatifs à la vraie valeur du quantile sont donnés en pourcentage TAB. 6-6 pour les nitrates et TAB. 6-7 pour les nitrites. L'écart type résiduel permet d'évaluer la qualité des estimations car il met en évidence la variabilité des estimations autour du vrai quantile. Son importance donne une idée de la qualité d'une estimation sur une unique réalisation. Le biais quant à lui indique si on sous estime ou surestime systématiquement le quantile.

Contrairement au cas de la moyenne annuelle, on ne peut pas systématiquement prévoir le sens du biais pour le quantile lorsque l'échantillonnage est préférentiel. Par exemple, pour un échantillon de 9 mesures préférentielles en été, le quantile est surestimé alors qu'on s'attendrait à l'inverse. Ensuite, les segments d'influence et le krigeage n'améliorent pas systématiquement l'estimation du quantile 90. L'analyse des résultats conduit aux conclusions suivantes :

- Avec 6 mesures par an et un échantillonnage préférentiel en été, la pondération n'a pas d'influence dans le cas des nitrates. Pour les nitrites, les valeurs fortes étant plus nombreuses, le biais est diminué. Dans le cas des nitrates, l'interpolation linéaire augmente le biais et l'écart type d'estimation, contrairement aux nitrites. Le biais est important pour les nitrates, il dépasse les 10%, le krigeage associé à l'interpolation linéaire pour les nitrites fournit un biais de 2% et un écart type d'estimation de 14% contre un biais de 9% pour la règle des 90% avec un écart type d'estimation de 15%.
- Pour 9 mesures par an préférentielles en été, comme précédemment, les poids modifient les estimations pour les nitrites et non pour les nitrates. L'interpolation linéaire améliore les estimations pour les nitrates et les dégrade pour les nitrites. Les biais restent importants.

- Pour 12 mesures par an, la pondération améliore les estimations dans le cas des nitrites et ne change rien pour les nitrates. Associée à l'interpolation linéaire, les estimations moyennes pour les nitrates sont assez proches de la réalité. Pour les nitrites les résultats ne sont pas satisfaisants.
- Pour 15 mesures préférentielles en hiver, la pondération améliore les estimations à la fois pour les nitrites et les nitrates, les estimations restant très biaisées pour les nitrites. L'interpolation linéaire augmente les biais mais diminue les écarts type d'estimation.
- Pour 18 mesures préférentielles en été, pondération et interpolation améliorent les biais et écarts type d'estimation. Les résultats sont améliorés en faisant un calcul sur 3 ans. Pour l'ensemble des méthodes, le biais et l'écart type résiduel sont réduits, excepté pour la règle des 90%.
- Pour 18 mesures préférentielles en hiver, pondération et interpolation améliore les biais et écarts type d'estimation.

En moyenne sur l'ensemble des stratégies d'échantillonnage, les pondérations par krigeage et segments d'influence fournissent des résultats proches (TAB. 6-8). Dans les deux cas, nitrates et nitrites, le krigeage associé à une interpolation linéaire semble la meilleure méthode, fournissant un compromis entre le biais le plus faible et l'écart type d'estimation le plus faible. Cette méthode réduit en moyenne sur l'ensemble des échantillonnages le biais de la règle des 90% (on passe de 8% à 4.5% pour les nitrates et de 13.16% à 7.83% pour les nitrites) mais aussi l'écart type d'estimation relatif (de 14.5% à 11.8% pour les nitrates et de 21.8% à 15% pour les nitrites). Pour ces deux stations, le nombre de mesures reste insuffisant pour estimer un quantile.

Il est impossible de prétendre qu'une méthode est systématiquement meilleure que l'autre car l'estimation dépend de la configuration des données et de l'allure des chroniques temporelles. Les méthodes proposées restent biaisées avec des écarts types d'estimation importants, mais nous jugeons qu'en moyenne, elles améliorent les estimations et leurs variations et permettent de comparer plus pertinemment des stations aux stratégies d'échantillonnage différentes. Dans l'objectif d'étudier les risques toxiques à effet immédiat, les agences s'intéressent au quantile 90, signalons tout de même que l'estimation des moyennes annuelles est nettement plus fiable dans la mesure où les quantiles estimés sont toujours biaisés, quel que soit l'échantillonnage, ce qui n'est pas le cas pour la moyenne, et pour 12 mesures par an, les biais et écarts types résiduels sont plus faibles pour la moyenne que pour le quantile 90.

Quantile 90 annuel de référence : 12.97	Estimation moyenne du quantile 90% annuel	Ecart type résiduel	Intervalle de confiance expérimental moyen contenant 95 % des estimations
Statistique	13.99	1.62	[11.51 ; 17.11]
Krigeage	12.96	1.05	[11.49 ; 16.09]
Segments d'influence	13.07	1.16	[10.75 ; 15.75]
Interpolation linéaire	13.76	1.33	[11.77 ; 16.57]
Interpolation linéaire + krigeage	13.07	0.97	[11.42 ; 15.78]
Interpolation linéaire + segments d'influence	13.11	1.05	[11.13 ; 15.28]

TAB. 6-5 : Synthèse statistique sur le quantile 90 sur 3000 échantillons. Echantillonnage de 18 mesures par an, préférentielles en hiver. Station 50500, nitrates, année 1985.

Nombre de prélèvements		6		9		12		15		18 été		18 hiver	
B= biais en % ET= écart type résiduel relatif en %		B	ET	B	ET	B	ET	B	ET	B	ET	B	ET
Règle 90	Sans interpolation	-10	18	10	19	-4	11	6	12	-10	14	8	13
	Avec interpolation	-16	21	1	12	2	10	7	12	-2	14	6	10
Krigeage	Sans interpolation	-10	18	10	19	-4	11	5	12	-5	16	0	8
	Avec interpolation	-14	19	3	13	2	10	4	9	-3	12	1	8
Segments d'influence	Sans interpolation	-10	18	8	18	-2	12	1	10	-2	17	1	9
	Avec interpolation	-14	19	2	12	1	10	1	10	-5	14	1	8

TAB. 6-6 : Biais relatif moyen du quantile 90 et écart type résiduel sur 3000 échantillons. Station 50500, nitrates, année 1985.

Nombre de prélèvements		6		9		12		15		18 été		18 hiver	
B= biais en % ET= écart type résiduel relatif en %		B	ET	B	ET	B	ET	B	ET	B	ET	B	ET
Règle 90	Sans interpolation	9	15	9	15	-17	30	-17	30	10	12	-17	29
	Avec interpolation	1	13	-3	15	-12	19	-24	27	8	10	-34	37
Krigeage	Sans interpolation	6	15	-3	19	-6	22	-7	22	4	10	-6	21
	Avec interpolation	-2	14	-9	16	-11	17	-11	17	3	9	-11	17
Segments d'influence	Sans interpolation	-4	22	-8	23	-8	23	-8	23	1	12	-7	22
	Avec interpolation	-5	15	-12	18	-11	17	-11	17	1	11	-11	17

TAB. 6-7 : Biais relatif moyen du quantile 90 et écarts type résiduel sur 3000 échantillons. Station 57800, nitrites, 1996.

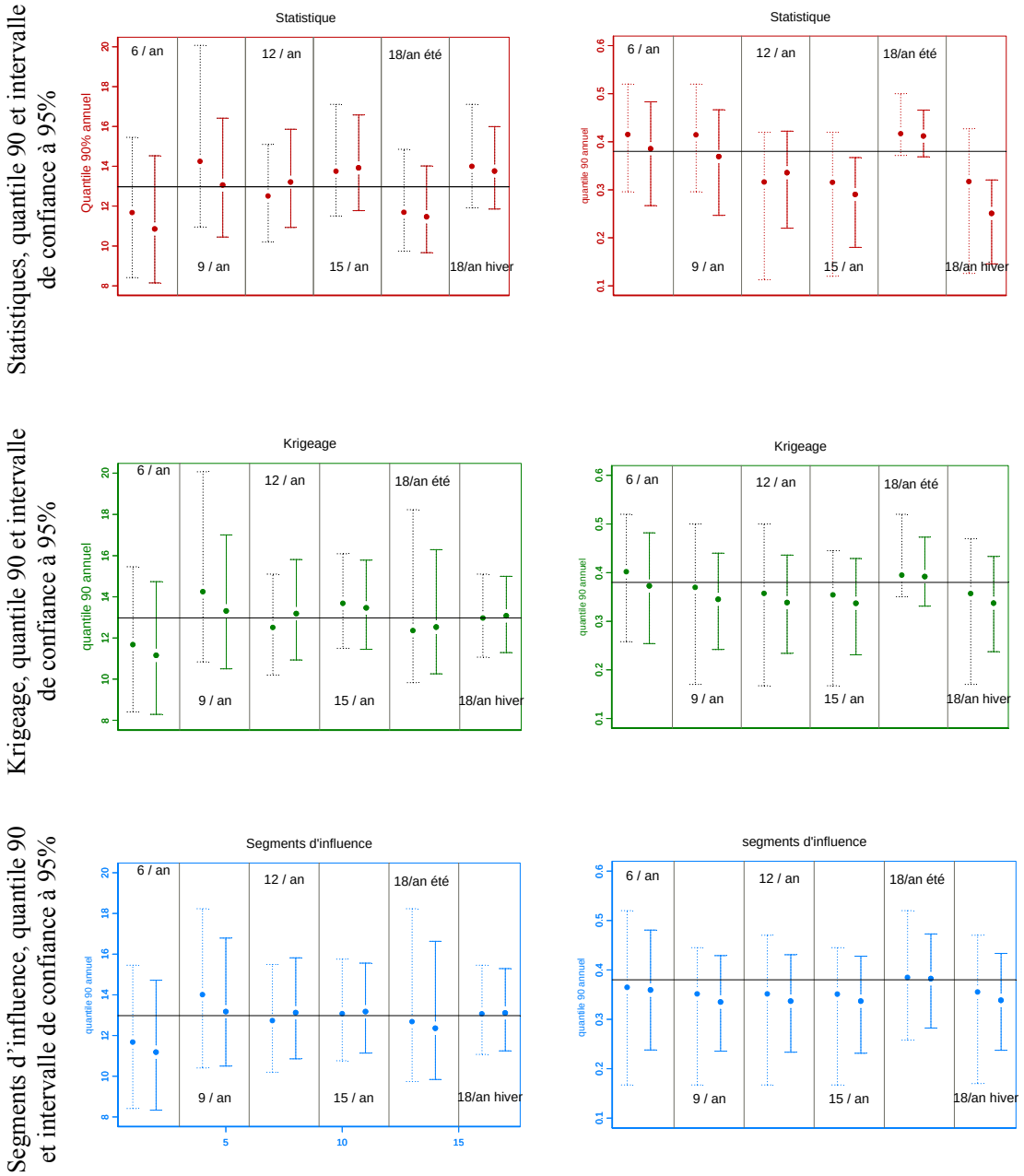


FIG. 6-8 : Quantile 90 annuel et intervalle de confiance expérimental à 95%. A gauche : Station 50500, nitrates, 1985. A droite : Station 57800, nitrites, 1996. Estimation par statistique, krigeage et segments d'influence (de haut en bas). Influence de l'interpolation linéaire du quantile empirique (les pointillés concernent les estimations sans interpolation linéaire). Echantillonnage préférentiel et régulier

B= biais moyen en % ET= écart type moyen d'estimation expérimental relatif en %		B	ET	B	ET
		Nitrates		Nitrites	
Règle 90	Sans interpolation	8	14.5	13.16	21.8
	Avec interpolation	5.6	13.16	13.6	20.1
Krigeage	Sans interpolation	5 .6	14	5.33	18.16
	Avec interpolation	4.5	11.8	7.83	15
Segments d'influence	Sans interpolation	4	14	6	20.83
	Avec interpolation	4	12.1	8.5	15.83

TAB. 6-8 : Biais et écarts types d'estimations du quantile 90 moyennés sur l'ensemble des échantillonnages pour chacune des méthodes. A gauche, nitrates, station 50500, 1985. A droite, nitrites, station 57800, 1996.

6.4 Etude par station

Au paragraphe précédent, le biais des méthodes a été étudié en se basant sur une unique simulation et en faisant varier les dates d'échantillonnage. Ainsi on peut se rendre compte pour une station et une année donnée du biais des méthodes en fonction de la taille d'échantillonnage et de l'amplitude des variations des estimations. On peut se poser aussi la question suivante : pour une station considérée, quelles sont les propriétés statistiques des estimateurs proposés ? Peut-on évaluer de manière générale le biais et l'écart type de dispersion des estimations pour chaque indicateur selon la stratégie d'échantillonnage ?

Toutes les années sont alors prises en compte. Nous allons donc présenter ici une application sur la station 171010, à Ploufragan sur le Gouet, qui peut être généralisée à toute station présentant en particulier des chroniques temporelles périodiques ou facilement modélisables. L'idée est de modéliser la série temporelle, d'étudier la structure variographique des résidus afin de se ramener à des simulations non conditionnelles. Ainsi, chaque simulation correspond alors à une série chronologique envisageable pour cette station. Le biais et l'écart type expérimental des estimations n'ont alors plus le même sens que dans la démarche précédente. Ici, on regarde la qualité des estimateurs pour un ensemble de chroniques reflétant les données d'une station sur plusieurs années. Cela donne une vue globale des erreurs commises sur le calcul d'indicateurs pour cette station, l'étude par année précédente n'étant pas possible en regard au faible nombre de mesures.

Soit $Z(t)$ la fonction aléatoire des concentrations en nitrates. On décompose $Z(t)$ comme la somme d'une composante déterministe $m(t)$ périodique représentant la moyenne des concentrations et d'un résidu aléatoire $R(t)$ résultant des variations interannuelles des concentrations

$$Z(t) = m(t) + R(t)$$

La composante déterministe peut être calculée par une régression directe sur l'ensemble de la chronique ou par régression sur les moyennes mensuelles de l'ensemble des années. Dans le cas de la station 171010, les résultats sont similaires. Ils sont présentés FIG. 6-9. On constate que la dispersion des concentrations est plus faible quand les concentrations décroissent.

Le variogramme expérimental des résidus est calculé puis modélisé.

Les résidus sont simulés conformément au variogramme choisi et non conditionnellement aux mesures. Les simulations sont validées par comparaison du variogramme théorique choisi et du variogramme expérimental moyen sur les 2000 simulations (voir FIG. 6-9 en bas à droite)

Ici, les types d'échantillonnage testés sont les mêmes que pour l'étude des chroniques reconstituées, mais contrairement à celle-ci, les échantillons ne sont pas tirés aléatoirement. Les dates d'échantillonnage ont été fixées et sont retenues pour l'ensemble des simulations. Les dates retenues correspondent à la date se situant au milieu de chaque strate décrite paragraphe 6.4.

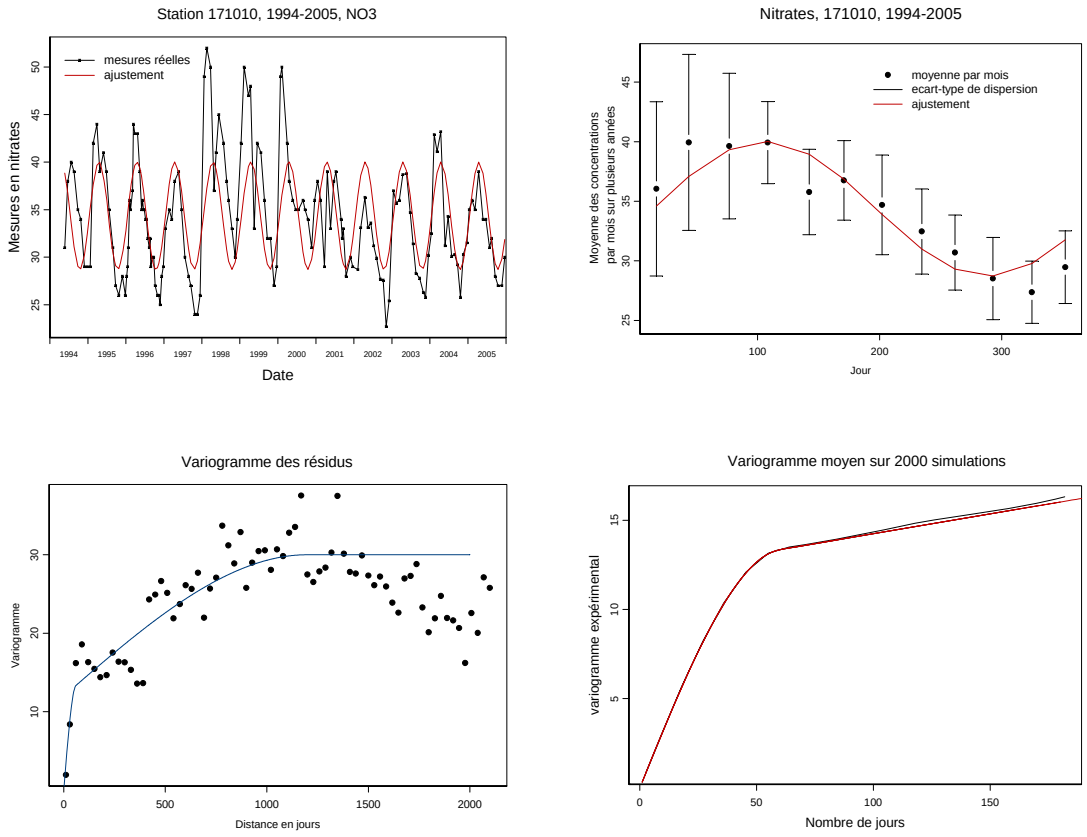


FIG. 6-9 : Stations 171010, nitrates. Entre 1994 et 2006. En Haut : à gauche : concentrations en nitrates en fonction du temps et modélisation de la partie déterministe de la série temporelle, à droite, moyennes mensuelle sur les 12 ans. En bas, à gauche, variogramme expérimental et modélisé des résidus, à droite, variogramme moyen sur 2000 simulation et variogramme théorique.

Moyennées sur 2000 simulations, les estimations de la moyenne annuelle montrent alors des biais très faibles FIG. 6-10. En revanche, l'intervalle de confiance prévu par krigeage est cinq fois moindre que l'intervalle calculé par statistique classique sur 2000 simulations. De plus, l'intervalle prévu par krigeage correspond bien à l'intervalle expérimental. L'écart type d'estimation expérimental diffère ici de celui calculé dans le cas d'une unique simulation conditionnelle. Ici, pour chaque simulation, la

moyenne annuelle et le quantile 90 estimés sont comparés à la valeur calculée sur l'ensemble de la simulation considérée.

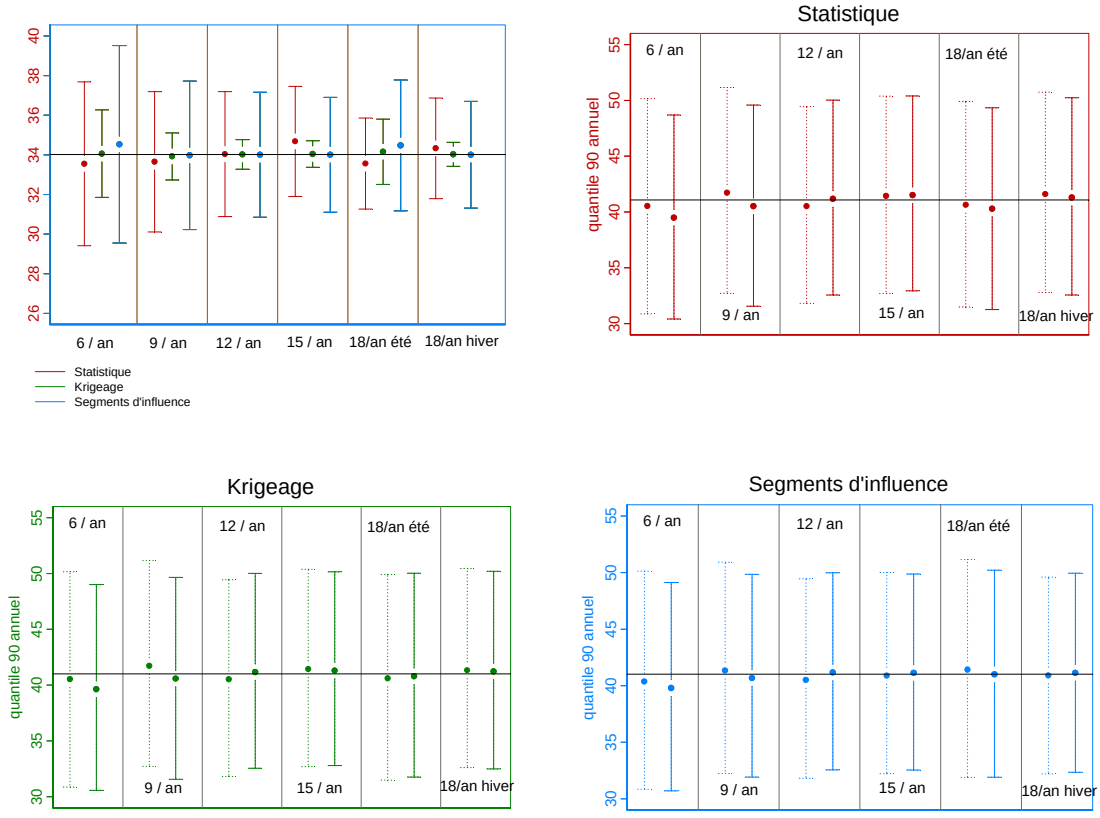


FIG. 6-10 : Station 171010. En haut à gauche : Moyenne annuelle et intervalle de confiance prédit à 95%. Ailleurs : quantile 90 annuel. Echantillonnage régulier et préférentiel. Rouge : statistique, vert : krigeage, bleu : segments d'influence. En pointillé : quantile empirique. Trait plein : avec interpolation linéaire

En moyenne, les estimations du quantile 90 entre les différentes méthodes présentent des intervalles de confiance à 95% similaires. Les estimations varient entre $\pm 25\%$ de l'estimation moyenne, ce qui est très important et peut engendrer une affectation erronée à une classe de qualité quelle que soit la taille de l'échantillon. L'écart type d'estimation expérimental moyen atteint jusqu'à 2.6 mg /l NO_3 (TAB. 6-9) ce qui représente environ 7% de la valeur moyenne du quantile 90.

Nombre de prélèvements		6		9		12		15		18 été		18 hiver	
B= moyenne des biais relatifs sur 2000 simulations ET= écart type résiduel		B	ET	B	ET	B	ET	B	ET	B	ET	B	ET
Règle 90	Sans interpolation	-1	2.26	2	1.78	-1	1.19	1	1.10	-1	1.80	1	1.18
	Avec interpolation	-4	2.52	-1	1.47	0	0.85	1	0.94	-2	1.81	1	10.90
Krigeage	Sans interpolation	-1	2.6	2	1.78	-1	1.19	1	1.10	-1	1.81	1	1.07
	Avec interpolation	-3	2.43	-1	1.43	0	0.85	1	0.84	-1	1.71	0	0.80
Segments d'influence	Sans interpolation	-2	2.36	1	1.77	-1	1.19	0	1.04	1	1.97	0	0.91
	Avec interpolation	-3	2.37	-1	1.38	0	0.84	0	0.75	0	1.72	0	0.71

TAB. 6-9 : Station 171010, nitrates. Biais relatif moyen du quantile 90 et écarts type d'estimation expérimental relatif sur 2000 échantillons.

En résumé, la meilleure solution pour estimer au mieux le quantile est d'utiliser un échantillonnage de taille régulier. Le biais et les écarts types d'estimations expérimentaux étudiés sont alors minimaux.
On peut d'ailleurs se demander si il ne serait pas préférable de prélever des mesures régulières plutôt que préférentielles quand il y a des restrictions budgétaires. Nous avons donc étudié le biais et l'écart type résiduel pour des échantillons réguliers de taille n allant de 1 à 36 mesures.
Les figures FIG. 6-11 et FIG. 6-12 présentent l'espérance des indicateurs et l'écart type résiduel en fonction du nombre de prélèvements pour des échantillons réguliers (chaque année, même échantillonnage de *n* mesures réparties régulièrement dans l'année). On voit alors que pour des échantillons réguliers, les différents estimateurs de la moyenne annuelle sont sans biais. L'écart type résiduel n'est par contre pas réduit. Pour le quantile, un échantillonnage régulier réduit à la fois le biais et l'écart type résiduel. D'où l'intérêt de prélever des mesures régulières plutôt que préférentielles.

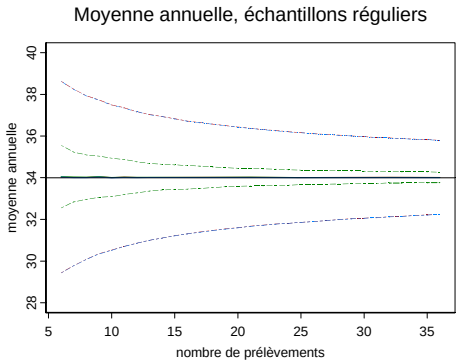


FIG. 6-11 : Evolution de la moyenne annuelle sur 1000 simulations et intervalles de confiance moyens prédits par les méthodes. Station 171010, nitrates. Echantillons réguliers. En rouge : statistique. En bleu : segments d'influence. En vert : krigeage.

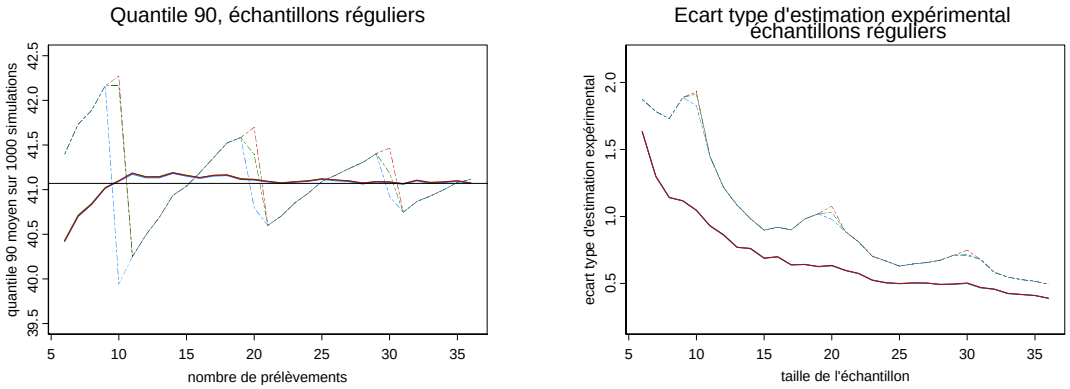


FIG. 6-12 : Station 171010. Echantillons réguliers. A gauche : quantile 90 moyen sur 1000 simulations. A droite : écart type d'estimation expérimental. En rouge : statistique. En bleu : segments d'influence. En vert : krigeage. En pointillé : quantile empirique. En continu : quantile empirique linéarisé.

6.5 Conclusion

Dans tous les cas, le krigeage fournit en théorie et en pratique les meilleures estimations de la moyenne au sens du biais et de la variance d'estimation minimale. Dans certains cas, le biais n'est pas parfaitement corrigé lorsque le caractère préférentiel de l'échantillonnage est trop marqué (par exemple 6 mesures par an principalement estivales).

L'estimation du quantile 90 est moins fiable. Cependant, les simulations montrent qu'une interpolation linéaire du quantile empirique associée à une pondération permet en moyenne de diminuer les biais et écarts types résiduels, et ceci pour les différents types d'échantillonnage. Cependant, les estimations sont à analyser avec beaucoup de précaution, car elles peuvent être très éloignées de la réalité, en particulier pour moins de 10 mesures par an. Pour estimer le quantile, il serait préférable d'avoir des mesures plus nombreuses, de préférence régulières mais surtout il faut éviter de prélever préférentiellement aux périodes de valeurs faibles. Si toutes les mesures sont concentrées sur la même saison, le calcul des indicateurs relatifs aux trois années n'améliore pas toujours significativement les estimations. Enfin un calcul d'incertitude devrait être associé à chaque estimation avant l'affectation à une classe. Il peut être évalué par simulation. En pratique, des approximations pourraient être aussi recherchées.

Sur les simulations étudiées, le krigeage et les segments d'influence fournissent pour le quantile des résultats très proches, parfois meilleurs par segments d'influence. Cependant, nous préconisons de retenir le krigeage car d'une part, pour la moyenne annuelle, il est la meilleure des méthodes proposées, d'autre part les résultats sont présentés ici en moyenne, ce qui a pour effet de lisser tous les artefacts sur un unique échantillon, présentés aux paragraphes 4.3 et 5.5. Ensuite, en l'absence de corrélation temporelle, comme c'est le cas pour de nombreuses stations, la méthode des segments d'influence n'est pas adaptée.

Chapitre 7

Application au réseau RNB Loire-Bretagne

Les stations RNB retenues aujourd'hui par les agences pour évaluer la qualité présentent des échantillons réguliers, plus nombreux et donc plus fiables. Dans ce cas, le krigeage de la moyenne diffère peu du calcul statistique, mais permet un meilleur calcul d'incertitudes.

Si ce type de stratégie est à espérer pour toutes les agences dans le futur ce qui improbable faute de moyens, les méthodes proposées restent cependant intéressantes pour les années passées dans le but d'observer l'évolution des concentrations. Les méthodes proposées, prouvées meilleures, changent-elles radicalement les estimations sur l'ensemble du bassin ? C'est ce que nous regardons dans ce chapitre pour les stations RNB du 8ème programme d'intervention entre 2002 et 2004 et entre 1985 et 2005, pour l'agence Loire Bretagne.

7.1 Modélisation des variogrammes

Nous avons modélisé les variogrammes pour les 269 stations du RNB8 pour les nitrates, les nitrites et les orthophosphates. La modélisation est effectuée à l'aide du programme semi-automatique de régression non linéaire par minimisation des moindres carrés par l'algorithme de Gauss-Newton (Becker et al., 1988 ; Pinheiro et al., 2000 ; Venables and Ripley, 1994).

Les variogrammes expérimentaux ont été calculés à l'aide des mesures entre 1995 et 2005 pour éviter de prendre en compte inutilement trop d'annés pour établir les corrélations, la pollution ayant changé significativement à partir des années 1990. Selon les variogrammes expérimentaux, nous avons ajusté un modèle sur ces 5 années ou moins car certains modèles s'ajustent mal sur 5 ans. De plus, pour les calculs d'indicateurs annuels, seule une année est utile. En revanche pour un calcul sur trois ans, trois années sont nécessaires. Il est arrivé que certains variogrammes ne soient pas calculables sur 3 ans (certaines stations n'ayant que 3 années de mesures), nous avons alors considéré que le variogramme calculé sur une année était valable sur 3 ans. A priori ces variogrammes doivent être utilisés pour des mesures effectuées entre 1995 et 2005, nous les avons toutefois utilisés pour tracer l'évolution des indicateurs sur 20 ans.

Globalement, quatre types de modèles sont rencontrés pour tous les paramètres étudiés :

- Absence de corrélation (FIG. 7-1; station 74200). Le modèle ajusté pépitique. A peine 3% des stations sont concernées pour les nitrates, 17% pour les nitrites et 13% pour les orthophosphates.
- Le modèle linéaire avec effet de pépite (FIG. 7-1, station 24300). Il concerne 2% des stations pour les nitrates, 20% pour les nitrites et 12% pour les orthophosphates.
- Le modèle sphérique avec effet de pépite (FIG. 7-1, station 11000). Il concerne 6% des stations pour les nitrates, 7% pour les nitrites et 9% pour les orthophosphates
- Des modèles périodiques avec une composante en cosinus. Ils concernent la majorité des stations, 89% pour les nitrates, 56% pour les nitrites et 66% pour les orthophosphates, alors que dans de nombreuses chroniques, la périodicité n'est pas visible (FIG. 4-8).

Les différences entre les variogrammes à composante périodique se résument par les points suivants:

- ✓ Amplitudes différentes
- ✓ Présence d'une structure sphérique (FIG. 7-1, station 172890). Elle peut être de courte ou longue portée auquel cas le modèle équivaut à un modèle linéaire.
- ✓ Absence de structure sphérique (FIG. 7-1, station 155500).
- ✓ Période d'une année ou d'une demi-année (FIG. 7-1, station 11300). La période d'une demi-année est plus particulièrement rencontrée dans les le cas des nitrites

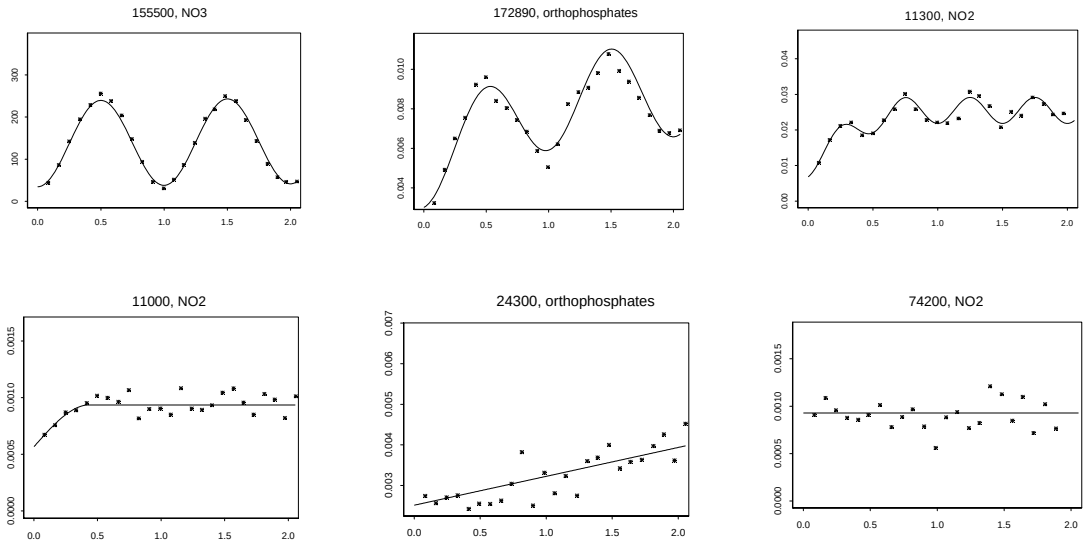


FIG. 7-1 : Les différents modèles de variogrammes rencontrées sur les données RNB en nitrates, nitrite et orthophosphates.

La modélisation des variogrammes étant partiellement automatique, les modèles retenus pour la modélisation ne sont pas toujours les plus pertinents. Une étude de sensibilité sur le choix des paramètres et du modèle pourrait être effectuée. En particulier, (Emery and Arnaud, 2000) montrent au travers d'exemples que la variance d'estimation peut nettement différer selon le modèle et les paramètres choisis. Cependant, dans le cas des rivières, l'incertitude résultant du faible nombre de mesures ainsi que de l'échantillonnage préférentiel est plus importante que celle induite par le choix du modèle.

7.2 Statistiques globales sur le réseau

La question se pose de l'influence globale de l'estimation géostatistique sur l'ensemble du réseau Loire Bretagne :

- Quels écarts entre les moyennes annuelles calculées par statistiques classiques ou par krigeage ?
- Quel gain sur les écarts types et intervalles de confiance associés ?
- Quels écarts entre le quantile 90 estimé par la règle des 90% ou par le quantile empirique « linéarisé » et pondéré ?

On note par station :

- La moyenne annuelle arithmétique M_A et l'écart type d'estimation associé E_A
- L'estimation par krigeage M_K de la moyenne annuelle et l'écart type d'estimation associé notée E_K
- Le quantile estimé par la règle des 90% : Q_A
- Le quantile estimé par linéarisation du quantile empirique et avec pondération par poids de krigeage de la moyenne annuelle : Q_K
- Le quantile estimé sans linéarisation avec pondération par poids de krigeage de la moyenne annuelle : Q_K^{SL} (simplement dans l'objectif de voir l'effet de la pondération.

Les résumés statistiques pour évaluer les différences sont alors les suivants :

- L'écart entre les moyennes estimées par statistique et krigeage $M_A - M_K$ puis l'écart relatif absolu entre les moyennes en pourcentage $\left|\frac{M_A - M_K}{M_A}\right| \times 100$.
- L'écart entre les écarts types d'estimation estimés par statistique et krigeage $E_A - E_K$ puis l'écart relatif absolu entre les moyenne en pourcentage $\left|\frac{E_A - E_K}{E_K}\right| \times 100$. Nous avons ici divisé l'écart par l'écart type d'estimation estimé par krigeage car certains écarts types d'estimation estimés par statistique sont nuls.
- L'écart entre les quantiles estimés par statistique et géostatistique (sans interpolation) $Q_A - Q_K^{SL}$ puis l'écart relatif absolu entre les moyenne en pourcentage $\left|\frac{Q_A - Q_K^{SL}}{Q_A}\right| \times 100$.
- L'écart entre les quantiles estimées par statistique et géostatistique $Q_A - Q_K$ puis l'écart relatif absolu entre les moyenne en pourcentage $\left|\frac{Q_A - Q_K}{Q_A}\right| \times 100$.

Ces résultats sont présentés pour des estimations annuelles entre 2002 et 2004, et 1985 et 2005, mais aussi pour une estimation tri-annuelle entre 2002 et 2004. Les résultats sont présentés bruts ou en moyenne en fonction de la taille de l'échantillonnage.

7.2.1 Par année entre 2002 et 2004

Les figures se lisent de la manière suivante (voir par exemple FIG. 7-2):

De haut en bas, sont comparés :

- La moyenne annuelle arithmétique et l'estimation par krigeage
- L'écart type d'estimation de la moyenne annuelle estimé par statistique et par krigeage
- Le quantile 90 calculé par la règle des 90% et le quantile empirique pondéré par poids de krigeage
- Le quantile 90 calculé par la règle des 90% et par interpolation linéaire du quantile empirique pondéré par poids de krigeage

De gauche à droite :

- écarts entre les indicateurs en fonction de l'effectif par an
- écarts relatifs en valeur absolue entre les indicateurs en fonction de l'effectif par an
- écarts relatifs moyennés par classe d'effectif en fonction de l'effectif

7.2.1.1 Nitrates

Moyenne annuelle (FIG. 7-2)

Pour moins de 10 mesures par an le krigeage a pour effet d'augmenter la valeur estimée de la moyenne annuelle et inversement pour plus de 10 mesures. Les écarts varient au maximum de ± 4 mg / NO_3 ce qui représente un écart relatif de 50% par rapport à la valeur moyenne. C'est pour un prélèvement de 6 mesures par an que les différences entre les méthodes sont les plus importantes avec en moyenne une différence de 12%. Vient ensuite le cas d'un échantillonnage préférentiel en hiver avec 18 mesures par an ou la différence moyenne atteint 7%. En toute logique, l'écart est minimal pour 12 mesures par an. Le gain sur l'écart type d'estimation par le krigeage atteint jusqu'à 600%, ce qui conduit à un intervalle de confiance à 95% 6 fois plus petit. En moyenne, le gain sur l'intervalle de confiance varie de 2 à 4, avec un maximum pour 18 mesures par an.

Quantile 90 (FIG. 7-2)

Les différences observées sur l'estimation du quantile sont plus faibles avec un maximum de 8% pour 12 mesures par an. Dans ce cas, c'est principalement la linéarisation qui modifie les estimations et non la pondération. De même pour moins de 10 mesures par an. L'effet de la pondération est maximal pour les échantillons préférentiels en période de fortes concentrations et de taille supérieure à 12 mesures par an. Les différences entre les méthodes peuvent atteindre jusqu'à 20 mg NO_3/ L ce qui est une différence très importante et implique probablement un changement de classe.

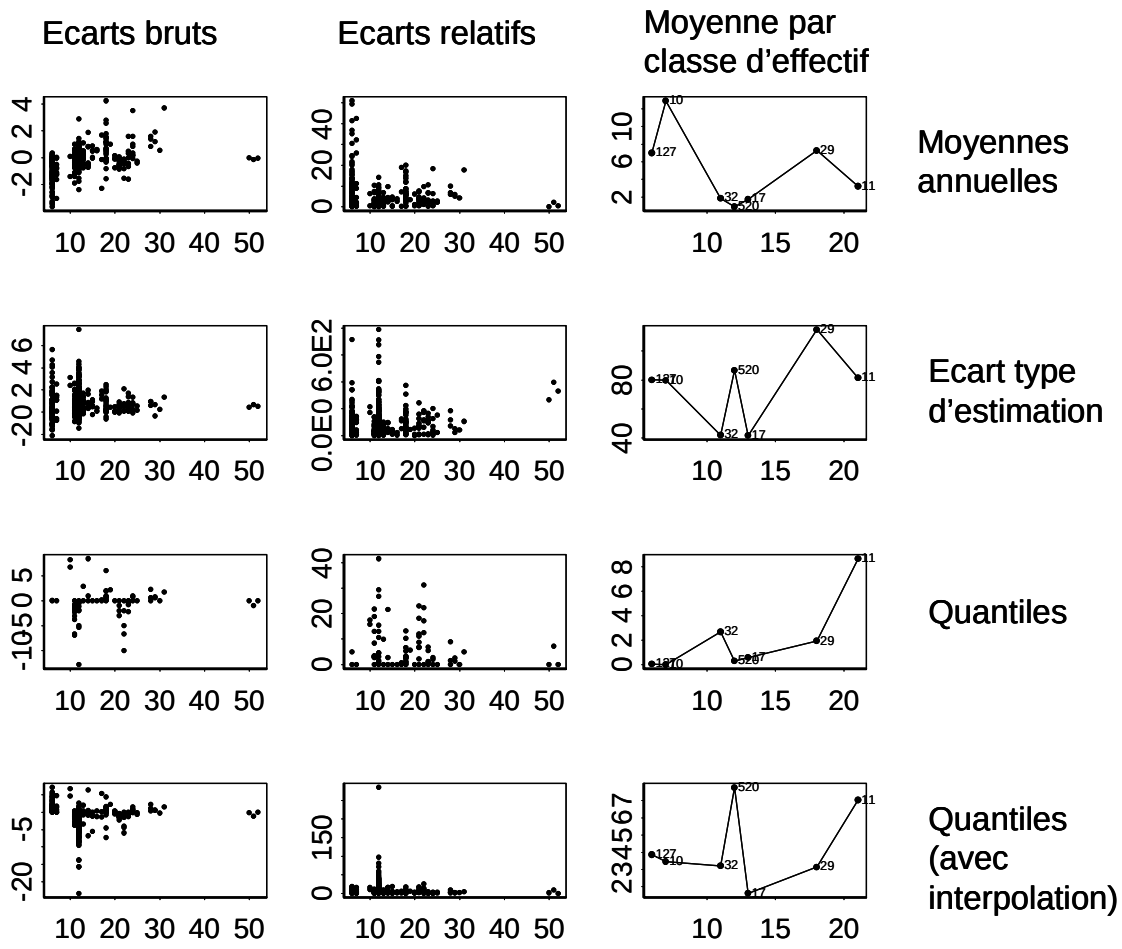


FIG. 7-2 : Ecart entre les estimations statistiques et géostatistiques par année pour les nitrates sur les 269 station RNB8 entre 2002 et 2004. Se reporter au début du paragraphe pour la lecture des graphiques.

7.2.1.2 Nitrites et orthophosphates

Moyenne annuelle (FIG. 7-3)

Pour les nitrites, les différences des estimations de la moyenne annuelle sont moins importantes que pour les nitrates. Ceci est dû au fait que les chroniques en nitrites présentent des variations saisonnières nettement moins marquées que les nitrates, voire inexistantes. Pour de nombreuses stations, les corrélations sont pépitiques (17%).

Cependant les écarts observés peuvent atteindre jusqu'à 40% pour un prélèvement de 6 mesures par an avec une différence d'en moyenne 3.5%. Contrairement aux nitrates, la différence entre les méthodes n'est pas systématiquement positive ou négative pour un type de prélèvement donné. Les différences entre les écarts types d'estimation sont importantes, atteignant jusqu'à 1200%. Ces différences « hors proportion » s'expliquent en général par la présence d'une donnée aberrante. Dans ce cas, il ne s'agit pas nécessairement d'une indication que le krigeage est beaucoup plus précis. On aurait pu songer à utiliser un estimateur « robuste » de la variance expérimentale si les distributions sont fortement asymétriques et ainsi obtenir un écart type d'estimation plus faible.

En moyenne, les différences varient entre 30 et 60% avec un maximum pour 6 mesures par an.

Le quantile 90 (FIG. 7-3)

Pour le quantile, les écarts sont importants pour les nitrites variant en moyenne de 5 à 20%, le maximum s'observant à nouveau pour une fréquence de 12 mesures par an. Dans ce cas, le quantile la règle des 90% retient l'avant dernière mesure la plus forte. Si l'échantillonnage est irrégulier, la pondération peut facilement provoquer un changement de classe avec comme conséquence l'estimation du quantile 90 par la valeur la plus forte (11/12 vaut 0.91, ce qui est très proche de l'ordre considéré). Ensuite, les nitrites présentent généralement des variations très marquées par une ou deux valeurs fortes chaque année, ce qui conduit à de nettes différences lors de l'application d'une interpolation linéaire du quantile.

Pour les orthophosphates, les conclusions sont analogues (FIG. 7-4).

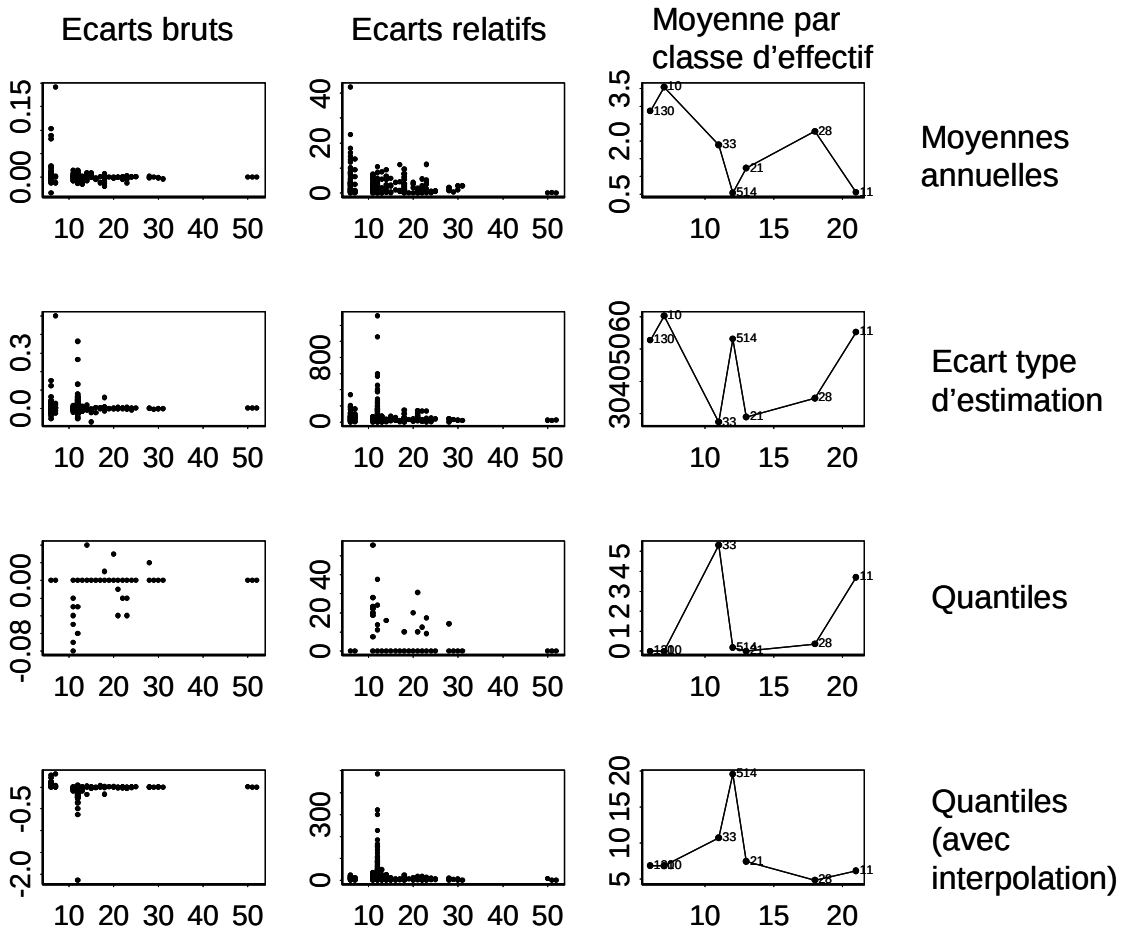


FIG. 7-3 : Ecart entre les estimations statistiques et géostatistiques par année pour les nitrites sur l'ensemble du réseau RNB du 8ème programme d'intervention entre 2002 et 2004. Se reporter au début du paragraphe pour la lecture des graphiques.

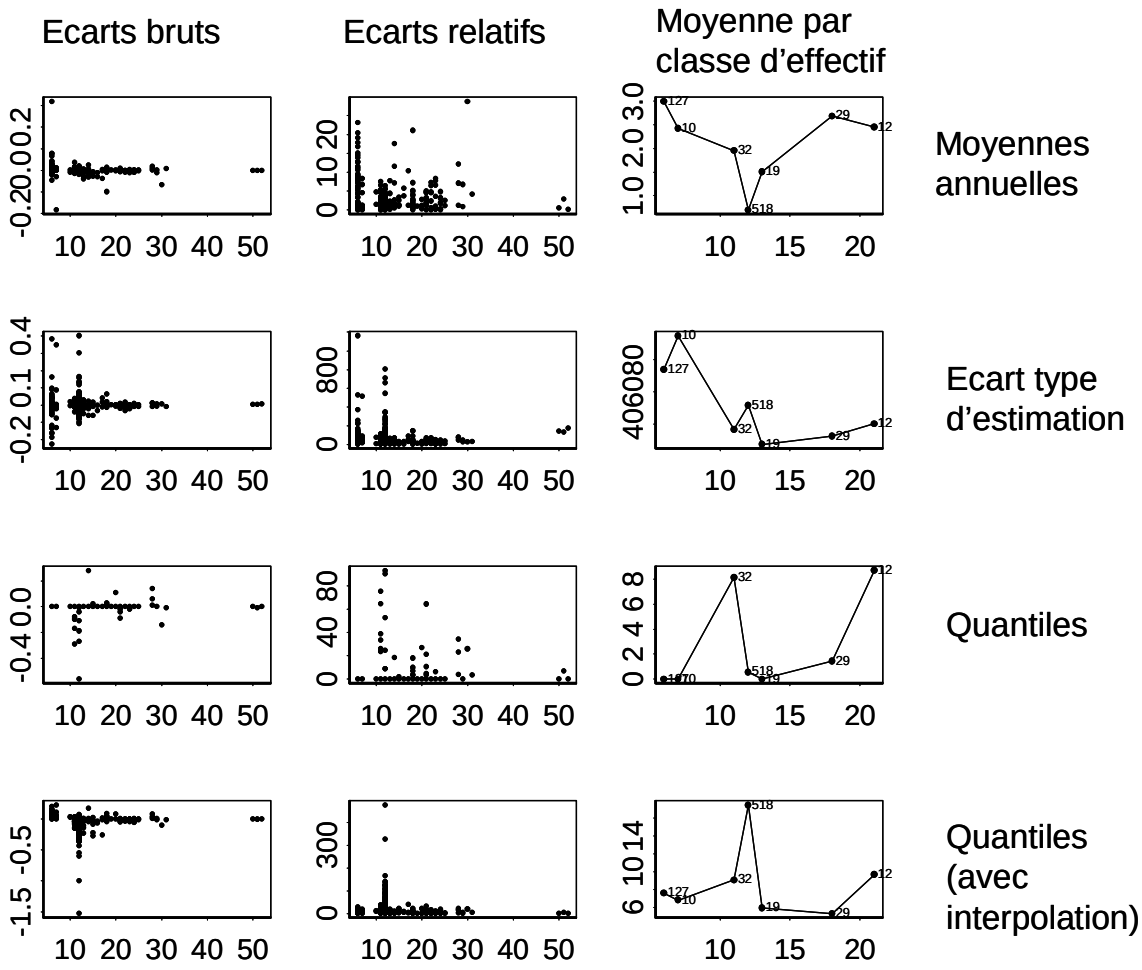


FIG. 7-4 : Ecart entre les estimations statistiques et géostatistiques par année pour les orthophosphates sur l'ensemble du réseau RNB du 8ème programme d'intervention entre 2002 et 2004. Se reporter au début du paragraphe pour la lecture des graphiques

7.2.2 Estimation sur trois années (2002-2004)

Les résultats du SEQ-Eau sont parfois présentés sur 3 ans, essentiellement à cause des stations peu échantillonnées. La même étude que précédemment est présentée pour un calcul des indicateurs tri-annuels. Les résultats sont présentés en annexe B.

Les conclusions sont les suivantes :

- Pour la moyenne annuelle, les différences entre les méthodes sont réduites à la fois pour l'estimation de la moyenne temporelle ainsi pour l'écart type d'estimation théorique. Pour tous les paramètres, l'écart maximal s'observe pour un échantillonnage de 6 mesures par an avec en moyenne une différence de 6% pour les nitrates, 4% pour les nitrites et 3% pour les orthophosphates. Les mesures sont plus nombreuses et globalement mieux réparties car la stratégie d'échantillonnage peut différer selon l'année

- Pour les trois substances, le gain sur les intervalles de confiance reste important.
- Pour le quantile, la différence maximale est atteinte avec un échantillonnage de 6 mesures préférentielles en été. Les différences sont plus importantes pour les nitrites et orthophosphates, en moyenne de 10 / 12 %, pour lesquels les valeurs fortes sont plus souvent en été.

7.2.3 Evolution à long terme (1985-2005)

Les résultats présentés en annexe B sont les suivants :

- Pour la moyenne annuelle, les différences relatives moyennes restent semblables avec toutefois une légère augmentation pour les nitrites et orthophosphates. En revanche les estimations varient beaucoup avec des différences allant jusqu'à 120% pour les nitrates. De même, la différence sur les écarts type d'estimation reste semblable en moyenne mais atteint jusqu'à 2000% pour les nitrites. Les courbes tracées sur l'ensemble des années traitent beaucoup plus de types d'échantillonnage et les moyennes sont calculées sur ces classes avec un effectif beaucoup plus grand qu'entre 2002 et 2004.
- Pour le quantile 90, les différences les plus importantes (12% en moyenne) concernent les échantillonnages à 15 mesures par an entre 2002 et 2004. Les années et stations concernées par un échantillonnage préférentiel en hiver (18 mesures) passent de 29 pour la période 2002-2004 à 175 pour la période 1985-2004. L'écart observé entre les estimations est alors de 6%. Pour les nitrites et orthophosphates, les différences les plus importantes restent pour un échantillonnage de 12 mesures par an avec des amplitudes moyennes similaires.

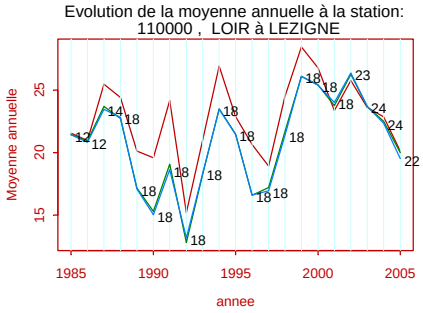
Nous présentons FIG. 7-5 quelques évolutions à long termes par station pour les trois substances étudiées.

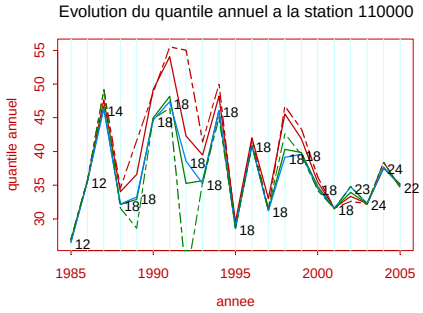
Nous avons tracé pour l'ensemble des stations RNB, l'évolution entre 1985 et 2005 des deux indicateurs, moyenne annuelle et quantile 90, par les estimateurs suivants :

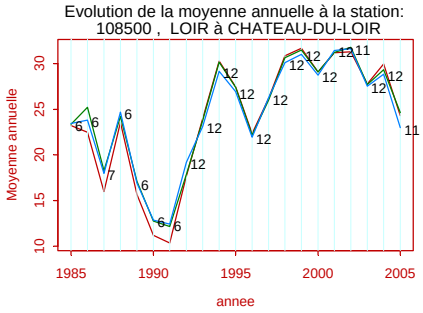
- Moyenne annuelle estimée par :
    - la moyenne arithmétique (en rouge sur les graphes de gauche)
    - segments d'influence (en bleu sur les graphes de gauche)
    - krigeage (en vert sur les graphes de gauche)
- Quantile 90 estimé par :
    - la règle des 90 % (en rouge pointillé sur les graphes de droite)
    - la règle des 90% linéarisée (en rouge continu sur les graphes de droite)
    - la règle des 90% pondérée par les poids de krigeage (en vert pointillé sur les graphes de droite)
    - la règle des 90% linéarisée et pondérée par les poids de krigeage (en vert continu sur les graphes de droite)
    - la règle des 90% linéarisée et pondérée par les segments d'influence (en bleu continu sur les graphes de droite)

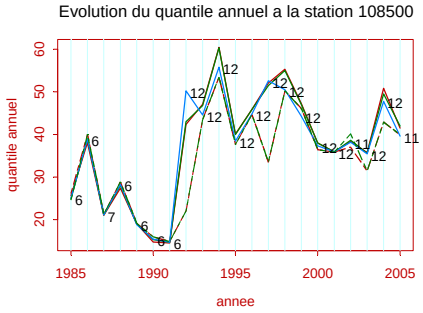
Le nombre de mesures dans l'année est indiqué à coté de chaque estimation.

Nitrates

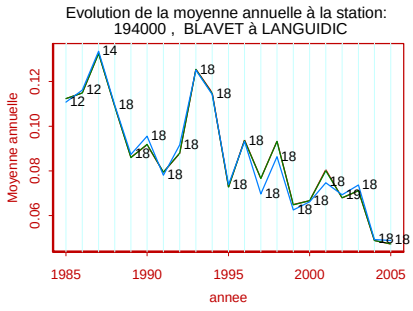


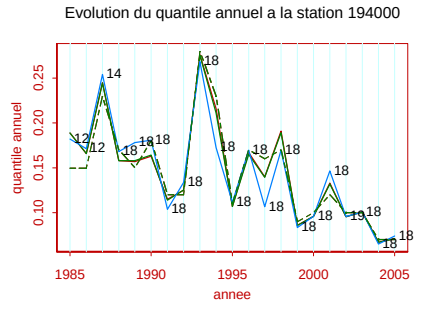


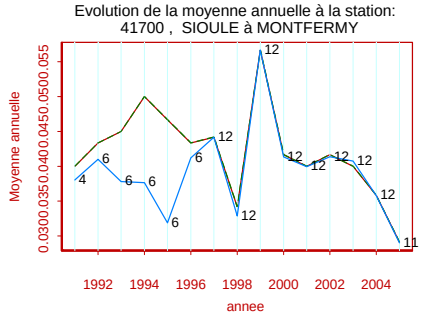


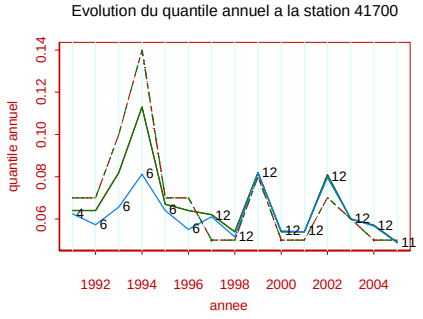


Nitrites









Orthophosphates

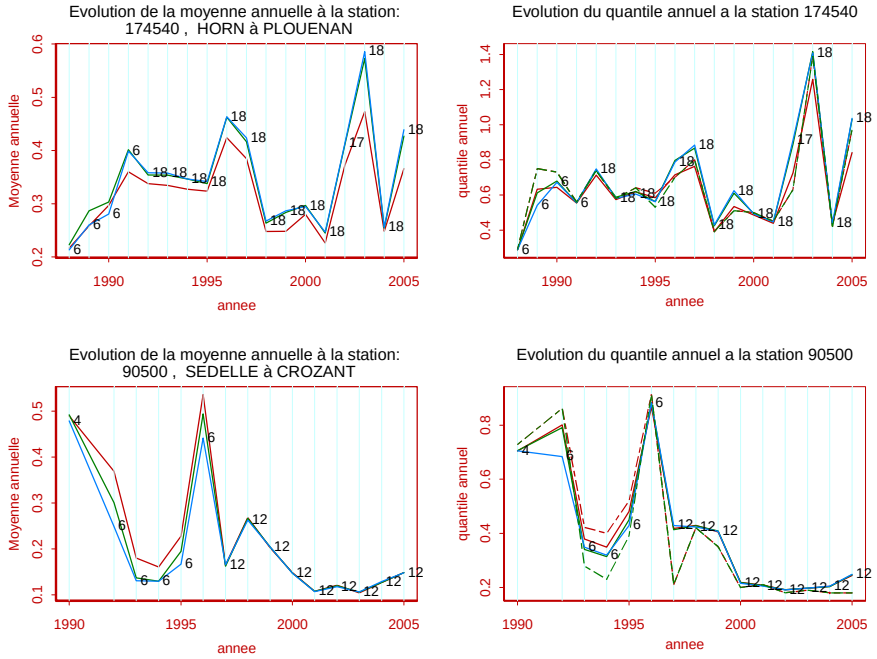


FIG. 7-5 : Evolution des moyennes annuelles (à gauche) et quantiles 90 (à droite) en nitrates, nitrites et orthophosphates pour différentes stations et techniques d'estimation. En rouge, statistiques classiques. En vert, pondération par krigeage. En bleu, pondération par segments d'influence. Pour les courbes des quantiles : en pointillé, quantile empirique, en continu, linéarisation du quantile empirique.

7.3 Analyse par région

Il est intéressant de regarder si les différences entre les méthodes sont plus marquées pour certaines régions. Les calculs précédents ont été repris pour chacune des régions hydrographiques. Les calculs effectués par année entre 2002 et 2004, sur un bloc de trois années ou depuis 1985, conduisent à des résultats différents surtout en ce qui concerne l'amplitude des écarts, mais nous présentons ici des conclusions globales.

Les nitrates

- Pour la moyenne annuelle, les différences les plus marquées s'observent dans la région Loire aval et côtiers vendéens, avec une différence moyenne de 25% pour les échantillons de 6 mesures par an. La Mayenne-Sarthe-Loir, la Loire moyenne et la Bretagne présentent elles aussi des différences importantes à la fois pour des échantillons de 6 mesures, 18 mesures mais aussi 14 mesures en Loire Moyenne. Les différences entre les méthodes sont à l'opposé réduites en Vienne Creuse et l'Allier-Loire Amont avec en moyenne au maximum 5%.
- Pour le quantile, les différences sont importantes atteignant jusqu'à 20% en moyenne pour la Loire moyenne et la Vendée, pour les échantillons de 14 mesures par an. Les différences sont moins

importantes pour le quantile que pour la moyenne. Les différences restent les plus faibles pour la Vienne Creuse et l'Allier-Loire Amont.

Globalement, les différences les plus importantes correspondent aux régions les plus polluées.

Les nitrites

- Pour les moyennes annuelle, les écarts entre les méthodes d'estimation restent relativement faibles (la différence moyenne maximale observée est de l'ordre de 5%) et identiques selon la région étudiée.
- Pour les quantiles, les écarts entre les méthodes sont importantes mais du même ordre quelle que soit la région.

Les orthophosphates

- C'est en Bretagne et Loire Moyenne que les écarts les plus importants sont observés pour les orthophosphates, mais les différences restent plus faibles que pour les nitrates.
- Les différences observées pour les quantiles sont similaires quelle que soit la région et sont importantes.

7.4 Conclusion

L'estimation de la moyenne annuelle est significativement modifiée par le krigeage en particulier pour les nitrates à l'ouest du bassin, là où les concentrations sont fortes. Pour les nitrites et les orthophosphates, l'écart est plus réduit. Dans toutes les régions, l'écart type d'estimation est systématiquement réduit.

Pour le quantile 90, les estimations diffèrent fortement pour les nitrites et orthophosphates. Ces différences sont particulièrement importantes pour les échantillons de 12 mesures par an.

Il faudrait accompagner chaque résultat de son incertitude, calculée préalablement par station avant de l'associer à une classe de qualité. Notons d'ailleurs un point important : en de nombreuses stations et années, la conversion d'une estimation à la classe de qualité associée estompe les différences entre les méthodes, l'écart conduisant rarement à un changement de classe. En résumé, le codage par classe de couleur proposé par le SEQ-Eau reste le plus souvent valide, alors que les estimations peuvent être fortement améliorées.

Chapitre 8

Application au réseau RNB du bassin de la Moselle

Pour l'étude des corrélations spatiales, le réseau de stations de l'agence Rhin Meuse a été préféré à celui de l'agence Loire Bretagne, les stations étant plus nombreuses et plus régulièrement échantillonnées. Pour la majorité de ces stations, les mesures sont mensuelles, mais ces données ne sont pas parfaitement régulières. En particulier, il manque souvent une mesure au mois de décembre. Nous avons donc estimé par krigeage les moyennes annuelles des débits et des concentrations en nitrates, que nous assimilerons par la suite aux moyennes annuelles.

Le réseau hydrographique de la Moselle comporte 99 stations. Les variogrammes temporels moyens entre 1992 et 2003 pour ces stations sont similaires à ceux rencontrés pour le bassin Loire Bretagne :

- Pour les débits, la majorité des variogrammes (84%) comportent une composante périodique, 13% sont pépitiques, 2% sont de type « pépite + linéaire » et 1% «pépite + sphérique ». Lorsque la station ne comporte pas suffisamment de mesures, par exemple une seule année de mesures, la corrélation temporelle n'a pas été modélisée.
- Pour les concentrations en nitrates, 75% des variogrammes comportent une composante périodique, 11% sont de type « pépite + linéaire », 11% de type « Pépite + sphérique » et 3% pépitiques.

Les moyennes annuelles des débits et des concentrations en nitrates sur l'ensemble des stations entre 1992 et 2004 sont comparées aux estimations statistiques FIG. 8-1. Les différences peuvent être parfois importantes, atteignant 25% pour les débits, 15% pour les nitrates et 40% pour les flux de nitrate. Dans le cadre du SEQ-Eau, les moyennes annuelles sont calculées sur une année calendaire civile. Dans la plupart des études hydrologiques, la moyenne annuelle porte sur l'année hydrologique qui débute le 1er octobre et se termine le 30 septembre de l'année suivante ce qui permet entre autres de considérer la période hivernale en un seul bloc. Les hydrologues parlent de débit annuel dans le cas d'une année civile et de module dans le cas de l'année hydrologique. Nous appellerons année hydrologique « année A » l'année qui débute le 1er octobre de l'année A dans les chapitres suivants. Les différences entre les estimations statistiques et géostatistiques des modules annuels présentées FIG. 8-1 (à droite) sont encore plus importantes que sur les débits annuels : jusqu'à 40% d'écart pour les débits et pour les flux de nitrate. Cela s'explique par une configuration parfois inadaptée des mesures.

Prenons l'exemple de la station (2074000) pour laquelle la différence entre les estimations des débits par krigeage ou statistique est maximale et regardons la répartition des données entre 1999 et 2000 (FIG. 8-2). Des données manquent en janvier et février. Pour l'année hydrologique 1999, le poids de krigeage est très élevé au mois de décembre 1999, qui présente une très forte valeur de débit, ce qui a pour effet d'augmenter fortement la moyenne annuelle. En revanche, pour l'année civile 1999, le poids affecté au mois de décembre ne sera pas particulièrement élevé puisque tous les mois de l'année sont informés.

Différences relatives krigeage / statistique

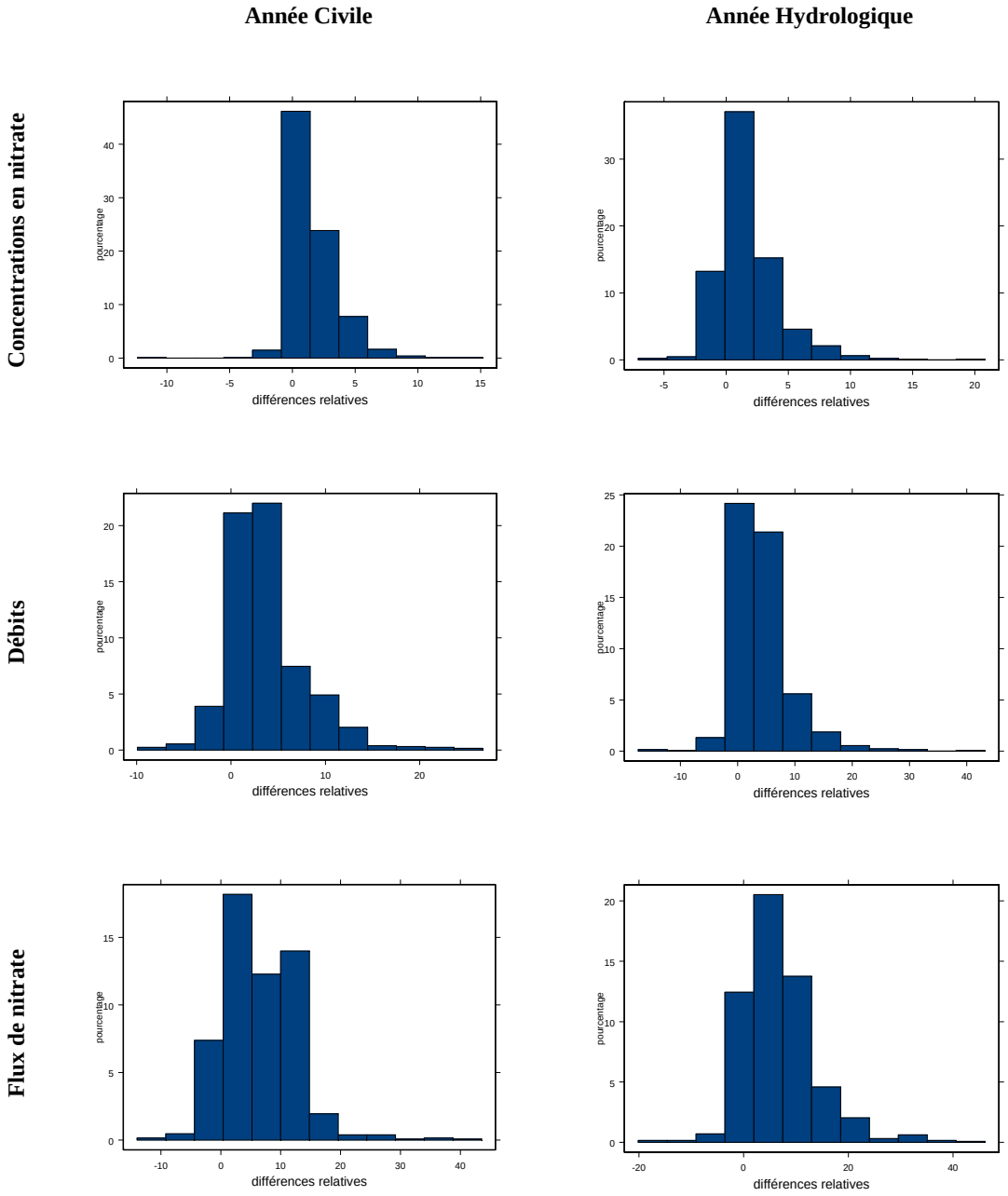


FIG. 8-1 : Histogrammes de la différence relative en pourcentage ($100 \times \frac{M_K - M_S}{M_S}$) entre les moyennes annuelles calculées par statistiques classiques M_S et par krigeage M_K. A gauche : année civile. A droite : année hydrologique. De haut en bas : concentrations en nitrate, débits, flux de nitrate.

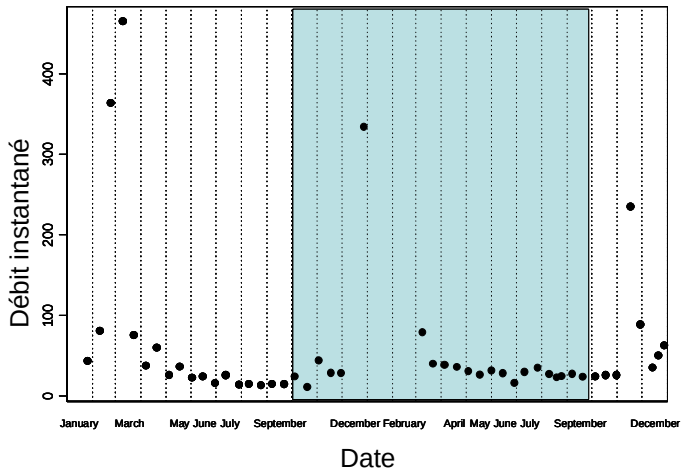


FIG. 8-2 : Stations 2074000, Répartition des mesures des débits entre 1999 et 2000. En bleu, année hydrologique.

Les différences entre les estimations de la moyenne annuelle statistiques ou géostatistiques ont des conséquences. En particulier, les techniques de régressions sont fréquemment utilisées pour mettre en relation les concentrations, les débits, ou les flux de concentrations avec des variables auxiliaires comme l'occupation du sol. Par exemple, la relation entre les modules par station et la surface drainée à l'amont de ces stations est présentée figure FIG. 8-3 pour l'année 1996 pour l'ensemble des stations du réseau hydrographique drainé par la Moselle. La modélisation sur des estimations statistiques ou géostatistiques de modules par krigeage ou statistique diffèrent assez fortement : par rapport à l'estimation statistique, le coefficient de régression linéaire augmente de 15% par le krigeage et l'intercept est réduit de 68%.

Pour l'étude des corrélations spatiales des modules annuels le long des cours d'eau, nous assimilons les valeurs estimées des modules annuels à leur valeur réelle. Cette approximation est importante, puisque l'écart type de krigeage sur l'ensemble des stations et des années est en moyenne de 10% pour les nitrates, de 38% pour les débits et de 31% pour les flux de nitrates. Cet écart type est un résultat en moyenne sur l'ensemble des chroniques annuelles rencontrées, qui comprend donc à la fois des échantillons réguliers ou irréguliers, mais rarement moins de 10 mesures. Selon l'année et la station, les incertitudes diffèrent fortement, comme le montre l'histogramme des écarts types d'estimation FIG. 8-4. Les incertitudes sur les débits et flux de nitrate sont plus importantes que pour les concentrations en nitrates. Cet écart est dû au fait que les débits et flux de nitrates présentent des chroniques moins structurées que les concentrations en nitrates, des « pics » de valeurs fortes et varient fortement d'une année à l'autre.

L'incertitude sur l'estimation des débits et des flux de nitrate pourrait être réduite en utilisant les stations hydrométriques qui mesurent les débits par jour. L'incertitude sur les débits serait alors négligeable et celle sur les flux de nitrate fortement réduite. En effet, (Moatar and Meybeck, 2005) ont montré expérimentalement que le biais et les incertitudes sur les flux sont diminués en combinant l'information continue sur les débits et une interpolation linéaire des concentrations entre les mesures aux stations. Par exemple pour la station 50500 à Orléans en 1981, l'écart type de précision expérimental relatif est réduit de 13% à 3% pour les flux de nitrate. Dans la même idée, des techniques géostatistiques peuvent être appliquées, comme le krigeage avec dérive externe. Il serait d'ailleurs

intéressant de les comparer à celles utilisées dans (Moatar and Meybeck, 2005) afin d'observer la pertinence de l'écart type d'estimation annoncé par krigeage.

Ces données auraient pu être prises en compte dans l'étude spatiale des débits en échantillonnant les stations hydrométriques à même fréquence que les stations RNB. On aurait ainsi plus de stations à disposition entre deux confluences. Cela n'a pas été fait, les données étant arrivées tardivement.

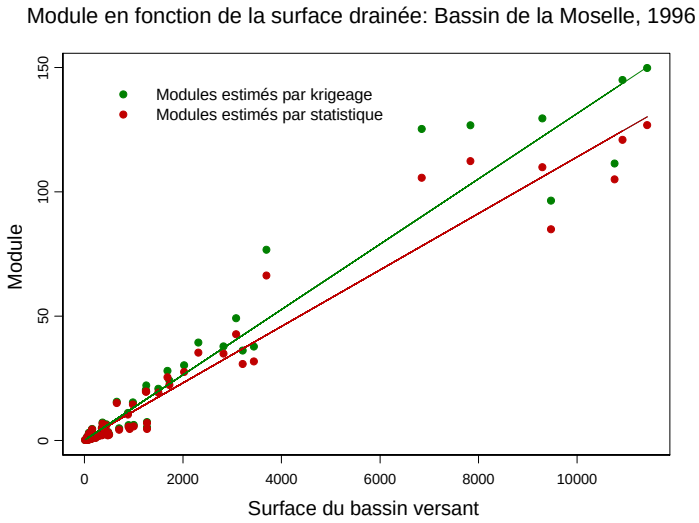


FIG. 8-3 : Régression linéaire du module 1996 sur le bassin de la Moselle en fonction de la surface drainée. En vert, les modules ont été estimés par krigeage, en rouge, par statistique classique.

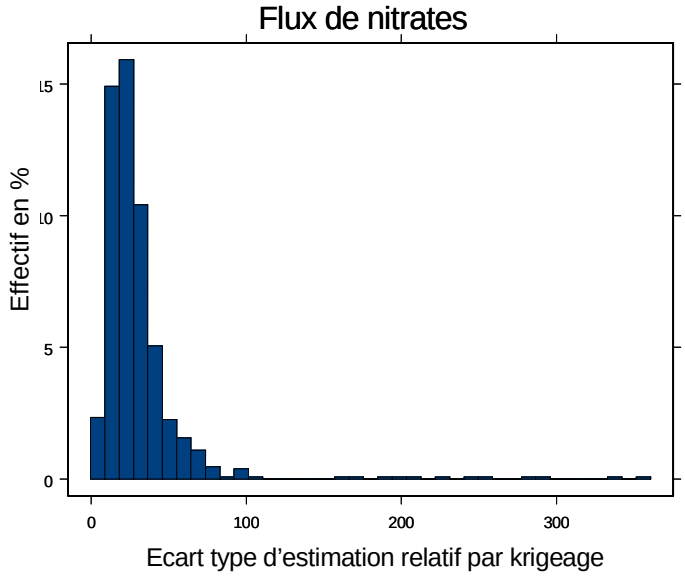


FIG. 8-4 : Ecart type d'estimation relatif (Ecart type d'estimation estimé divisé par la moyenne annuelle estimée en %) des flux de nitrate sur l'ensemble des stations du bassin de la Moselle entre 1996 et 2001.

Partie III
Indicateurs de qualité entre stations

Les modèles de corrélation usuels, développés en géostatistique sur des espaces euclidiens ne sont pas valides sur des structures arborescentes. De nouveaux modèles doivent donc être développés avec les nombreuses questions que pose l'étude de fonctions aléatoires sur un réseau hydrographique : quelle distance choisir ? Comment gérer les discontinuités aux confluences ? Après une analyse expérimentale des débits, des concentrations en nitrate et des flux de nitrate, un modèle général de covariance et variogramme est proposé. En tout point du réseau, le cours d'eau est considéré comme la somme de « filets d'eau élémentaires » définis de la source à l'exutoire. La pertinence du modèle est étudiée au travers d'un exemple d'application pour les débits spécifiques. On cherche principalement à vérifier les hypothèses du modèle. La question de l'inférence du modèle est discutée.

Chapitre 9

Position du problème et premières analyses expérimentales sur les données.

Une fois les indicateurs calculés par station, la question se pose de savoir comment les interpoler le long des cours d'eau, ce qui revient en fait à estimer les indicateurs en tout point du réseau. Ceci nécessite de modéliser la corrélation spatiale de la variable étudiée. De nouveaux modèles doivent être développés, les modèles usuels n'étant plus valables sur des supports arborescents (Ver Hoef et al., 2006). Des questions se posent: quelle distance considérer ? Comment tenir compte de la discontinuité aux confluences ? Y a-t-il indépendance des mesures sur les différents affluents à l'amont des confluences ? Les variables étudiées peuvent elles être supposées stationnaires ?

Nous présentons dans ce chapitre une étude expérimentale dans l'objectif de guider la modélisation qui sera présentée dans le chapitre suivant.

9.1 Problèmes posés le long des réseaux

Pour estimer des débits, concentrations ou flux moyens annuels le long d'un réseau hydrographique, il est essentiel de prendre en compte la géométrie du réseau. Or, les modèles géostatistiques usuels ont été développés pour des espaces euclidiens. En effet, les théorèmes de Bochner et de Schoenberg font explicitement intervenir la distance euclidienne dans la caractérisation spectrale des covariances et des variogrammes (Chilès and Delfiner, 1999).

Ver Hoef et al.(2005) démontrent à travers un exemple simple d'arbre binaire à 63 nœuds que le choix d'une covariance sphérique ou linéaire conduit à des matrices de covariance non définies positives. D'où la nécessité de développer de nouveaux modèles de covariance le long des graphes reposant sur des hypothèses cohérentes avec la réalité. Dans ce manuscrit, nous étudions à la fois les débits et les nitrates dans l'objectif de mettre en place un modèle de corrélation pour les flux de concentrations en nitrates.

L'étude d'une variable le long d'un graphe soulève les questions suivantes :

- **Quelle distance choisir ?** Habituellement, la distance euclidienne est choisie pour décrire les corrélations en fonction de la distance. Le long d'un graphe, différentes distances peuvent être envisagées : la distance le long du graphe non orienté pour laquelle tous les couples de points sont considérés qu'ils soient à l'amont-aval l'un de l'autre ou non, ou la distance le long du graphe orienté ou seuls les couples de points au fil de l'eau sont considérés. Par exemple, le couple de points formé par les stations 2077500 et 2070500 (voir FIG. 9-1) est considéré dans le calcul du variogramme sur un graphe non orienté, mais ne l'est pas pour un graphe orienté. Quelle est l'influence de la distance sur l'analyse des corrélations spatiales ? Quelles hypothèses se cachent en réalité derrière ce choix ?
- **Stationnarité ?** La géostatistique repose la plupart du temps sur une hypothèse de stationnarité. Une covariance est dite stationnaire si elle ne dépend que de la distance curviligne entre deux points du graphe et non stationnaire lorsqu'elle dépend des deux points séparément. Cette hypothèse est elle valable pour les concentrations et les débits ?

- **Confluences ?** Comment gérer la discontinuité aux confluences ? Pour résoudre le problème de la continuité aux confluences et en particulier de la non stationnarité qu'elle engendre, certains auteurs (Ver Hoef et al., 2006) proposent une pondération des fonctions aléatoires considérées avec des poids choisis de manière à rendre la fonction aléatoire finale stationnaire sur l'ensemble du réseau hydrographique. Mais cette pondération n'est pas toujours cohérente avec la réalité physique que les paramètres doivent respecter. Les débits sont par exemples additifs aux confluences : un débit immédiatement à l'aval d'une confluence est égal à la somme des débits immédiatement à son amont. De même pour les flux de concentrations sous réserve de la conservation de la matière au fil de l'eau.

Nous proposons une analyse exploratoire non seulement pour répondre à ces questions, mais surtout pour nous guider vers un modèle statistique cohérent avec les données. Ce modèle est décrit dans le chapitre suivant (Chapitre 10), après avoir présenté les différents modèles actuels.

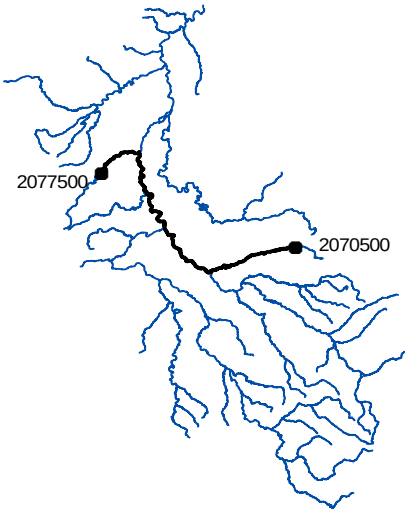


FIG. 9-1 : Couple de stations non reliées au fil de l'eau sur le réseau hydrographique simplifié de la Moselle.

9.1.1 Choix de la distance

Dans le cas d'un réseau hydrographique, le question se pose de savoir comment mettre en évidence les corrélations spatiales (Gardner et al., 2003 ; Ganio et al., 2005). Pour calculer les variogrammes des débits, des concentrations en nitrate et des flux de concentrations sur le réseau hydrographique composé par la Moselle, trois distances sont considérées :

- La distance usuelle euclidienne 2D, qui ne tient pas compte de l'arborescence
- La distance curviligne 1D, définie sur le graphe orienté. Pour deux points non reliés au fil de l'eau, la distance est infinie.
- La distance curviligne définie 1D sur le graphe non orienté.

Le variogramme est alors estimé par la formule usuelle :

$$\overline{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}\left[Z(x_\alpha) - Z(x_\beta)\right]^2 \text{ avec } N(h) = \left\{(\alpha,\beta), x_\alpha - x_\beta = h\right\}$$

Afin d'interpréter les différences entre les résultats obtenus, le variogramme expérimental calculé uniquement à partir des couples de points non connectés au fil de l'eau a été ajouté. Cela permet de comparer la variabilité inter cours d'eau par rapport aux variations le long du cours d'eau.
Les variogrammes sont présentés pour l'année 1995 avec un pas égal à 20km. Les conclusions sont similaires pour l'ensemble des années étudiées avec toutefois des amplitudes de variation différentes.

Relation entre les distances

La position relative de chacun des variogrammes (FIG. 9-3, FIG. 9-5 et FIG. 9-6) s'explique aisément. Les différents variogrammes sont calculés à partir des mêmes écarts de moyennes annuelles, mais ces écarts ne sont pas associés à la même distance. Par exemple une distance euclidienne courte correspond parfois à deux points proches, sur des rivières différentes dont la confluence est éloignée, ce qui implique une distance curviligne importante entre les deux points. La figure FIG. 9-2 présente le nuage de corrélation de la distance curviligne calculée le long d'un graphe non orienté en fonction de la distance euclidienne. Afin de rendre lisible le graphique, le nuage est présenté pour des distances moyennes calculées par groupe de 100 couples, les quantiles 5% et 95% sont présentés sous la forme d'un intervalle entourant chaque valeur. La relation est linéaire avec une pente importante jusqu'à 40km. Après un palier, elle augmente à nouveau avec une pente plus faible. Les quantiles 5% croient linéairement alors que les quantiles 95% restent constants entre 20 et 80km

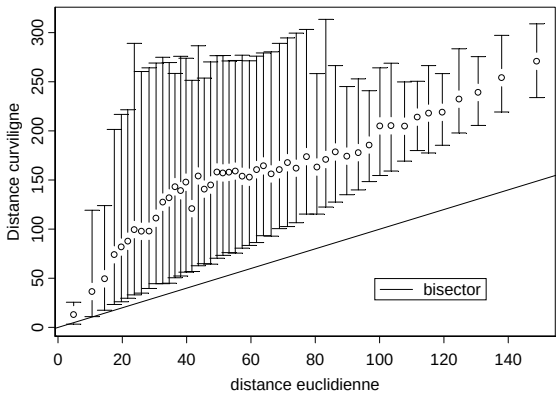


FIG. 9-2 : nuage de corrélation entre la distance euclidienne et la distance curviligne calculée le long d'un graphe non orienté. Résultats moyens par groupe de 100 couples

Débits

Les débits augmentent de l'amont vers l'aval. La figure (FIG. 9-3) présente les modules annuels en fonction de la surface drainée (en bas à gauche) ou de la distance à l'exutoire (à droite). Le nuage de corrélation des débits en fonction de la distance à l'exutoire montre deux ensembles, l'un très fortement décroissant, l'autre avec des faibles valeurs presque constantes en fonction de la distance. Cette configuration du nuage s'explique par la confluence de la Moselle avec la Meurthe à environ 115 km de l'exutoire. A l'aval de cette confluence, donc près de l'exutoire, pour une distance donnée, se situent des stations sur la Moselle, donc avec de forts débits et d'autres sur de petits affluents de la Moselle, donc avec de faibles débits (voir FIG. 9-3 en haut à droite, les stations A et B). Cette séparation du nuage n'intervient pas lorsque l'on exprime les débits en fonction de la surface drainée

ou de la distance à la source principale, qui lui est fortement liée (FIG. 9-4). Un point du réseau ayant généralement plusieurs sources, il est préférable de retenir comme critère d'analyse des variables la surface drainée qui englobe les contributions de toutes les sources.

Le variogramme (FIG. 9-3 en haut à gauche) calculé pour la distance curviligne sur l'arbre orienté est fortement non stationnaire. Les distances fortes correspondent alors à des couples de points dont l'un est proche de l'exutoire et l'autre proche d'une source, donc à une différence importante de débit. A l'inverse, dans le cas de l'arbre non orienté, les plus grandes distances correspondent en général à des couples de points proches de sources, donc avec des débits faibles et une différence entre ces débits faible aussi. La distance le long d'un graphe non orienté a donc tendance à affaiblir l'amplitude du variogramme. Le variogramme calculé sur le graphe orienté est supérieur à celui calculé sur le graphe non orienté, lui-même supérieur à celui calculé sur les couples de points non connectés : les variations inter rivières sont moins importantes que les variations le long du cours d'eau lui-même. Une étude géostatistique directement sur les débits annuels moyens n'est pas adaptée et c'est pourquoi nous proposons par la suite de prendre en compte leur relation avec la surface drainée.

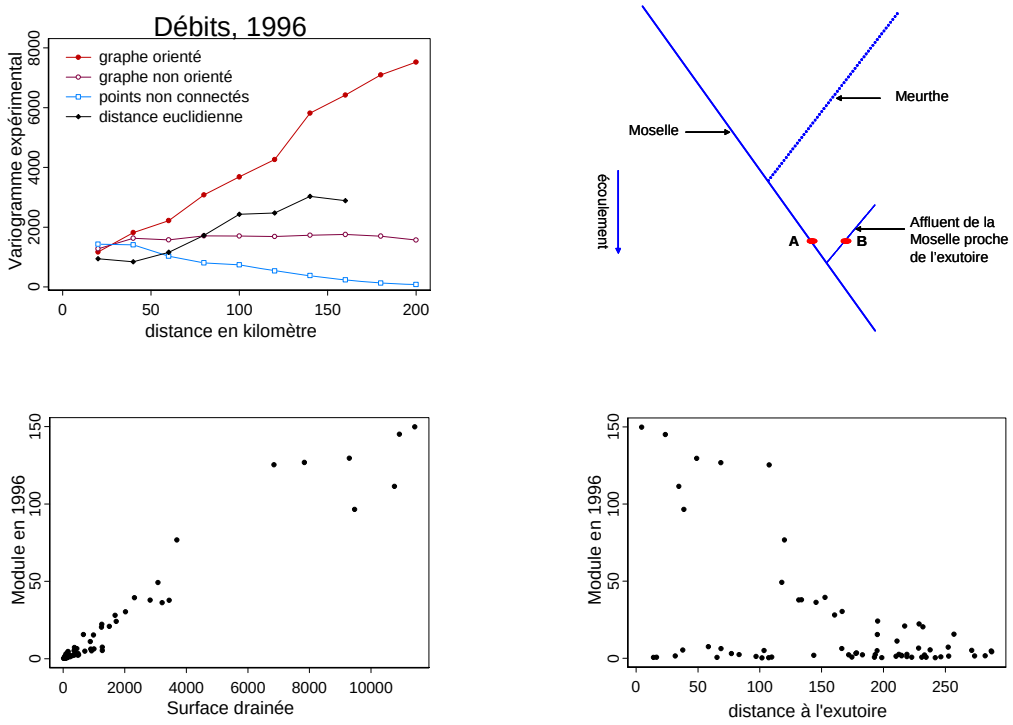


FIG. 9-3 : En haut, à gauche, variogrammes expérimentaux pour les débits pour différentes distances. En haut, à droite : exemple pédagogique. En bas, à gauche : modules en fonction de la surface drainée. En bas, à droite : modules en fonction de la distance à l'exutoire.

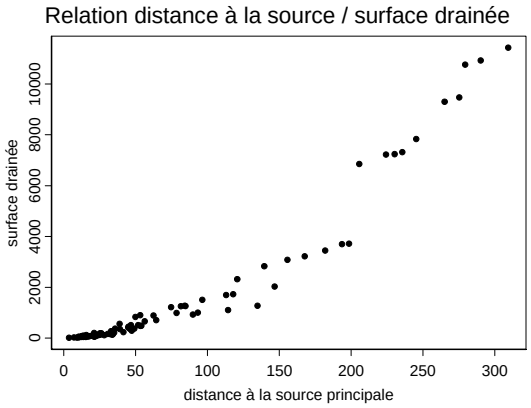


FIG. 9-4 : Relation entre surface drainée et distance à la source principale

Concentrations en nitrates

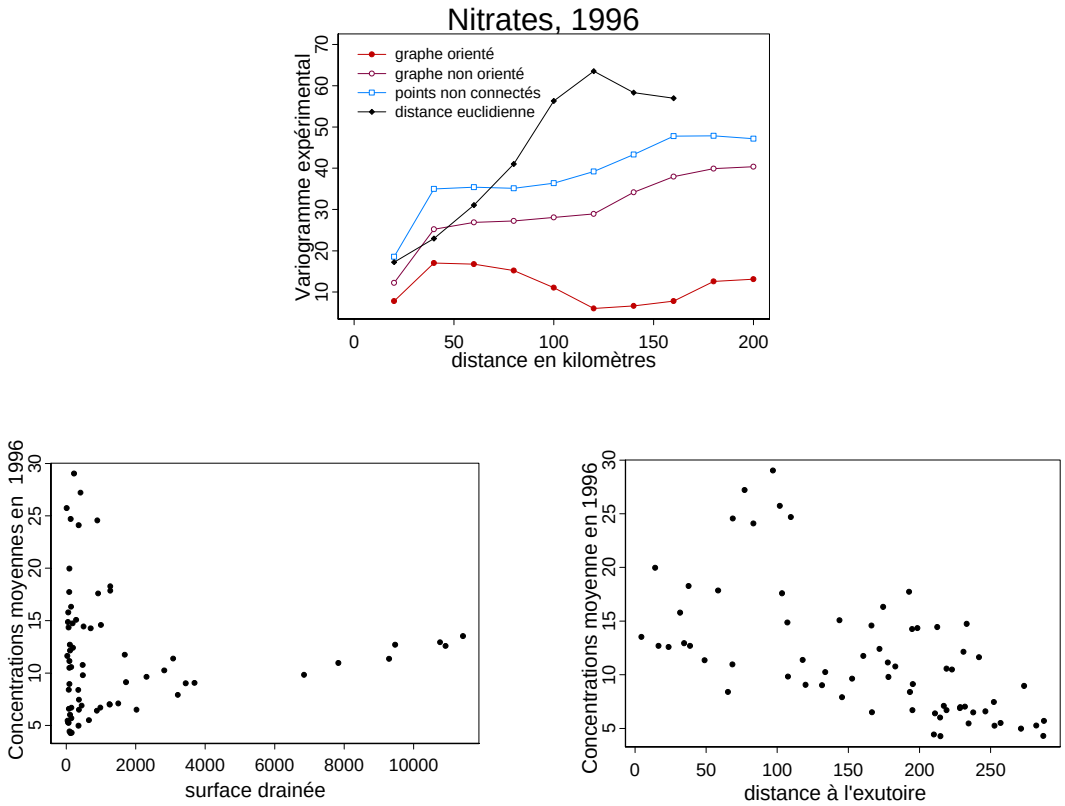


FIG. 9-5 : En haut, variogrammes expérimentaux pour les nitrates pour différentes distances. En bas, à gauche : concentrations moyennes en nitrates en fonction de la surface drainée. En bas, à concentrations annuelles moyennes en fonction de la distance à l'exutoire.

Les nitrates (FIG. 9-5) augmentent de l'amont vers l'aval mais de manière moins prononcée que les débits. Les variogrammes montrent que les variations entre rivières sont ici plus importantes que le long du cours d'eau, le variogramme pour le graphe orienté étant inférieur à celui calculé pour le graphe non orienté. Tous les variogrammes montrent un palier atteint à environ 40km sauf pour la distance euclidienne où le palier ne semble atteint qu'à 120 km, ce qui pourrait s'interpréter comme le fait que les concentrations sont plus liées via les propriétés du sol que par le transport de polluant le long de la rivière. Le variogramme est d'ailleurs quasi pépitique pour la distance curviligne sur arbre orienté. L'indépendance des concentrations sur des rivières parallèles n'est pas vérifiée. Comme pour les débits, il ne semble pas approprié de travailler directement sur les concentrations mais de prendre en compte dans un premier temps leur relation avec l'occupation du sol. Il est d'ailleurs montré que les nitrates sont fortement liés au pourcentage d'agriculture et de forêt. La non stationnarité est une propriété du modèle qu'il est délicat de valider par les données.

Flux de nitrates

La croissance des débits le long des cours d'eau étant beaucoup plus importante que celle des nitrates, les flux de nitrates ont un comportement similaire à celui des débits. Leur analyse variographique conduit donc à des conclusions similaires.

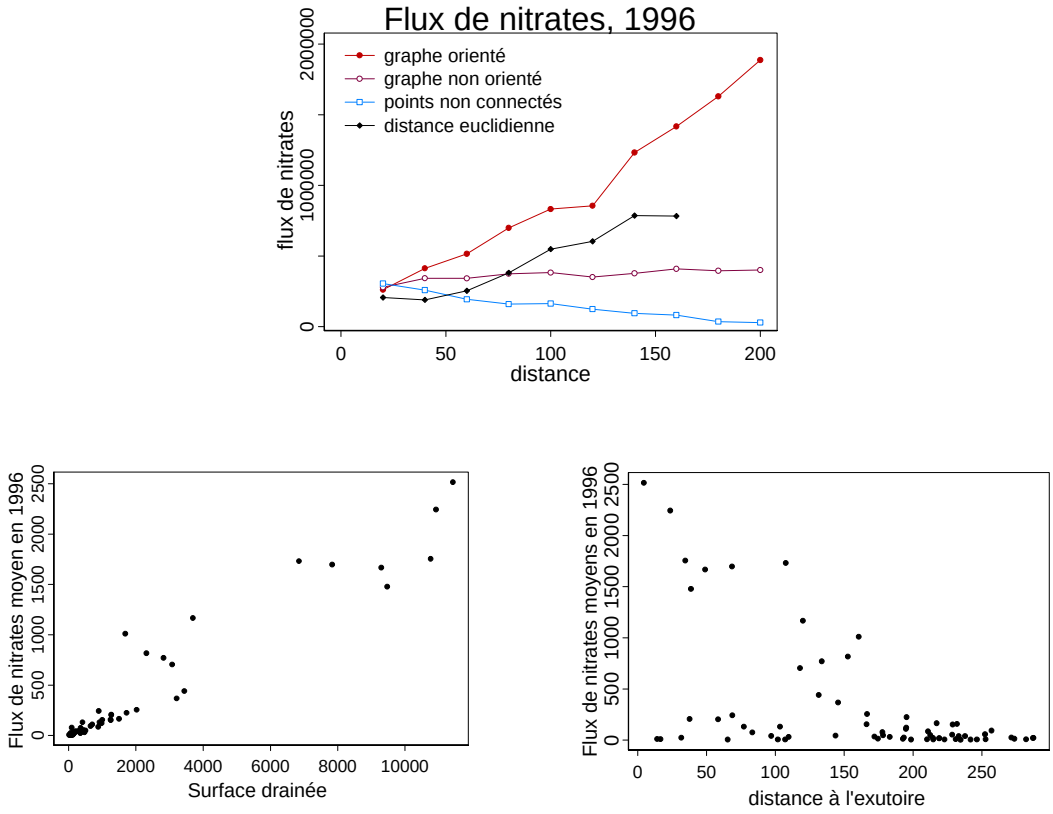


FIG. 9-6 : En haut : variogrammes expérimentaux des flux de nitrates pour différentes distances. En bas, à gauche : flux de nitrates moyens en fonction de la surface drainée. En bas, à droite : flux de nitrates moyens en fonction de la distance à l'exutoire.

9.1.2 Stationnarité

L'étude expérimentale des débits, des concentrations en nitrates et des flux de nitrate montre que pour aucune de ces variables la fonction aléatoire associée ne peut être supposée stationnaire sur l'ensemble du réseau. Pour se ramener à l'étude de fonction «stationnaires», il faut prendre en compte les relations entre ces paramètres et des variables explicatives, comme la surface drainée pour les débits et les concentrations en nitrates liées à l'occupation du sol (forêt, agriculture). Ces relations expérimentales sont présentées dans ce paragraphe.

Débits

Les débits dépendent essentiellement de la surface drainée. Une relation simplifiée, non linéaire est communément utilisée par les hydrologues pour décrire des débits moyens sur une période supérieure au mois (Strahler, 1964):

$$D_i = f(S_i) = c(S_i)^d$$

avec D_i le débit à la station *i*, S_i le bassin versant drainé à la station *i.*

Cette relation est vérifiée sur le bassin de la Moselle (FIG. 9-7). On distingue deux parties correspondant à la confluence de la Meurthe et de la Moselle. Selon l'année étudiée, les paramètres du modèle diffèrent significativement et dans certains cas la relation est purement linéaire.

Les paramètres des modèles sont estimés par régression linéaire ou non linéaire avec le choix d'un modèle à résidus multiplicatifs ou additifs :

- **Le modèle multiplicatif** $D_i = f(S_i) \times R_i$ est celui usuellement utilisé par les hydrologues et les statisticiens car il permet un calcul facile des coefficients par passage au logarithme ; il explique l'augmentation de la variance des débits en fonction de la surface drainée.

  Ce modèle doit être utilisé avec précaution, car d'une part, on ne sait pas si l'erreur sur les mesures est additive ou multiplicative, d'autre part, l'estimateur des débits par passage au log est biaisé. Si on suppose les résidus de la forme e^R avec $R \sim N(m,\sigma)$, alors :

  $E\left(\hat{D}_i\right) = E(D_i) \times e^{(Var(\hat{g}+\hat{d}\log(D_i))/2)-\sigma^2/2}$ où $g = \log(c)$, $\hat{D}_i$, $\hat{g}$ et $\hat{d}$ estimateurs de D_i, g et d

  Le calcul de ce biais nécessite la connaissance des corrélations entre les estimateurs des coefficients de régression $\hat{g}$ et $\hat{d}$ et la loi des débits, ce dont nous ne disposons pas. Travailler sur le logarithme des débits directement peut être envisagé par un krigeage log normal. Le biais des estimations pour un krigeage à moyenne connue est alors connu et égal à $\frac{\sigma^2 - \sigma_{KS}^2}{2}$ où σ_{KS}^2 est la variance de krigeage.

- **Le modèle additif** $D_i = f(S_i) + R_i$ est utilisé dans l'hypothèse d'homoscédasticité, c'est-à-dire de variance constante, qui ici n'est pas respectée. On peut dans ce cas utiliser une régression pondérée $Q_i = f(S_i) + g(S_i)R_i$ où $1/g(S_i)$ est une pondération, fonction de la surface drainée qui exprime la variance des résidus.

La décomposition d'un phénomène en deux composantes, l'une déterministe appelée dérive, l'autre aléatoire appelée résidu, est connue sous le nom de krigeage universel. Il soulève deux difficultés :

- Réaliser l'estimation optimale de la dérive

- Le variogramme théorique du résidu est biaisé. Ce problème est moins grave qu'il n'y parait pourvu que l'on soit près à délaisser le variogramme traditionnel (Marcotte and Powojowski, 2004).

Un moyen de contourner ces difficultés est d'étudier directement les débits spécifiques D_i/S_i qui correspondent à la hauteur d'eau répartie sur la surface drainée par unité de temps. Il faudra cependant vérifier que les hypothèses du modèle sont cohérentes pour cette variable.

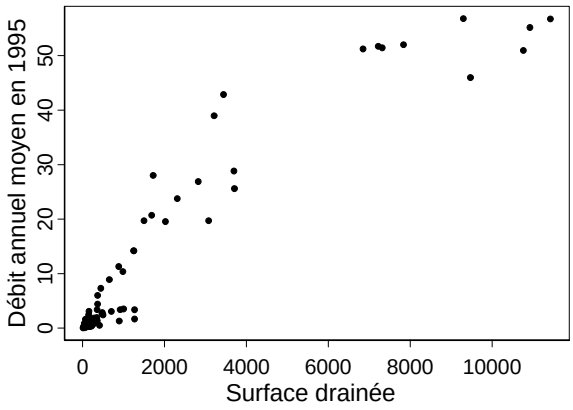


FIG. 9-7 : Année hydrologique 1995. Bassin de la Moselle. Relation entre les débits et la surface drainée.

Concentrations en nitrates

Les moyennes annuelles des concentrations en nitrate dépendent essentiellement de l'agriculture et des forêts (de Fouquet and Bez, 2001). Un exemple présentant ces relations est donné FIG. 9-8. Le pourcentage de forêts et d'agriculture a été calculé sur l'ensemble de la surface drainé à l'amont d'un point à partir des données Corine Land Cover de l'IFEN. Ces relations peuvent être prises en compte par un modèle multivariable (Chilès and Delfiner, 1999).

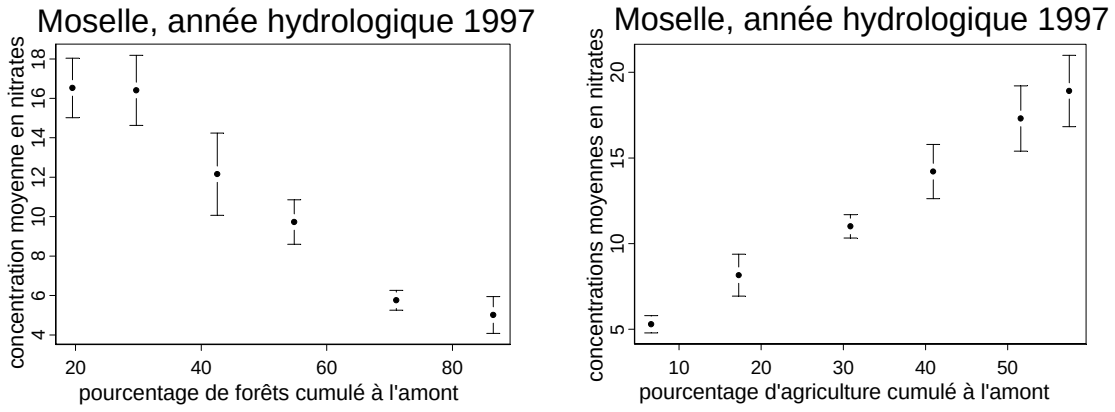


FIG. 9-8 : Moselle, 1997. A gauche, relation entre les concentrations moyennes en nitrates et le pourcentage de forêts cumulé à l'amont. A droite, relation entre les concentrations moyennes en nitrates et le pourcentage d'agriculture.

Flux de nitrates

Les flux de nitrates s'expliquant essentiellement par les débits sont liés comme ces derniers à la surface drainée amont. Deux solutions sont envisageables pour se ramener à une fonction stationnaire : étudier les résidus de la régression des flux de nitrate par les débits ou considérer des « flux spécifiques » qui correspond au rapport des flux par la surface drainée. Un modèle multivariable prenant en compte à la fois la surface drainée, le pourcentage d'agriculture et de forêts serait encore plus approprié.

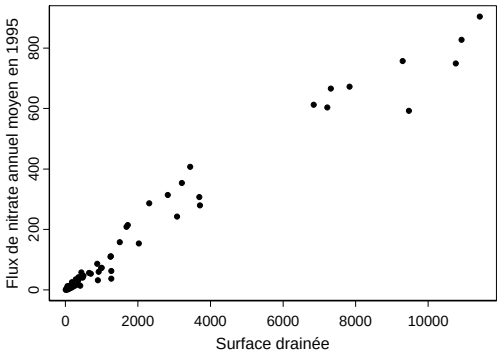


FIG. 9-9 : Année hydrologique 1995. Bassin de la Moselle. Relation entre les flux de nitrate et la surface drainée

Cette troisième partie ne vise pas à fournir un modèle complexe des débits, des flux de nitrates ou des concentrations en nitrates prenant en compte tous les paramètres influents. De nombreux modèles phénoménologiques ont été développés par les agences (PEGASE, DECLIC, SENEQUE). L'objectif se limite à la construction d'un modèle de covariance valide pour ces paramètres et la question de son inférence. Parce que les débits, les flux de nitrates et les concentrations en nitrates présentent en général une non stationnarité très marquée (Dent and Grimm, 1998), nous avons cherché leur lien avec d'autres variables afin de se ramener à une variable d'étude « plus stationnaire ».

9.1.3 Conditions de cohérence aux confluences

La construction d'un modèle de fonctions aléatoires le long d'un réseau hydrographique doit nécessairement respecter les conditions suivantes :

- Les débits sont additifs aux confluences : à l'aval immédiat d'une confluence, le débit est égal à la somme des débits à l'amont immédiat de la confluence. Notons D_3 le débit à la confluence et D_1, D_2 les débits immédiatement à l'amont. Alors la relation (9-1) doit être respectée.

$$D_3 = D_1 + D_2 \quad (9\text{-}1)$$

- Les flux sont additifs aux confluences sous l'hypothèse de non dégradation de la matière le long des cours d'eau.
- Par déduction de l'additivité des flux, à la confluence, une concentration est égale à la somme des concentrations amont pondérées par les débits:

$$C_3 = \frac{D_1}{D_1 + D_2} C_1 + \frac{D_2}{D_1 + D_2} C_2 \,,\ C_i \text{ concentration au point i} \quad (9\text{-}2)$$

Sous ces contraintes, on ne se ramène pas à l'étude de variables stationnaires sur l'ensemble du réseau hydrographique.

Prenons l'exemple simple d'une confluence. La plupart des modèles de régression proposés au paragraphe précédent, associés à la condition d'additivité des débits, conduit à l'étude de fonctions aléatoires non stationnaires :

- Le modèle de régression linéaire simple $D_i = cS_i + R_i$ conduit à la condition suivante sur les résidus $R_3 = R_1 + R_2$ qui implique que l'espérance, si elle n'est pas nulle, et la variance des résidus ne sont pas constantes. Les résidus se somment à chaque confluence avec le même poids égal à 1. Dans ce raisonnement, les résidus sont considérés comme un écart réel entre la vraie valeur du débit et la valeur prédite par le modèle de régression.
- Le modèle de régression non linéaire multiplicatif $D_i = cS_i^d \times R_i$ conduit à la condition sur les résidus $R_3 = \frac{S_1^d}{(S_1 + S_2)^d} R_1 + \frac{S_2^d}{(S_1 + S_2)^d} R_2$. A chaque confluence, le résidu est une somme pondérée des résidus à l'amont. L'espérance et la variance des résidus ne sont pas constantes. Notons que faire l'hypothèse d'indépendance des mesures avant confluence et de stationnarité implique que d est égal à ½, ce qui n'est absolument pas vérifié sur les données.
- Pour les débits spécifiques, à chaque confluence, le débit spécifique doit vérifier la relation (9-3).

$$T_3 = \frac{S_1}{S_1 + S_2} T_1 + \frac{S_2}{S_1 + S_2} T_2 \text{ où } T_i = {D_i}/{S_i} \text{ est le débit spécifique (9-3)}$$

Nous verrons au Chapitre 10 comment s'intègre la pondération dans le modèle général de corrélation développé pour les graphes.

Maintenir la stationnarité aux confluences par une pondération des fonctions aléatoires définies sur chaque bief (Ver Hoef et al., 2005) n'est donc pas toujours justifié car ces pondérations ne sont pas toujours réalistes. Pour les concentrations, par exemple, le raisonnement est similaire aux débits. A la confluence, une concentration est égale à la somme des concentrations amont pondérées par les débits.

Dans ce cas, la somme des poids est égale à 1, ce qui conduit une espérance et variance croissantes et donc à un processus non stationnaire. Cela contredit le choix de pondérer les concentrations à l'amont d'une confluence par des fonctions de la surface drainée ou de l'ordre du cours d'eau dont la somme des carrés est égale à 1 (Ver Hoef , 2005 ; Cressie et al., 2006).

Cependant prendre en compte les relations de la variable étudiée avec des facteurs d'influence permet de diminuer la non stationnarité sur l'arbre et éventuellement de se ramener à un phénomène stationnaire par bief avec indépendance à l'amont d'une confluence.

Chapitre 10

Modélisation des flux et des débits le long des cours d'eau

Une synthèse bibliographique est présentée sur les modèles géostatistiques à supports arborescent. Un modèle simple de covariance est ensuite proposé et appliqué aux débits spécifiques. De manière pratique, il paraît peu réaliste de ne s'appuyer que sur des mesures pour estimer les moyennes annuelles d'une substance polluante ou d'un flux sur l'ensemble du réseau. Le nombre de mesures étant insuffisant pour une interpolation (moins d'une mesure par bief), l'idéal serait de prendre en compte des variables exogènes telles que la pluie, la surface drainée, l'occupation du sol comme le font les modèles phénoménologiques tels que PEGASE, SENEQUE ou DECLIC. Jusqu'à présent, même si elle n'est pas explicite dans les rapports, l'ensemble de ces modèles reposent sur une hypothèse d'indépendance spatiale des mesures. Cette indépendance est par exemple sous entendue par l'utilisation de régressions statistiques entre les mesures et les paramètres d'influence. Il serait donc intéressant d'intégrer la corrélation spatiale et les techniques géostatistiques à ces modèles pour les caler aux mesures, améliorer les estimations mais surtout leurs intervalles de confiance.

Faute de disposer d'un tel modèle phénoménologique, notre étude privilégie la modélisation des corrélations spatiales à l'estimation elle-même.

10.1 Bibliographie

Les quelques travaux publiés sur la prise en compte de la géostatistique dans la modélisation de concentrations de polluants le long d'un réseau hydrographique sont récents. Par ailleurs, ils ne sont jamais liés aux modèles phénoménologiques utilisés par les agences qui eux ne prennent pas en compte les corrélations spatiales. Il est donc intéressant de commencer par présenter rapidement les articles principaux sur les modèles développés pour des supports arborescents.

De façon générale, on distingue quatre catégories d'articles:

- Des problèmes similaires sur la construction de fonctions noyaux ou de covariance sur des graphes sont en développement (Kondor and Lafferty, 2002 ; Smola and Kondor, 2003 ; Vert et al., 2004 ; Kondor and Vert, 2004) avec en particulier des applications en bioinformatique. Nous ne les présenterons pas ici, car les variables étudiées sont définies uniquement sur des nœuds.
- Plusieurs modèles phénoménologiques sont développés pour les agences de l'eau par diverses universités afin de rendre compte de l'état des cours d'eau pour différents paramètres physiques, chimiques ou biologiques. Ces modèles proposent une modélisation basée sur un nombre considérable de facteurs d'influence et une connaissance approfondie de l'hydrologie. Mais il est difficile de trouver la documentation scientifique permettant de retracer les équations utilisées dans le modèle. De plus, ces modèles ne prennent pas en compte les corrélations spatiales dans les estimations. Ils ont pour intérêt dans notre cas d'étude, de nous guider sur les relations à prendre en compte dans la modélisation.
- Des modèles hydrologiques pour les débits avec prise en compte des corrélations spatiales par la géostatistique et s'appuyant sur la structure de graphe du réseau hydrographique. Pour la plupart,

l'objectif n'est pas de calculer un débit sur le linéaire du cours d'eau mais de ramener les débits à une surface.

- Des modèles géostatistiques pour divers paramètres étudiés sur un réseau hydrographique. L'objectif est de construire des modèles de covariance valides pour ce type de support et de les utiliser pour une estimation par krigeage.

Le détail bibliographique et une description brève sont donnés dans les paragraphes suivants.

10.1.1 Approche Agences de l'eau

Les différents modèles développés pour les agences dans l'objectif d'évaluer la qualité des cours d'eau (passée, présente ou future) sont présentés en grande majorité dans un rapport de l'INERIS de J-M Brignon (2004). Les deux modèles principaux sont SENEQUE pour le bassin Seine Normandie et PEGASE.

10.1.1.1 Pégase

Le modèle mathématique PEGASE, planification et gestion de l'assainissement des eaux, a été développé par le centre d'Etudes et de Modélisation de l'Environnement de l'Université de Liège (Smitz et al., 1997 ; Everbecq et al., 2001). Il propose le calcule prévisionnel et déterministe de la qualité des eaux en fonction des apports et des rejets de polluants, dans des conditions hydrologiques diverses et permet par le biais de simulations de comparer différents scénarios d'assainissement et de dépollution afin d'en dégager des politiques optimales en matière de gestion des eaux par les agences. C'est un modèle très complet qui référence la structure d'arbre du réseau hydrographique, calcule les bassins versants en tout point d'une rivière au kilomètre près et contient une représentation structurée des rejets urbains, industriels, des réseaux de collecteurs, des stations d'épuration, des rejets dus aux activités d'élevage, des apports diffus des sols. Il permet d'obtenir sur un bassin versant déterminé, sur un linéaire de cours d'eau ou sur une « masse d'eau », une évaluation des rejets y compris diffus, une simulation des impacts liés à diverses actions politiques, mais aussi des cartes de qualité s'appuyant sur les gilles du SEQ-Eau. Il serait très intéressant de se baser sur ce modèle pour étudier les corrélations spatiales des débits, concentrations ou flux afin de les prendre en compte dans le modèle lui-même. Mais cette démarche a posé plusieurs difficultés : d'une part, seules des cartes de résultats de scénarios pouvaient nous être fournies par les agences alors que nous ne disposons pas techniquement de l'ensemble des informations qui sont utilisées pour le calcul de ces estimations. De plus, la documentation scientifique décrivant les modèles mathématiques utilisés est rare, ce que confirme J-M Brignon dans son rapport de bilan des outils utilisés pour la Directive Cadre (Brignon, 2004). Cela rend difficile l'utilisation de ce modèle pour tester un éventuel gain de la géostatistique dans les simulations et l'incertitude qui leur est associée.

10.1.1.2 Senèque

Par ailleurs, le PIREN Seine a développé le modèle Senèque (Ruelland, 2004) qui simule la qualité physico chimique des cours d'eau. Proches dans leur fonctionnement et leur conception, les modèles Pégase et Senèque sont complémentaires, Pégase étant plus tourné vers la caractérisation quantitative de la qualité de l'eau, Sénèque vers la compréhension des mécanismes hydrobiochimiques à l'échelle du bassin versant.

10.1.1.3 Autres modèles phénoménologiques

D'autres modèles sont utilisés par les agences de l'eau. Pour la linéarisation des débits, le modèle DECLIC (Francois, 1997) développé par le Centre d'Etudes géographiques de l'université de Metz est utilisé par l'Agence de l'eau Rhin Meuse. Il permet d'estimer les valeurs d'écoulement caractéristiques et débits d'étiage avec une densité de points beaucoup plus élevée que celle fournie par le réseau hydrométrique. Les résultats de cette méthode sont publiés dans des catalogues et actuellement mis en ligne sur le site de l'agence Rhin Meuse.

Enfin, le modèle NOPOLU du Beture-Cerec propose un outil de linéarisation des débits et classes de qualité SEQ-Eau, basé uniquement sur des observations et utilisant des fonctions d'interpolations simples (interpolation linéaire pour les débits par exemple).

10.1.2 Modélisation géostatistique.

10.1.2.1 Par les hydrologues

Aucun des modèles utilisés par les agences de l'eau ne prend en compte les corrélations temporelles et spatiales des mesures. A priori, les modèles statistiques décrivant les phénomènes de pollution ne prennent pas en compte la structure arborescente et l'ensemble des mesures sont considérées comme les réalisations de variables aléatoires identiquement distribuées et indépendantes. C'est une hypothèse sous jacente lorsqu'on utilise par exemple une régression. Plusieurs auteurs se sont intéressés à ces problèmes et à l'utilisation de la géostatistique pour les résoudre. Dans un premier temps, certains auteurs (Villeneuve et al, 1979 ; Huang and Yang, 1998) préconisent l'utilisation du krigeage pour l'estimation des débits moyens sur une période supérieure à un mois dans l'objectif d'intégrer la variabilité spatiale dans les estimations et de proposer le meilleur estimateur linéaire au sens de la variance d'estimation minimale. L'estimation porte sur les débits spécifiques (débit divisé par la surface drainée) pour se ramener à une fonction aléatoire stationnaire. Huang et al. se distinguent par le choix d'un point représentatif du bassin pour définir les distances dans le calcul des variogrammes.

Le problème de ces démarches est que la structure d'arbre liée à la géométrie d'un réseau hydrographique, n'est pas prise en compte. Les distances considérées dans les calculs sont euclidiennes (distance à « vol d'oiseau ») et non les distances au fil de l'eau, et tous les couples formés par des points sont pris en compte qu'ils soient reliés au fil de l'eau ou non. Ceci implique par exemple une hypothèse de dépendance des concentrations sur des rivières parallèles. La distance euclidienne, comme montré, peut fortement différer de la distance curviligne car deux stations peuvent être très proches géographiquement mais sur deux rivières dont la confluence est éloignée. Considérer la distance euclidienne et non la distance curviligne ne semble pas aberrant, les concentrations étant par exemple fortement influencées par l'occupation du sol, deux stations proches peuvent être très corrélées. Cependant, si on cherche à mettre en évidence des corrélations spatiales liées au transport de polluant sur une rivière, il faut travailler avec la distance curviligne. Enfin des interrogations sur les discontinuités aux confluences subsistent.

E. Sauquet (Sauquet, 2000a ; Sauquet et al., 2000b et 2000c) présente une synthèse bibliographique sur l'étude des corrélations et de l'interpolation des débits le long d'un réseau hydrographique, renvoyant à divers articles de L. Gottschalk (Gottschalk, 1993a et 1993b ; Gottshalk and Krasovskaia, , 1994 et 1998 ; Gottschalk and Tveito, 1997) et Krasovskaia (Krasovskaia et al., 2003) dans lesquels sont développés des modèles pour calculer une fonction de covariance des débits le long d'une rivière basés sur une équation de conservation du volume d'eau généré à l'échelle du bassin.

Dans sa thèse, E. Sauquet présente les résultats d'estimation d'écoulements annuels par krigeage des lames d'eau sur l'ensemble du réseau formé par le bassin Rhône-Méditerranée-Corse en comparant la distance euclidienne à celle calculée au fil de l'eau (deux points dont la confluence est la mer sont aussi pris en compte avec une distance égale à la somme de la distance entre les deux exutoires et la distance de chaque point à l'exutoire). Mais une validation croisée montre que les deux méthodes sont peu performantes. Il privilégie alors une autre piste selon laquelle le débit interannuel observable en chaque point du linéaire est l'intégrale sur le bassin versant d'un champ ponctuel de génération d'écoulement spatialement homogène. Plus simplement, le débit en un point est égal à la somme des contributions à l'écoulement en rivière en amont de ce point, divisé par la surface drainée. Le bassin versant est découpé en plusieurs zones adjacentes qui forment une partition du bassin. La contribution de chacune de ces zones est estimée par krigeage 2D à l'aide du réseau de référence inclus dans le bassin versant pour lequel module et surface drainée sont connus.

10.1.2.2 Quelques modèles récents géostatistiques

Intégration le long des cours d'eau

(Bruno et al., 2000)

R. Bruno et al. proposent une application du krigeage à l'analyse des sédiments en rivière. L'objectif est d'estimer la contribution locale du sol sur la concentration de sédiments mesurée en rivière. Pour cela, les valeurs mesurées sont considérées comme les réalisations d'une fonction aléatoire $Z(u)$ non stationnaire, définie sur $\mathbb{R}$ et égale à la combinaison linéaire de:

- Une fonction aléatoire stationnaire Y définie sur $\mathbb{R}^2$, représentant la contribution locale du sol
- La contribution de la rivière elle-même, par transport des sédiments de l'amont à l'aval : $Z(u-du)$

Le modèle est développé en considérant une distance curviligne calculée depuis chaque source et se sommant aux confluences. Ainsi à chaque confluence, les deux contributions amont sont prises en compte. Une écriture de Y en fonction de Z se déduit alors aisément de la modélisation et l'étude de la corrélation spatiale de Y et un krigeage peuvent être effectués. Le modèle est appliqué au vanadium sur le bassin versant du Reno en Italie. L'amplitude de la variance de krigeage s'avère assez élevée. Ce modèle conduit aux quelques interrogations suivantes :

- Le choix de la distance. Sur une rivière, c'est la distance curviligne depuis la source, mais après confluence, les distances amont se somment. La distance est donc discontinue et ne respecte pas non plus l'inégalité triangulaire si on considère deux points à l'amont-aval l'un de l'autre et séparé par un point de confluence.
- Le modèle est écrit pour une confluence, mais que devient la généralisation à l'arbre entier ?

Moyennes mobiles sur le réseau

(Ver Hoef et al., 2006)

Ver Hoef et al montrent que les modèles usuels géostatistiques de covariance ne sont plus valables sur les graphes (voir aussi Krivoruchko and Mateu, 2003). Ils proposent alors la construction d'un modèle basé sur le procédé des moyennes mobiles qui répartit sur les arrêtes un noyau défini sur la demi-droite amont (voir aussi Collins et al., 2001 ; Higdon, 2002).

Sur chaque bief, les concentrations sont supposées stationnaires. La construction de la fonction aléatoire sur l'ensemble du réseau hydrographique se fait alors de la source vers l'exutoire : la concentration en un point dépend uniquement des concentrations à l'amont. A chaque confluence, la concentration immédiatement à l'aval est égale à la somme pondérée des concentrations amont. Les poids correspondent aux débits ou à une information équivalente telle que l'ordre du cours d'eau, et ont la somme de leurs carrés égale à un dans l'objectif de maintenir la stationnarité. Un modèle de covariance est alors déduit de cette construction sous l'hypothèse d'indépendance entre rivières à l'amont de leur confluence, et une application du krigeage est donnée pour des concentrations de métaux lourds. Les résultats obtenus améliorent ceux obtenus par un krigeage classique avec distance euclidienne. Plusieurs questions restent en suspend à l'issue de cet article : au vu des données, peut on vraiment considérer les concentrations stationnaires sur l'ensemble du graphe, voire sur un bief ? Peut-on considérer les concentrations indépendantes sur deux rivières parallèles ?

Un modèle mixte pour les distances

(Cressie et al. , 1997 et 2005)

Deux articles de Cressie et al. traitent de l'estimation des concentrations sur un réseau hydrographique. Le premier présente une modélisation spatio-temporelle de la concentration en nitrates le long d'une rivière drainant un bassin versant. La structure du réseau hydrographique n'intervient pas dans la modélisation effectuée en deux dimensions, le temps et une dimension d'espace, via l'abscisse curviligne. Une dérive multiple permet de prendre en compte la saisonnalité ainsi que de nombreuses variables décrivant le milieu.

Le deuxième, écrit en 2006, propose une autre façon de tenir compte des caractéristiques locales. Il reprend la démarche de modélisation par Ver Hoef, en utilisant les moyennes mobiles. L'idée supplémentaire dans cet article consiste à combiner au modèle de covariance développé pour la distance curviligne, un modèle basé sur la distance euclidienne pour tenir compte des corrélations dues à la proximité géographique et des corrélations liées au transport amont du polluant. L'application porte sur une variable transformée et normalisée construite à partir des concentrations en oxygène dissous. Mais l'ajustement du modèle de covariance par maximum de vraisemblance montre que seul le modèle développé pour la distance euclidienne doit être pris en compte.

Modèles statistiques pour les cours d'eau

(Monestiez et al., 1989 et 2005 ; Audergon et al., 1993 ; Bailly et al., 2006)

Les travaux de Monestiez et al. dans le domaine des fonctions aléatoires définies sur des graphes ont débuté par l'étude des estimateurs de la fonction de covariance et du variogramme sur un arbre fruitier. Les auteurs calculent selon quatre distances les variogrammes expérimentaux de divers paramètres tels que le poids et la teneur en sucre des fruits. Retenant comme distance le nombre d'embranchements entre deux points, une covariance exponentielle est ajustée et un krigeage effectué. Le choix de la distance est pointé comme une étape importante de la modélisation.

Les articles suivants traitent des fonctions aléatoires définies sur des supports arborescents : on ne s'intéresse plus uniquement à des paramètres définis sur les nœuds source ou exutoires d'un graphe, mais à l'estimation de paramètres le long d'un réseau hydrographique, donc aussi sur les arrêtes. Pour modéliser la largeur du fluvisol dans la partie aval du réseau hydrographique de l'Hérault ou les fossés de drainage du bassin de Roujan, les auteurs construisent une fonction aléatoire de l'exutoire vers les sources, en posant une hypothèse d'indépendance conditionnelle entre points situés sur des cours d'eau différents, connaissant l'ensemble des valeurs à l'aval de leur confluence. Par cours d'eau, tous

les modèles de covariance sont admissibles, entre points situés sur des cours d'eau différents, la covariance n'est pas stationnaire. Des simulations conditionnelles sont présentées.

10.2 Modèle général

La construction de modèles géostatistiques de concentrations ou de débits le long des cours d'eau a fait l'objet d'un article (de Fouquet and Bernard-Michel, 2006) dans les comptes rendus géosciences de l'académie des sciences, reporté annexe C. Le principe général du modèle est rappelé dans ce paragraphe.

La relation d'additivité des débits ou des flux aux confluences d'un réseau hydrographique conduit dans la majorité des cas à l'étude de fonctions aléatoires non stationnaires, ce qui signifie que la variabilité de la variable régionalisée n'est pas uniforme dans l'espace.

En général, le principe des méthodes géostatistiques comme le krigeage universel ou le krigeage intrinsèque d'ordre k, consiste à extraire des ces fonctions une composante stationnaire. Par exemple, on peut chercher la relation de la variable avec des facteurs influents, la prendre en compte à l'aide d'une régression et travailler sur les résidus de cette régression. Mais dans le cas d'un réseau hydrographique, les résidus d'une régression sont non stationnaires.

Nous proposons une construction par «filets d'eau », en combinant des variables définies sur les chemins reliant les sources à l'exutoire. La pondération est alors appliquée pour respecter l'additivité des débits ou des flux aux confluences et non pour maintenir la stationnarité. Ces deux choix sont d'ailleurs incompatibles.

Cas d'une rivière sans confluence

Supposons une rivière isolée sans aucune confluence entre sa source et son exutoire,. On peut alors définir sur cette rivière une fonction aléatoire de covariance stationnaire autorisée en dimension 1 ou intrinsèque.

Cas d'un réseau hydrographique

Pour un réseau hydrographique à *N* sources, considérons les *N* fonctions aléatoires définies sur les *N* chemins allant des sources vers un exutoire unique (voir FIG. 10-1, à gauche). Ces fonctions aléatoires notées $Y_J,\ 1 \leq J \leq N$ définies sur des « filets élémentaires » en dimension 1, sont supposées stationnaires ou intrinsèques. Elles sont assimilées à des rivières sans confluence. Le modèle consiste alors à écrire la variable étudiée (concentration, flux) comme une combinaison linéaire des variables définies sur les filets élémentaires, supposées indépendantes ou non.

A l'aval des confluences successives depuis les sources, les filets se cumulent pour former le débit total et la concentration dans la rivière se ramène à une combinaison des concentrations des différents filets, en proportion de leur débit relatif.

De manière générale, en un point x du réseau à *n* sources, la fonction aléatoire étudiée Z s'écrit :

$$Z(x) = \sum_{j \in \{1,..n\}} W_j(x) Y_j(x)$$

où $W_j(x)$ est le poids affecté en x au filet j. Il est nul si x n'appartient pas au filet j, et dans le cas contraire, il se calcule par conservation des débits ou flux de concentrations aux confluences. Le poids

affecté en un point x au filet j est alors égal au produit des poids des arrêtes qui composent le filet j et qui sont situées à l'amont de x (voir FIG. 10-1, à droite).

Des exemples de calculs de ces poids ont déjà été donnés dans le cas d'une unique confluence au paragraphe 9.1.3 :

- Pour les débits, les poids affectés aux deux biefs à l'amont d'une confluence sont égaux à 1.
- Pour les débits spécifiques, ils dépendent de la surface drainée et ont leur somme égale à 1.
- Pour des concentrations, sous hypothèse de « non dégradation » de la matière le long du cours d'eau, ils sont fonctions des débits et ont aussi leur somme égale à 1.

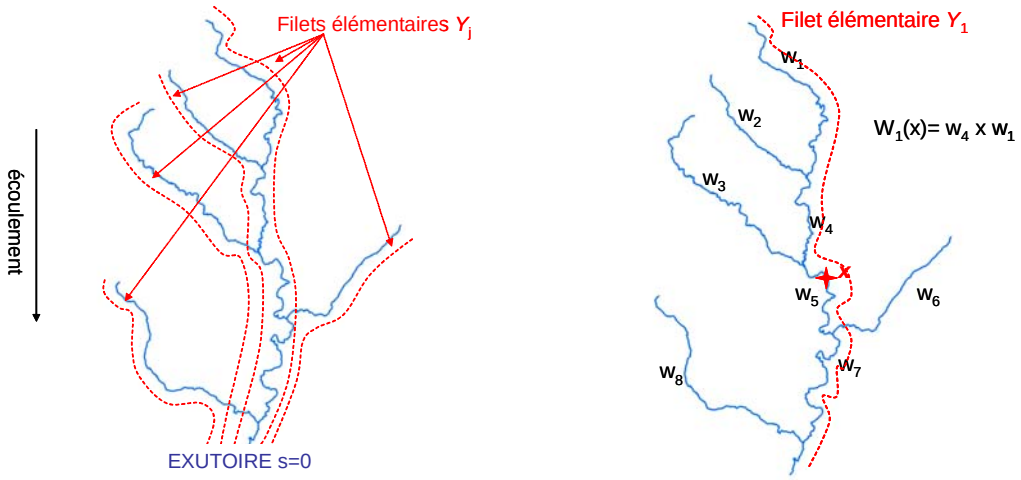


FIG. 10-1 : A gauche, filets élémentaires pour un réseau simplifié à 5 sources. A droite, calcul de la pondération en un point x pour ce même réseau.

Modèle de covariance

La décomposition de la fonction aléatoire en filets élémentaires permet alors le calcul d'un modèle valide de covariance non stationnaire entre deux points du graphe. Différents cas doivent être considérés : indépendance ou non des filets, stationnarité ou non des filets. Ces modèles sont décrits dans (de Fouquet and Bernard-Michel, 2006).

Dans le cas de filets aléatoires stationnaires indépendants de même moyenne, variance et covariance C, la covariance s'écrit :

$$C(x,y) = \begin{cases} 0 \text{ si } x \text{ et } y \text{ ne sont pas connectés} \\ C(0) \sum_{i\in\{1,..n\}} W_i(x)^2 \text{ si x=y} \\ C(|x-y|) \sum_{i\in F_c(x,y)} W_i(x) W_i(y) \text{ si } x \text{ et } y \text{ sont connectés} \end{cases}$$

avec $|x-y|$ la distance curviligne entre x et y et $F_c(x,y)$ les filets élémentaires communs à x et y

Inférence

L'application aux données réelles pose la question importante de l'inférence des modèles. Deux difficultés se présentent :

- Les hypothèses portent sur les filets aléatoires et non sur la fonction aléatoire associée à la variable étudiée. En pratique, aucune réalisation de ces filets n'est disponible excepté à proximité des sources avant confluence.
- Les données sont insuffisantes avec seulement une mesure par bief. Comment dans ces conditions vérifier la stationnarité par bief et l'indépendance avant confluences ?

Des outils permettant de vérifier la pertinence du modèle et de ses hypothèses doivent être développés.

10.3 Application aux débits spécifiques

Ce paragraphe présente une application des modèles aux débits spécifiques. Les débits n'on pas été retenus ici comme variable d'étude car ils présentent une non stationnarité par bief, ce qui est incompatible avec l'hypothèse de stationnarité des filets élémentaires qui les composent. Deux solutions sont envisageables pour se ramener à des filets potentiellement « stationnaires » : étudier les débits spécifiques ou les résidus d'une régression entre les débits et la surface drainée. La première solution a été retenue, l'approche par décomposition conduisant à un variogramme biaisé (Chilès and Delfiner, 1999).

Des outils de vérification des hypothèses sont proposés, ainsi qu'une méthode d'inférence de la covariance ou du variogramme des débits spécifiques. Les flux de nitrates divisés par la surface drainée que l'on appellera « flux spécifiques » sont présentés en parallèle.

10.3.1 Définitions

Nous introduisons ici des définitions pour l'étude des débits spécifiques, qui simplifient celles utilisées dans le modèle général (de Fouquet and Bernard-Michel, 2006). Soient :

- R_G le réseau hydrographique considéré. Ici il s'agit de la structure arborescente formée par les drains principaux du bassin de la Moselle définis par l'agence de l'eau Rhin Meuse.
- $I(x)$: Ensemble des filets d'eau passant par x
- $J(i,x)$: Ensemble des confluences rencontrées sur le filet *i* à l'amont de x inclus
- Y_i : Fonction aléatoire définie sur le filet d'eau *i,* issu de la source *i*. Ces fonctions représentent les débits spécifiques élémentaires sur les filets. On les considère indépendantes, stationnaires de même :
  - ✓ moyenne $E(Y_i(x)) = m \ \forall x \in R_G$
  - ✓ variance $Var(Y_i(x)) = \sigma^2 \ \forall x \in R_G$
  - ✓ covariance $C_1(h)$ autorisée en dimension 1.
- S_j^i : Surface drainée à l'amont de la confluence *j*, sur la branche contenant le filet *i.* Une même surface est donc définie par plusieurs noms, par exemple FIG. 10-3, S_2^1 et S_2^2 représentent la même surface drainée.
- S_j : Surface drainée immédiatement à l'aval de la confluence *j*.

- $S_j \setminus S_j^i$: Surface drainée immédiatement à l'amont de la confluence j sur la branche ne contenant pas le filet *i*
- $w_j^i = \frac{S_j^i}{S_j}$: Poids affecté au filet *i* à l'aval immédiat de la confluence *j*.
- $W(i,x)$: Poids affecté au filet *i* en *x*.

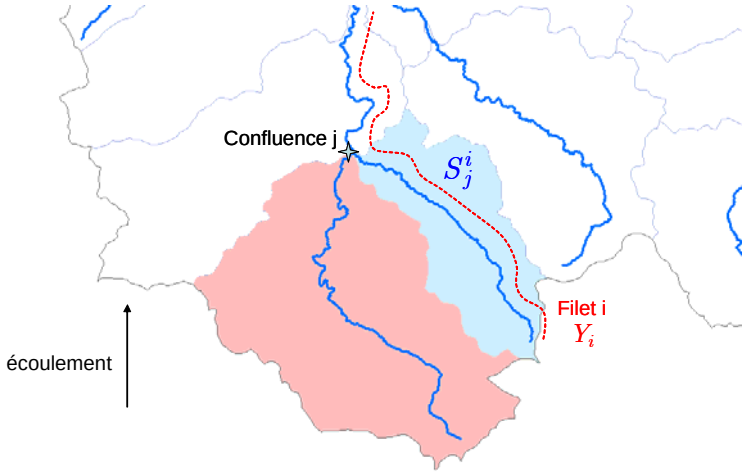


FIG. 10-2 : Notations : cas d'une unique confluence.

$T(x)$: Débit spécifique au point x appartenant au réseau hydrographique.

10.3.2 Modèle pour les débits spécifiques

L'additivité des débits aux confluences conduit à la pondération suivante :

$$T(x) = \sum_{i \in I(x)} W(i,x) Y_i(x) \text{ avec } W(i,x) = \begin{cases} \prod_{j \in J(i,x)} w_j^i \\ 1 \text{ si } J(i,x) = \{\varnothing\} \end{cases}$$

Dans l'exemple de la FIG. 10-3, on obtient:

$$T(x) = \frac{S_2^1 S_1^1}{S_2 S_1} Y_1(x) + \frac{S_2^2 S_1^2}{S_2 S_1} Y_2(x) + \frac{S_2^3}{S_2} Y_3(x)$$

En une confluence, la somme des poids vaut 1 ce qui conduit à la non stationnarité du débit spécifique observable sur l'ensemble du graphe.

Le modèle ainsi développé prend en compte l'ensemble des confluences du réseau hydrographique, c'est-à-dire que tous les cours d'eau référencés, quelle que soit leur taille ou importance, sont considérés. Ici, l'étude est restreinte à un arbre formé par les cours d'eau principaux définis par l'agence de l'eau Rhin Meuse. La question se pose d'ailleurs de savoir s'il est vraiment utile de considérer tous les cours d'eau : dans le cas d'un réseau routier, faut-il par exemple prendre en compte toutes les petites routes pour évaluer le flux de voitures sur une autoroute ? Quelle taille minimale doit être considérée ? Le chemin ? La route de campagne ? Appliquer le modèle sur un réseau simplifié

revient en fait à considérer la pondération constante par bief alors qu'en réalité ce bief est rejoint par des affluents. En revanche, le choix du réseau hydrographique doit être effectué avec prudence : les filets d'eau qui débutent aux sources doivent se comporter de manière similaire près des sources sinon l'hypothèse de fonctions aléatoires identiques pour les filets ne sera pas vérifiée.

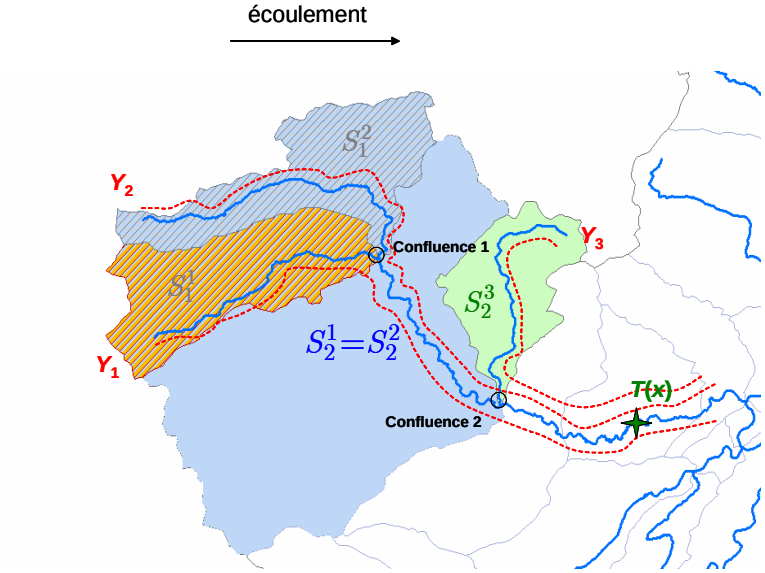


FIG. 10-3 : Notations et calcul du débit spécifique dans le cas de deux confluences.

Sous les hypothèses de mêmes fonctions aléatoires stationnaires indépendantes d'espérance *m* et de variance σ^2 pour chaque filet élémentaire, on déduit les propriétés suivantes pour les débits spécifiques:

Espérance :

$$E(T(x)) = m \quad (10\text{-}1)$$

Variance :

$$Var(T(x)) = \sigma^2 \times \sum_{i \in I(x)} W(i,x)^2 \quad (10\text{-}2)$$

Covariance :

$$C(x,y) = \begin{cases} 0 \text{ si x et y sont non connectés} \\ C_1(0) \sum_{i \in I(x)} W(i,x)^2 \text{ si } x = y \\ C_1(|x-y|) \sum_{i \in I(x) \cap I(y)} W(i,x) W(i,y) \text{ si } x \text{ et } y \text{ sont connectés au fil de l'eau} \end{cases} \quad (10\text{-}3)$$

Dans le cas où x et y sont sur un même bief sans confluence qui les sépare, la formule reste valable :

$$C(x,y) = C_1(|x-y|) \sum_{i \in I(x)} W(i,x)^2$$

Variogramme :

$$Var(T(y)-T(x)) = \begin{cases} 0 \text{ si } x = y \\ \sigma^2 \left(\sum_{i \in I(y)} W(i,y)^2 + \sum_{i \in I(x)} W(i,x)^2 \right) \text{ si } x \text{ et } y \text{ ne sont pas connectés} \\ 2\gamma_1(y-x) \left[K(x,y) \sum_{i \in I(x)} W(i,x)^2 \right] \\ \quad + \sigma^2 \left[\sum_{i \in I(y) \setminus I(x)} W(i,y)^2 + (K(x,y)-1)^2 \sum_{i \in I(x)} W(i,x)^2 \right] \\ \quad si\ x \text{ et } y \text{ sont connectés, } y \text{ à l'aval de } x \end{cases} \qquad (10\text{-}4)$$

avec $K(x,y) = \prod_{j \in J(i,y) \setminus J(i,x)} w_j^i = \frac{W(i,y)}{W(i,x)}$

10.3.3 Inférence du modèle

L'inférence du modèle repose sur plusieurs étapes :

- D'une part la vérification des hypothèses sur les données
- Ensuite, le développement d'outils pour l'inférence du variogramme des fonctions aléatoires définies sur les filets.
- Enfin, l'inférence elle-même et l'évaluation de la qualité du modèle selon différents critères tels que la validation croisée, l'amplitude de la variance d'estimation annoncée et la comparaison des estimations à celles fournies par des modèles existants, pour mesurer l'apport de la géostatistique.

Le modèle que nous avons retenu pour l'étude des débits spécifiques et des « flux spécifiques » est le plus simple, pour lequel les fonctions aléatoires sont supposées stationnaires et indépendantes.

Son inférence est principalement confrontée aux difficultés suivantes :

- Les fonctions aléatoires sont définies sur des filets aléatoires, aucune réalisation n'est donc disponible à l'exception des stations proches des sources
- Peu de mesures sont disponibles, avec en moyenne moins d'une mesure par bief et des stations espacées de plus de 30 kilomètres, ce qui rend la vérification des hypothèses et l'inférence difficile.
- Les débits spécifiques et flux de nitrates dépendent de facteurs qui n'ont pas été pris en compte dans le modèle comme la pluie, le pourcentage de forêts, d'agriculture... La stationnarité des fonctions aléatoires risque donc de ne pas être vérifiée.

Nous proposons cependant diverses méthodes pour vérifier les hypothèses et procéder à l'inférence du variogramme. Les résultats montrent qu'une estimation par krigeage n'est pas envisageable pour le moment et qu'il serait nécessaire pour cela non seulement de disposer de plus de mesures, en particulier près des sources, mais aussi de combiner le modèle à un modèle phénoménologique.

10.3.3.1 Vérification des hypothèses

Vérifier les hypothèses du modèle choisi est difficile en particulier parce que les données sont peu nombreuses. Par exemple, sur le bassin de la Moselle en 1996, seules 68 stations de mesure sont disponibles. Les stations le long de la Moselle sont en moyenne espacées de 13 kilomètres, soit 17 stations en service en 1997 pour la rivière la mieux informée. De ce fait, les hypothèses deviennent alors impossibles à vérifier, en particulier la stationnarité de la fonction aléatoire entre deux confluences. A cela s'ajoute une difficulté: les hypothèses ne portent pas sur la fonction aléatoire (le débit spécifique), mais sur ses composantes élémentaires. Or à l'exception des mesures effectuées à l'aval immédiat des sources avant toute confluence (ce que nous appellerons les biefs de rang 1), aucune réalisation de ces fonctions aléatoires n'est disponible. Sur chaque bief de rang 1, une seule mesure a en général été effectuée. Il est donc impossible de vérifier la stationnarité pour chacun des filets élémentaires et encore plus difficile de vérifier l'hypothèse de fonctions aléatoires identiques sur les différents filets. Pour vérifier malgré tout ces hypothèses pour les débits spécifiques, nous avons groupé toutes les mesures des biefs de rang 1 avec l'idée suivante que si tous les filets sont les réalisations d'une même fonction aléatoire, alors pour une distance à la source donnée, les mesures sur chaque bief devraient se comporter de manière similaire. A priori, on s'attend donc à voir les débits augmenter en fonction de la distance à la source, ce qui confirmerait l'impossibilité de choisir des filets élémentaires stationnaires pour les débits. En revanche, les débits spécifiques devraient être constants au fil de l'eau avec une dispersion elle aussi constante. La figure FIG. 10-4 confirme dans l'ensemble ces résultats, avec des débits croissants et une variance croissante. Les débits spécifiques apparaissent légèrement décroissants lorsqu'on moyenne les données par classe de 15 points, mais globalement ils sont relativement constants avec des variations d'amplitudes constantes.

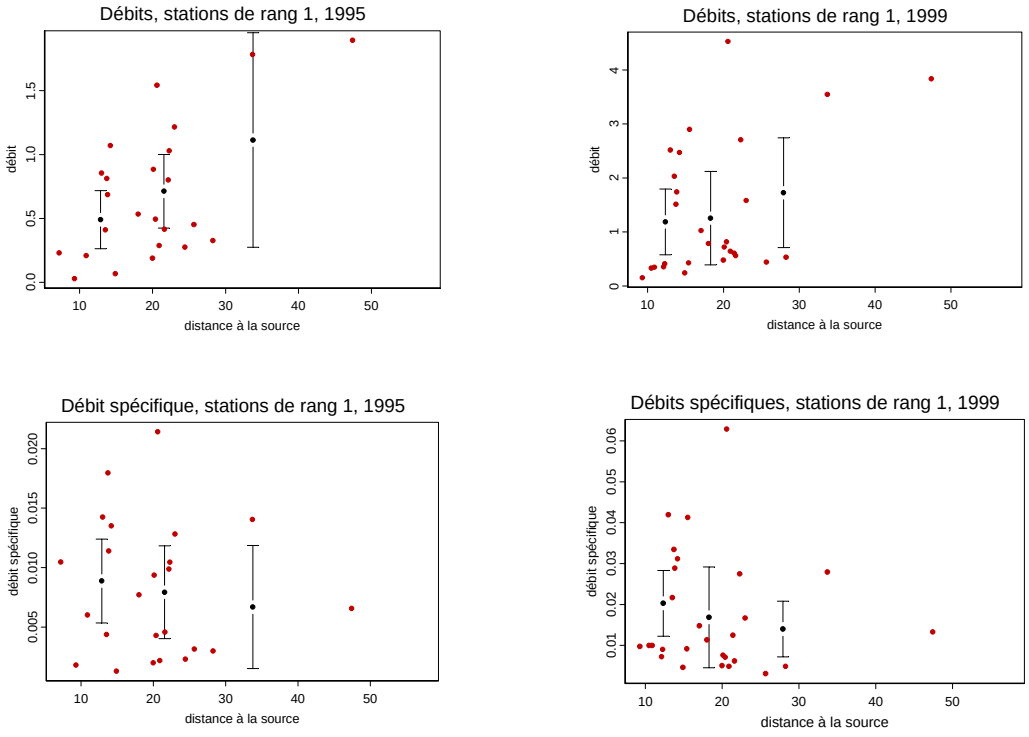


FIG. 10-4 : Débits (en haut) et débits spécifiques (en bas) pour les biefs de rang 1 en 1995 (à gauche) et 1999 (à droite). En noir, moyennes et intervalles de confiance à 95% par groupe de 15 points.

Sans avoir suffisamment de mesures à proximité des sources, il s'avère compliqué de vérifier les hypothèses du modèle choisi. En revanche, les hypothèses sur les filets impliquent certaines propriétés pour la fonction aléatoire définie sur le réseau entier, qui peuvent être testées.

Sous les hypothèses de fonctions aléatoires stationnaires indépendantes d'espérance *m* et de variance σ^2 identiques pour tous les filets élémentaires, on déduit les propriétés suivantes pour les débits spécifiques:

- L'espérance des débits spécifiques est constante en tout point du réseau hydrographique

$$E(T(x)) = m = cte \quad (10\text{-}5)$$

- La variance des débits spécifiques en un point x est fonction de la variance des filets élémentaires et d'un coefficient que nous notons $U(x) = \sqrt{\sum_{i \in I(x)} W(i,x)^2}$ qui dépend des poids affectés au point x. La variance de la fonction aléatoire $\frac{T(x)}{U(x)}$, appelée « Débit spécifique pondéré » est constante et égale à la variance des filets élémentaires.

$$Var\left(\frac{T(x)}{U(x)}\right) = \sigma^2 = cte \quad (10\text{-}6)$$

- Enfin, pour deux points x et y non connectés du réseau, la variance de la différence des débits spécifiques divisée par un coefficient noté $UNC(x,y) = \sqrt{\left(\sum_{i \in I(y)} W(i,y)^2 + \sum_{i \in I(x)} W(i,x)^2\right)}$ (dépendant des poids affectés à ces points) est elle aussi constante et égale à la variance des filets élémentaires. Cette relation montre que le variogramme expérimental usuel, que l'on appellera par la suite « variogramme d'accroissements pondérés», calculé sur des couples de points non connectés n'est pas constant, ce qui se vérifie par exemple figure FIG. 10-5.

$$Var\left(\frac{T(y) - T(x)}{UNC(x,y)}\right) = \sigma^2 = cte \text{ si } x \text{ et } y \text{ ne sont pas connectés} \quad (10\text{-}7)$$

Ces trois points sont vérifiés à l'aide des mesures disponibles pour les débits spécifiques. Cependant, en chaque point de mesure, on a accès à une unique réalisation pour chaque année. Il est donc difficile d'estimer l'espérance et la variance des débits spécifiques en ces points. Une solution pourrait consister à grouper plusieurs mesures proches d'une même rivière et d'en calculer la moyenne et la variance, ainsi on aurait une représentation sur le réseau entier de l'espérance et de la variance des débits spécifiques. Mais les stations sont éloignées et on a rarement plus d'une station par bief. Nous avons donc choisi de grouper les stations en fonction de leur surface drainée. Par exemple, pour vérifier l'hypothèse d'espérance constante des débits spécifiques sur le réseau, nous représentons les débits spécifiques en fonction de la surface drainée.

Une vérification des propriétés sur la Moselle est menée en parallèle pour étudier les propriétés au fil de l'eau.

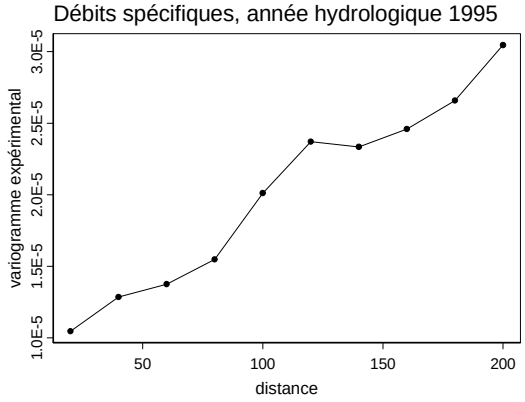


FIG. 10-5 : variogramme expérimental des débits spécifiques pour l'année hydrologique 1995 sur le bassin de la Moselle. Points non connectés.

Débits

Les hypothèses ont été vérifiées sur l'ensemble des années, mais seuls deux exemples représentatifs sont présentés ici.

Espérance des débits spécifiques :

Pour les années 1995 et 1999, les figures FIG. 10-6 (haut) et FIG. 10-8 (haut) montrent que le passage de l'étude des débits aux débits spécifiques réduit fortement la non stationnarité de l'espérance et de la variance des débits qui croissent en fonction de la surface drainée. L'hypothèse d'espérance constante pour les débits spécifiques semble a priori admissible si on regarde leur évolution moyenne en fonction de la surface drainée FIG. 10-6 et FIG. 10-8, en haut à droite.

Ces graphiques présentent les moyennes et intervalles de confiance à 95% des débits spécifiques pour des groupes de 15 points en fonction du logarithme de la surface drainée (pour plus de lisibilité).

En 1995, les points s'écartent au maximum de 30% de la moyenne des débits spécifiques calculée sur l'ensemble des points. En 1999, ils s'en écartent au maximum de 20%. En revanche, l'étude des débits spécifiques le long de la Moselle uniquement montre qu'en 1995 ils diminuent de l'amont vers l'aval (FIG. 10-7) ce qui est moins marqué en 1999 avec une pente diminuée de 11% (FIG. 10-9). Pour l'ensemble des années étudiées, de façon logique, l'hypothèse d'espérance constante pour les débits spécifiques est d'autant mieux vérifiée que la relation est linéaire entre les débits et la surface drainée. Dans le cas d'une relation non linéaire, une étude des résidus de la régression serait plus appropriée.

En résumé, les débits spécifiques sont nettement plus stationnaires que les débits. Pour certaines années pour lesquelles la relation entre débits et surface drainée est linéaire, l'hypothèse d'une espérance des débits spécifiques constante est acceptable.

Variance des débits spécifiques pondérés

- La variance des « débits spécifiques pondérés » (voir équation

$Var\left(\frac{T(x)}{U(x)}\right) = \sigma^2 = cte$ (10-6)) est présentée FIG. 10-6 et FIG. 10-8 en bas à gauche pour des groupes de 15 stations dont les surfaces drainées sont analogues. Les résultats ne sont pas très satisfaisants car les points s'éloignent jusqu'à 200% de la valeur moyenne des débits spécifiques pondérés en 1995 et 150% en 1999. L'étude de la variance le long de la Moselle (FIG. 10-7 et FIG. 10-9) montre une variance légèrement croissante en 1999 et plutôt constante en 1995. Ces deux graphiques présentent les débits spécifiques pondérés en fonction de la surface drainée plutôt que leur variance, le nombre de stations sur la Moselle étant trop faible (17 stations).

Variogramme pondéré des débits spécifiques : stations non connectées

Le « variogramme pondéré » des débits spécifiques pour les stations non connectées du réseau correspond à la variance d'un accroissement pondéré en fonction des emplacements des stations sur le réseau. Nous l'avons approché par la fonction *F* suivante dépendant de la distance entre les points :

$$F(h) = \frac{1}{|N(h)|}\sum_{N(h)} \frac{\left[Z(x_\alpha) - Z(x_\beta)\right]^2}{UNC(x,y)^2} \text{ avec } N(h) = \left\{(\alpha,\beta),\ x_\alpha \text{ et } x_\beta \text{ non connectés}, x_\alpha - x_\beta = h\right\}$$

Les résultats (FIG. 10-6 et FIG. 10-8 en bas à droite) montrent que pour l'année 1995, l'hypothèse d'une variance constante n'est pas vérifiée, la variance ayant plutôt tendance à augmenter avec la distance. A l'inverse, la variance semble constante pour l'année 1999.

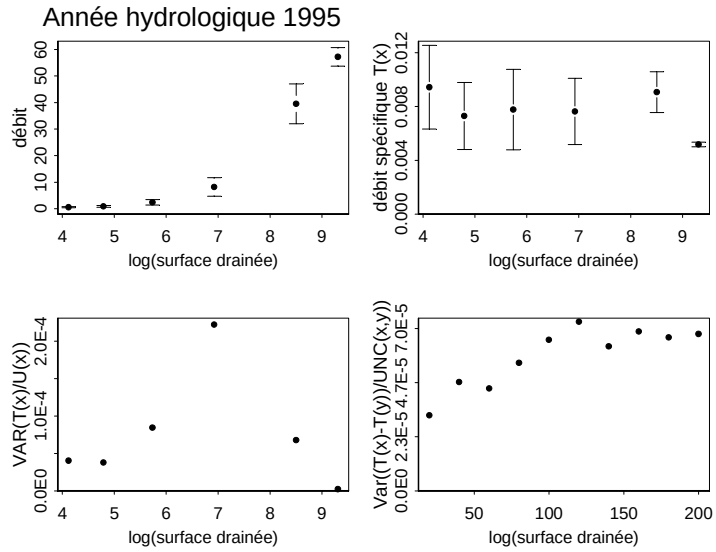


FIG. 10-6 : Vérification des hypothèses du modèle sur le bassin de la Moselle pour l'année hydrologique 1995.
En abscisse : surfaces drainées en échelle logarithmique
En haut : Moyennes et intervalles de confiance à 95% par groupe de 15 points. A gauche : débits. A droite : débits spécifiques.
En bas : A gauche : variance pondérée des débits spécifiques estimée par groupe de 15 points. A droite : variogramme expérimental pondéré des débits spécifiques.

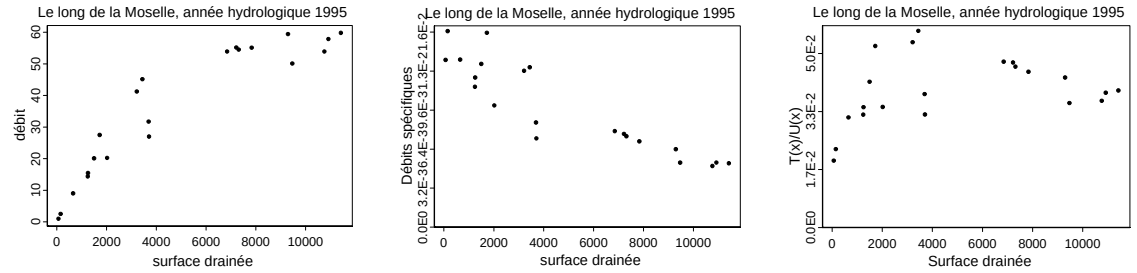


FIG. 10-7: De gauche à droite, débits, débits spécifiques, débits spécifiques pondérés en fonction de la surface drainée. Le long de la Moselle, année hydrologique 1995.

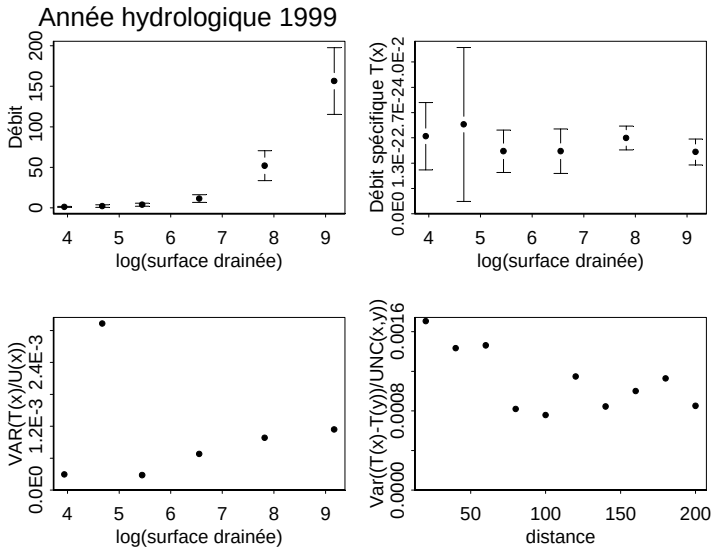


FIG. 10-8 : Vérification des hypothèses du modèle sur le bassin de la Moselle pour l'année hydrologique 1999.

En abscisse : surfaces drainées en échelle logarithmique

En haut : Moyennes et intervalles de confiance à 95% par groupe de 15 points. A gauche : débits. A droite : débits spécifiques.

En bas : A gauche : variance pondérée des débits spécifiques estimée par groupe de 15 points. A droite : variogramme expérimental pondéré des débits spécifiques.

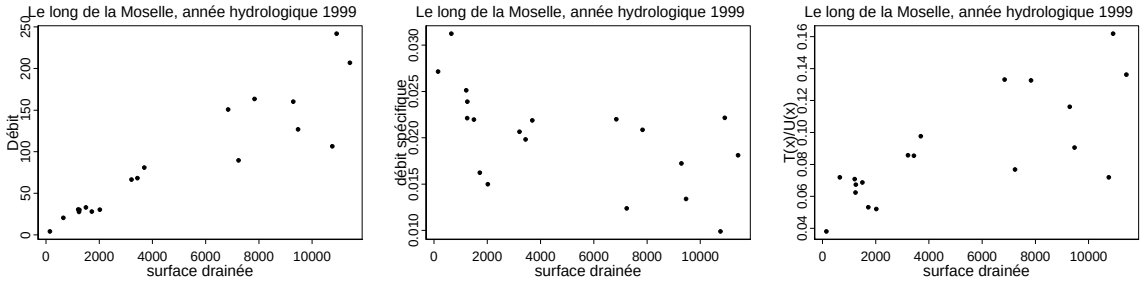


FIG. 10-9: De gauche à droite, débits, débits spécifiques, débits spécifiques pondérés en fonction de la surface drainée. Le long de la Moselle, année hydrologique 1999.

Flux de nitrates « spécifiques »

De façon analogue, le modèle peut être étendu aux flux de nitrate en divisant les flux de nitrates par les surfaces drainées de manière à se ramener à des filets élémentaires stationnaires. C'est ce que nous appellerons les flux de nitrate « spécifiques ». Etudier les flux de nitrate spécifiques revient quasiment à étudier les concentrations elles-mêmes, les débits et les surfaces drainées étant fortement liées.

Les hypothèses semblent nettement plus réalistes pour les flux spécifiques qui présentent une stationnarité plus marquée (FIG. 10-10 et FIG. 10-11) que les débits spécifiques. En particulier, le variogramme des accroissements pondérés pour les stations non connectés est relativement constant.

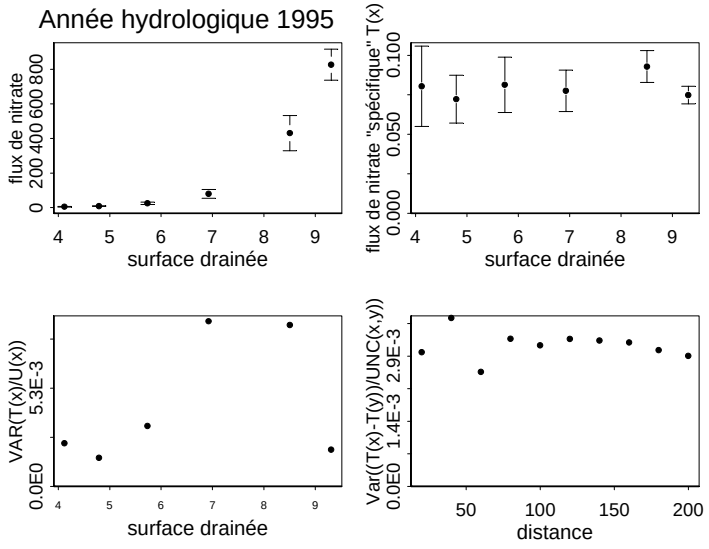


FIG. 10-10 : « Flux de nitrate spécifiques » : Vérification des hypothèses du modèle sur le bassin de la Moselle pour l'année hydrologique 1999.
En haut : Moyennes et intervalles de confiance à 95% par groupe de 15 points. A gauche : flux de nitrate en fonction du log de la surface drainée. A droite : « flux de nitrate spécifiques » en fonction de la surface drainée.
En bas : A gauche : variance pondérée des « flux de nitrate spécifique »s estimée par groupe de 15 points. A droite : variogramme expérimental pondéré des « flux de spécifiques » en fonction de la distance.

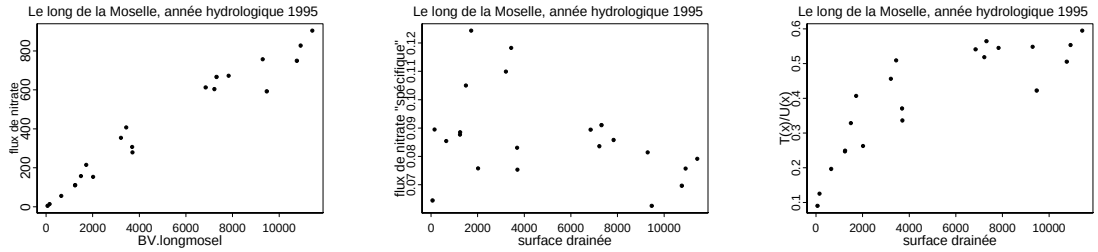


FIG. 10-11: Le long de la Moselle, année hydrologique 1995. De gauche à droite, flux de nitrate, « flux de nitrate spécifiques », « flux de nitrate spécifiques » pondérés en fonction de la surface drainée.

10.3.3.2 Inférence du modèle

Aux paragraphes précédents, nous avons montré comment construire, pour les débits, les débits spécifiques, les flux ou les concentrations de polluant, un modèle de fonctions aléatoires valide sur un réseau hydrographique. La question se pose alors d'inférer ces modèles, c'est-à-dire de confronter ces modèles théoriques aux données afin d'en déterminer les paramètres. Habituellement en géostatistique, on ajuste un variogramme calculé expérimentalement par une combinaison de modèles afin de dégager « visuellement » le modèle le plus approprié et d'en estimer les paramètres par diverses techniques statistiques, telles qu'une régression non linéaire. Cette approche est par exemple celle de (Sauquet, 2000a) qui étudie le variogramme expérimental des débits en considérant la distance au fil de l'eau et l'ajuste par un modèle sphérique. Comme il a été vu précédemment, ce modèle conduit parfois à des matrices de covariance non définies positives dans les équations de krigeage (Sauquet, 2000a).

Pour des variables définies le long des rivières, cette étape de visualisation est rarement possible car comme il a été vu précédemment, le cumul des fonctions aléatoires à chaque confluence conduit à l'étude de fonctions fortement non stationnaires sur l'ensemble du réseau hydrographique. Le variogramme ou la fonction de covariance entre deux points ne dépend plus de la distance qui les sépare, mais de la position des points eux-mêmes. Ces modèles sont donc difficilement « observables » puisqu'ils ne s'expriment plus en fonction de la distance.

Dans ce cas, une régression ou une estimation par maximum de vraisemblance sont utilisées pour caler les paramètres du modèle. Pour un modèle de covariance *C* s'exprimant comme le produit d'un modèle de covariance usuel et d'une fonction *K* connue, $C(x,y) = C_1(|x-y|) * K(x,y)$ (modèles (Ver Hoef et al., 2006 ; Cressie, 2005 ; de Fouquet and Bernard-Michel, 2006), on teste plusieurs modèles de covariance C_1, et le choix définitif est souvent déterminé par validation croisée. Le principe général est d'estimer tour à tour chaque observation à partir des autres données. On peut alors calculer l'erreur d'estimation en chaque point de donnée et la comparer à l'écart type d'estimation donné par le krigeage.

L'inconvénient de cette démarche est l'utilisation de modèles de type « boites noires » pour lesquelles on ne visualise absolument pas le phénomène étudié. De plus, le choix d'un modèle par rapport à un autre est déduit de la qualité des estimations, mais finalement peut-on vraiment se fier à un tel critère quand on dispose de seulement 23 points de mesures (Ver Hoef et al., 2005) ou comme ici 70 sur le bassin de la Moselle. En restreignant l'arbre étudié aux seuls drains principaux, moins d'une mesure par bief est disponible et les stations sont espacées de plus de 30 kilomètres. Comment dans ces conditions obtenir des estimations précises et inférer le modèle ? Cela nous semble irréaliste, et c'est pourquoi nous présentons ici, uniquement la méthodologie pour inférer le modèle.

Nous proposons l'inférence du variogramme donné par l'équation (8.4) de deux manières :

- Soit on cherche à inférer le variogramme de la fonction aléatoire définie sur le réseau hydrographique entier en utilisant les techniques dites « boites noires ». Les paramètres du variogramme γ_1 des filets élémentaires et leur variance sont estimés par maximum de vraisemblance. Le modèle final pour γ_1 est retenu par validation croisée.
- Soit on cherche à inférer le variogramme des filets γ_1, qui est stationnaire et se déduit de l'équation (8.4) par la formule suivante :

$$\gamma_1(|y-x|) = \frac{Var(T(y)-T(x)) - \sigma^2\left[\sum_{i\in I(y)\setminus I(x)} W(i,y)^2 + (K(x,y)-1)^2 \sum_{i\in I(x)} W(i,x)^2\right]}{2K(x,y)\sum_{i\in I(x)} W(i,x)^2} \quad (10\text{-}8)$$

Pour cela, deux cas sont à envisager.

A l'aide des biefs de rang 1, si l'on a suffisamment de données, on peut inférer le variogramme sur une distance limitée. C'est impossible en pratique à réaliser puisque seule une mesure par bief est disponible. En revanche, on peut approcher expérimentalement γ_1 par l'expression (10-8) en approchant la variance des écarts des débits spécifiques $Var(T(y)-T(x))$ par la moyenne des écarts quadratiques $(T(y)-T(x))^2$. Nous sommes conscients que cette manière de procéder entraîne des biais dans l'estimation du variogramme et des moyennes annuelles elles mêmes, mais l'idée est dans un premier temps de voir si une structure se dégage du variogramme observé. Les résultats sur le bassin de la Moselle en 1995 et 1999 sont en fait peu convaincants et le variogramme γ_1 calculé expérimentalement est négatif. Le calcul (10-8) repose sur le choix de la variance des filets élémentaires σ^2, relativement mal estimée par le faible nombre de mesures. Nous avons estimé cette variance à l'aide des mesures sur les biefs de rang 1. En 1995, la variance est estimée par 3.11 10^{-5}, en 1999 par 2.11 10^{-4}. Les variogrammes expérimentaux obtenus pour ces valeurs de variance montrent une structure croissante en fonction de la distance mais ont la majorité des valeurs négatives (voir FIG. 10-12 en haut). Si on augmente la valeur de la variance, alors le variogramme présente des valeurs de plus en plus fortement négatives. Si au contraire on diminue la variance des filets élémentaires, alors les variogrammes prennent des valeurs positives (voir FIG. 10-12 en bas), mais ne présentent pas de structure modélisables. Une étude de sensibilité a été menée sur le choix de la variance σ^2, en particulier nous avons cherché la valeur de σ^2 pour laquelle, les valeurs du variogramme expérimental deviennent toutes positives. Nous avons trouvé σ^2 égal à 2.2 10^{-6}. Le variogramme associé présente alors une structure sphérique (FIG. 10-13) de portée 100 et de palier 0.002. Ces résultats ne permettent pas de déduire un modèle pour le variogramme des filets élémentaires. Les causes possibles d'erreurs peuvent s'expliquer par le manque de données, des corrélations entre filets élémentaires engendrées par la dépendance des mesures au milieu qui n'ont pas été prises en compte.

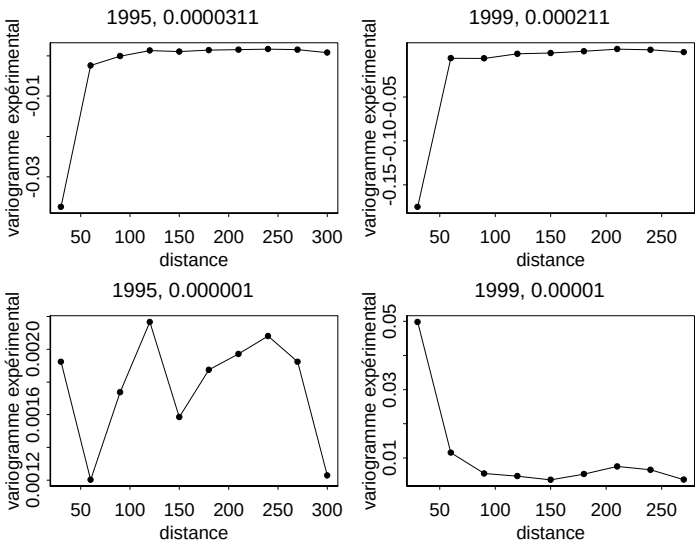


FIG. 10-12 : Variogrammes expérimentaux des filets élémentaires en 1995 et 1999 sur le bassin de la Moselle pour différentes variances σ^2 signalées à droite de l'année dans le titre du graphique

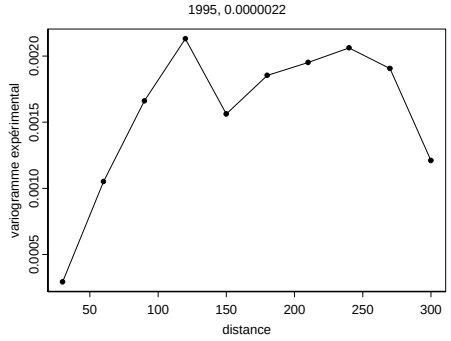


FIG. 10-13 : Variogramme des filets élémentaires en 1995 pour $\sigma^2 = 0.0000022$

Pour inférer les modèles géostatistiques le long des cours d'eau il semble impératif de disposer de plus de mesures spatialement. En particulier, un arbre composé de quelques sources et de quelques confluences, avec de nombreuses stations par bief permettrait d'évaluer la pertinence et l'apport de la géostatistique dans l'estimation des débits, concentrations et flux annuels. Dans un second temps, les concentrations et les flux le long des cours d'eau dépendent d'un nombre considérable de facteurs qu'il semble inévitable de prendre en compte dans la modélisation puisqu'ils expliquent en majeure partie les valeurs observées. Quelques relations ont été mises en évidence dans cette thèse : débits / surface drainée, nitrates / forêts-agriculture, mais il est indispensable de faire un lien avec des modèles phénoménologiques tels que PEGASE dans lequel de nombreux facteurs d'influence sont pris en compte.

On pourrait par exemple introduire une variabilité spatiale par simulation géostatistique sur les facteurs d'entrée de PEGASE et en déduire la variabilité des débits, concentrations et flux en sortie.

On pourrait aussi travailler en krigeage avec dérive externe : si le modèle phénoménologique est « bien fait », il explique une bonne partie de la variabilité spatiale, les données simulées par le modèle

peuvent alors être entrées comme dérive du modèle. Cette démarche permettrait de recaler le modèle aux mesures. Quelques problèmes de supports doivent cependant être discutés, et un plus grand nombre de stations doit être envisagé.
Enfin, et c'est probablement le plus difficile de manière pratique, il faudrait récupérer les équations du modèle PEGASE afin de répertorier les facteurs d'influence dans l'objectif de les prendre en compte par des modèles géostatistiques multivariables. Il serait alors intéressant de comparer les estimations et leurs incertitudes à celles simulées par PEGASE.

Conclusion

Conclusions et recommandations

Etude des indicateurs par station

Le calcul des indicateurs de qualité évalués ici pour les nutriments, est amélioré par le krigeage qui permet de tenir compte de la corrélation temporelle. Ainsi, les biais résultant des échantillonnages préférentiels sont corrigés pour la moyenne annuelle et réduits pour les quantiles et l'incertitude prévue par le modèle théorique est plus réaliste. Le calcul pratique de la précision pour le quantile 90, qui peut être approché par simulation, reste cependant une question ouverte.

La comparaison aux méthodes actuelles pour les 269 stations RNB du 8ème programme d'intervention en Loire Bretagne, montre que les différences sont parfois très importantes. Il y aurait donc un réel intérêt à mettre en œuvre ces indicateurs géostatistiques, facilement automatisables.

Bien évidement, les estimations sont à interpréter avec précaution. Les propriétés statistiques des estimateurs telles que le biais et la variance d'estimation ne doivent pas être confondues avec les résumés numériques obtenus pour un unique échantillon. C'est seulement en moyenne qu'on déduit qu'une méthode est meilleure qu'une autre.

En pratique, les incertitudes sur les indicateurs restent importantes, les mesures par année et par station étant peu nombreuses. Ainsi, avec 12 mesures par an en 1985 à Orléans, l'écart type d'estimation pour les nitrites atteint 11% pour la moyenne et 17% pour le quantile 90. Il serait donc utile de tenir compte de cette incertitude avant de procéder au codage actuel par classe de couleur. Nous conseillons de plus d'augmenter la fréquence des mesures et de retenir un échantillonnage régulier.

Pour la suite, pour améliorer les estimations, on pourrait tenir compte des analogies de milieu, indiquées par exemple par l'occupation du sol de façon à regrouper des stations dont le comportement est jugé similaire pour une substance donnée. Un cokrigeage entre stations d'un même groupe permettrait d'améliorer les estimations, en particulier sur les stations les moins bien informées.

Les méthodes ici proposées pour les nutriments peuvent bien évidement être étendues à d'autres paramètres et aux quantiles de tout ordre.

Dans le cas des flux de polluants, des techniques géostatistiques multivariables peuvent être mises en œuvre, les débits étant mesurés « en continu » aux stations hydrométriques.

La réflexion sur le calcul et l'utilisation des indicateurs mérite d'être poursuivie, ces indicateurs étant à la base des interprétations sur l'évolution des pollutions.

Etude des corrélations spatiales

L'étude des corrélations spatiales le long de supports arborescents est un sujet nouveau en géostatistique qui nécessite un retour sur les fondements de la géostatistique.

Un modèle général de fonctions aléatoires sur un réseau hydrographique a été développé, fournissant des modèles valides de covariance.

L'inférence de ce modèle s'est heurtée au nombre trop réduit de stations entre confluences. Ce modèle constitue donc une première approche vers des modélisations plus complexes qui devraient être l'objet de futurs travaux. En particulier, il serait nécessaire de tenir compte des modèles phénoménologiques tels que PEGASE.

Ces modèles pouvant expliquer une bonne partie de la variabilité spatiale, on pourrait travailler sur les écarts entre les mesures et ces modèles, ce qui permettrait de recaler le modèle aux mesures expérimentales.

Par ailleurs, la variabilité le long du réseau hydrographique pourrait être étudiée en introduisant une variabilité spatiale dans les principaux facteurs en entrée des modèles phénoménologiques.

En effet, la démarche la plus intéressante serait d'associer les deux modélisations, phénoménologique et géostatistique afin de mesurer l'apport de la géostatistique.
Car finalement, la question à laquelle on souhaite répondre est la suivante : quel est l'apport de la géostatistique dans la modélisation des concentrations le long des cours d'eau ? Dans le cas de l'étude des indicateurs par station, il a été vu qu'elle améliore les estimations mais surtout qu'elle fournit une précision plus réaliste que celle annoncée par statistique classique. Les conclusions pour la modélisation spatiale d'une variable le long d'un réseau hydrographique devraient être identiques, mais de nombreuses étapes sont encore nécessaires avant d'arriver à cette conclusion. Tout d'abord, il faut construire un modèle géostatistique valide respectant les propriétés physiques des variables étudiées et prenant en compte les facteurs d'influence. Mais de plus, les hypothèses d'un tel modèle doivent être vérifiées si l'on veut obtenir des estimations et des incertitudes fiables.

Beaucoup d'études restent donc envisageables pour mettre en évidence les corrélations spatiales sur un support arborescent, et les prendre en compte dans l'estimation des concentrations.

Bibliographie

Bibliographie

[1] Agence de l'eau Loire Bretagne (2001). EUROWATERNET. Construction d'un réseau représentatif de la qualité des cours d'eau. Rapport final, IFEN. BETURE CEREC.

[2] Agence de l'eau Loire Bretagne (2002). Système d'évaluation de la qualité de l'eau des cours d'eau. Rapport de présentation SEQ-Eau. Version 2.

[3] Agences de l'eau et du Ministère de l'Environnement (1994). Les traitements statistiques et graphiques utilisés par les agences de l'eau dans le cadre des données techniques physico-chimique, Etude Inter Agences.

[4] Audergon, J.-M., P. Monestiez, and R. Habib (1993). Spatial dependences and sampling in a fruit tree: A new concept for spatial prediction in fruit studies. Journal of Horticultural Science 68(1): 99-112.

[5] Bailly, J.-S., P. Monestiez, and P. Lagacherie (2006). Modelling spatial processes on directed trees with geostatistics: Application to artificial network properties on rural catchment areas. Mathematical Geology (in press).

[6] Becker, R. A., J. M. Chambers, and A.R. Wilks (1988). The new S Language, a programming environment for data analysis and graphics. Wadsworth and brooks/Cole, Pacific Grove.

[7] Bernard-Michel, C. and C. de Fouquet (2002). Remarques sur les calculs statistiques pour l'évaluation de la qualité de l'eau, Centre de Géostatistique. Ecole des Mines de Paris.

[8] Bernard-Michel, C. and C. de Fouquet (2003). Calculs statistiques et géostatistiques pour l'évaluation de la qualité de l'eau, Centre de Géostatistique, Ecole des Mines de Paris.

[9] Bernard-Michel, C. and C. de Fouquet (2004a). Calculs statistiques et géostatistiques pour l'évaluation de la qualité de l'eau: estimation du quantile 90, Centre de Géostatistique. Ecole des Mines de Paris.

[10] Bernard-Michel C. and C. de Fouquet (2004b). Estimating indicators of river quality by geostatistics. Geostatistics for Environmental Applications. (1): 443-454. P. Renard, H. Demougeot-Renard, R. Froidevaux (eds).

[11] Bernard-Michel C. and C. de Fouquet (2004c). Geostatistical indicators of waterway quality for nutrients. Geostatistics Banff 2004. (14): 907-912. Leuangthong O. and C.V. Deutsh (eds). Springer.

[12] Bernard-Michel, C. and C. de Fouquet (2005a). Calculs géostatistiques d'indicateurs des concentrations dans les cours d'eau. Centre de Géostatistique. Ecole des Mines de Paris.

[13] Bernard-Michel C. and C. de Fouquet (2005b). Geostatistical Indicators of Nutrients Concentrations in Streams. Proceedings of IAMG'05: GIS and Spatial Analysis, (2): 716-721. Quiming Cheng, Graeme Bonham Carter (eds).

[14] Borovkov A.A.(1987). Statistique mathématique, Mir Moscou.

[15] Brignon, J.-M. (2004). Les modèles pressions/impacts pour la directive-cadre Eau: bilan des outils actuellement utilisés et des besoins futurs. Rapport., Ineris/MECO/DRC.

[16] Bruno, R., V. Palumbo, and S. Bonduà (2000). Identification of regional variability component by geostatistical analysis of stream sediments. geoENV III- Geostatistics for Environmental Applications, Avignon, Kluwer Academic Publishers, 113-123.

[17] Chilès, J.-P. and P. Delfiner (1999). Geostatistics: modeling spatial uncertainty. New-York, Wiley.

[18] Collins, M. and N. Duffy (2001). Convolution Kernels for natural language. Neurol Information Proceeding Systems (NIPS).

[19] Cressie, N. and J. J. Masure (1997). Spatio-temporal Statistical Modeling of Livestock Waste in Streams. Journal of Agricultural, Biological, and Environmental Statistics 2(1): 24-47.

[20] Cressie, N., J. Frey, B. Harch, and M. Smith (2005). Spatial prediction on a river network. Department of statistics Preprint No. 747, The Ohio State University (to be published in JABES).

[21] De Fouquet, C. and N. Bez (2001). Construction d'un réseau représentatif de qualité des cours d'eau. Phase II. Rapport final, Centre de Géostatistique. Ecole des Mines de Paris.

[22] De Fouquet, C. and C. Bernard-Michel (2006). Modèles géostatistiques de concentrations ou de débits le long des cours d'eau. Comptes rendus Géoscience, 338(5) :307-318.

[23] Dent, L. C. and N. B. Grimm (1998). Spatial heterogeneity of stream water nutrient concentrations over successional time. Ecology 80(7): 2283-2298.

[24] Deschamps, P. (2004). Cours d'économétrie, Université de Fribourg.

[25] Emery, X. and M. Arnaud (2000) Estimation et interpolation spatiale. Méthodes déterministes et méthodes géostatistiques. Hermes science publications.

[26] Everbecq, E., J. F. Deliege, T. Bourouag, J-P. Dzisiak, and J. Smitz (2001). Consolidation et intégration de PEGASE au système de l'Agence. Rapport final, Centre d'étude et de modélisation de l'environnement de l'université de Liège (CEME) pour l'Agence de l'eau Rhin-Meuse.

[27] Florea Draghicescu, D. (2002). Non parametric quantile estimation for dependant data. Thèse n° 2592. Ecole polytechnique fédérale de Lausanne.

[28] Francois, D. (1997). Cartographie des écoulements caractéristiques. Etude méthodologique. Application aux bassins de la Meuse et de la Moselle françaises. 1ère phase: Etude méthodologique. A. d. l. e. R. Meuse.

[29] Ganio, L.M, C.E. Torgersen, and R. Gresswell. (2005). A geostatistical approach for describing spatial pattern in stream networks. Front Ecol Environ; 3(3):138-144.

[30] Gardner B., P.J. Sullivan, and A. J. Lembo (2003). Predicting stream temperatures : geostatistical model comparison using alternative distance metrics. Can J Fish Aquat Sci halieut. 60(3): 344-51.

[31] Gaudoin, O. (2001a). Méthodes statistiques pour l'ingénieur, notes de cours, ENSIMAG.

[32] Gaudoin, O. (2001b). Statistique non paramétrique, notes de cours, ENSIMAG.

[33] Gottschalk, L. (1993a). Correlation and covariance of runoff. Stochastic Hydrology and hydraulics 7: 85-101.

[34] Gottschalk, L. (1993b). Interpolation of runoff applying objective methods. Sochastic Hydrology and hydraulics 7: 269-281.

[35] Gottschalk, L. and I. Krasovskaia (1994). Interpolation of runoff to a regular grid net: theoritical aspects. FRIEND: flow regimes from international experimental and network data, Braunschweig, IAHS.

[36] Gottschalk, L. and I. Krasovskaia (1998). Grid estimation of ruoff data W. report of the WCP-Water project B3- Development of grid-related estimates of hydrological variables.

[37] Gottschalk, L. and O. E. Tveito (1997). Mapping of runnoff applying objective methods and GIS. Norsk Geogr. Tidskr. 51: 3-14.

[38] Higdon, D. M. (2002). "Space and space-time modeling using process convolution." Quantitative methods for Current Environmental Issues: 37-56.

[39] Huang, W.-C. and F.-T. Yang (1998). Streamflow estimation using kriging. Water resources research 34(No. 6): 1599-1608.

[40] Jack Chen, J. E. (2002). Two-phase quantile estimation. Winter Simulation Conference.

[41] Kishi, R. T., S. Fuchs, and H.H. Hahn (1998). Integrated use of spatial data and learning algorithms to detect water quality trends. ISPRS Commission IV Symposium on GIS, Stuttgart.

[42] Kondor, R. and J.-P. Vert (2004). Diffusion Kernels. Kernel methods in computational biology: eds. B. Scholkopf, K. Tsuda and J.-P. Vert. The MIT Press. 171-192.

[43] Kondor, R. I. and J. Lafferty (2002). Diffusion Kernels on graphs and Other Discrete Input Spaces. ICML 2002.

[44] Krasovskaia, I., L. Gottschalk, E. Leblois and E. Sauquet (2003). "Dynamics of river flow regimes viewed through attractors." Nordic Hydrolohy 34(5): 461-476.

[45] Krivoruchko, K. and J. Mateu (2003). Spatial statistics through non-Euclidean distances. The ISI International Conference on Environmental Statistics ans health.

[46] Lajaunie, C. (1993). L'estimation géostatistique minière. Cours C-152., Centre de Géostatistique. Ecole des Mines de Paris.

[47] Lantuejoul, C. (2001). Geostatistical simulations. Models and algorithms. Springer.

[48] Lee, H. K. H., C. H. Holloman, C.A. Calder and D.M. Higdon (2002). Flexible Gaussian Processes via convolution, Duke University.

[49] Leonard, J. (1999). Contribution au réseau "Eurowaternet" de l'agence Européenne de l'Environnement. Qualité des cours d'eau en France. Etude prototype nationale. Phase 1. Version 1. IFEN. Office international de l'eau.

[50] Littlejohn, C., S. Nixon, G. Cassazza, C. Fabiani, G. Premazzi, P. Heinonen, A. Ferguson and P. Pollard (2002). Guidance on monitoring for the water framework directive. Final draft. Working group 2.7., Agence de l'eau Loire Bretagne.

[51] Marcotte, D. and M. Powojowski (2004). Covariance models with spectral additive components. Geostatistics Banff 2004. (14): 115:124. Leuangthong O. and C.V. Deutsh (eds). Springer.

[52] Marechal, G. (1978). Analyse numérique des anamorphoses gaussiennes, Centre de Morphologie Mathématique. Ecole des Mines de Paris.

[53] Matheron G. (1965). Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux sciences de la nature. Paris, Masson.

[54] Matheron G. (1969). Cours de processus stochastiques, Ecole des Mines de Paris. Centre de géostatistique.

[55] Matheron, G. (1970). La théorie des variables régionalisées, et ses applications. , Les cahiers du centre de Morphologie Mathématique de Fontainebleau. Fasc. 5, ENSMP.

[56] Matheron, G. (1973). The intrinsic random functions and their applications. Adv. Appl. Prob. 5: 439-468.

[57] Meybeck M., J. Vogler, F. Moatar, H. Duerr, L. Laroche, and L. Lachartre. (2003). Analysis of temporal variability in river systems. E. C. Report, EUROCAT Programme, W. P. 5.1

[58] Moatar F. and M. Meybeck (2005). Compared performances of different algorithms for estimating annual nutrients load discharged by the eutrophic Loire. Hydrological processes, 19, 429-444.

[59] Monestiez, P., J.-S. Bailly, P. Lagacherie and M. Voltz (2005). Geostatistical modelling of spatial processes on directed trees: Application to fluvisol extent. Geoderma 128: 179-191.

[60] Monestiez, P., R. Habib and J.M. Audergon (1989). Estimation de la covariance et du variogramme pour une fonction aléatoire à support arborescent: application à l'étude des arbres fruitiers. Geostatistics, Kluwer Academic Publishers 1 : 39-56.

[61] Pereira, H. G., J. Ribeiro, A.J. Sousa, L. Ribeiro, A. Lopes, and J. Serôdio (2000). Forecasting river water quality indices. Geostats 2000 Cape Town.

[62] Pinheiro, J. C. and D. M. Bates (2000). Mixed effects models in S and Splus, Springer-Verlag.

[63] Rivoirard, J. (1994). Introduction to disjunctive kriging and non linear geostatistics, Clarenton.

[64] Ruelland D. (2004). SENEQUE, logiciel SIG de modélisation prospective de la qualité de l'eau. Revue Internationale de Géomatique. European Journal of GIS and Spatial analysis. 14 (**1**) : 1260-5875.

[65] SANDRE (Secrétariat d'administration national des données), (2002). Le référentiel hydrographique. Thème : INTER-THEMES. Version : 2002-1. Cellule d'animation SANDRE. Groupe référentiel SANDRE.

[66] Saporta, G. (1990). Probabilités. Analyse des données et statistique, Technip.

[67] Sauquet, E. (2000a). Une cartographie des écoulements annuels et mensuels d'un grand bassin versant structurée par la topologie du réseau hydrographique, Institut national polytechnique de Grenoble, Unité de Recherche Hydrologie-Hydraulique, Cemagref.

[68] Sauquet, E., L. Gottschalk, and E. Leblois (2000b). "Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme." Hydrological sciences journal 45(6): 799-815.

[69] Sauquet, E., I. Krasovskaia, and E. Leblois (2000c). "Mapping mean monthly runoff pattern using EOF analysis." Hydrology and earth system sciences 4(1): 79-93.

[70] Simonet, F. (2001). Le nouveau système d'évaluation de la qualité de l'eau des rivières: le SEQ-Eau. Revue de l'agence de l'eau Adour Garonne: 7-9.

[71] Smitz, J., E. Everbecq, J.F.. Deliege, J.P. Descy, R. Wollast, and J.P. Vanderborght. (1997). "PEGASE, une méthodologie et un outil de simulation prévisionnelle pour la gestion de la qualité des eaux de surface." tribune de l'eau 50(588): 73-82.

[72] Smola, A. J. and R. Kondor (2003). Kernels and Regularization on Graphs. COLT 2003.

[73] Strahler, A. N. (1964). Geology-Part 2, Handbook of applied Hydrology, McGraw-Hill, New York.

[74] Venables, W. N. and B. D. Ripley (1994). Modern applied statistics with Splus, Springer-Verlag.

[75] Ver Hoef, J. M., E. Peterson, and D. Theobald. (2006). Spatial Statistical Models that Use Flow and Stream Distance. Environmental and Ecological statistics (in press).

[76] Vert, J.-P., K. Tsuda, and B. Schölkopf (2004). A primer on kernel methods. Kernel Methods in Computational Biology, B. Schölkopf, K. Tsuda and J,-P vert (Eds), MIT Press, 131-154.

[77] Villeneuve, J.-P., G. Morin, B. Bobee, and D. LeBlanc (1979). Kriging in the design of streamflow sampling networks. Water resources research 15(6): 1833-1840.

[78] Wallin, M., T. Wiederholm, and R. Johnson (2002). Guidance on establishing reference conditions and ecological status class boundaries for inland surface waters. Third draft, CIS Working group 2.3- REFCOND. Agence de l'eau Loire-Bretagne.

[79] Zielinski, R. (2004). Optimal quantile estimators. Small sample approach. Institute of Mathematics Polish Academy of Sciences.

[80] Zielinski, R. (2005). L-Statistics as non parametric Quantile Estimators. Institute of Mathematics Polish Academy of Sciences.

Annexes

Annexe A :
Méthodes géostatistiques, variogramme et krigeage

Destinée au lecteur non familier à la géostatistique, cette annexe en présente les concepts principaux dans le cadre de séries temporelles. Volontairement, l'ensemble des développements théoriques utilisés et nécessaires à la compréhension du mémoire de thèse ne sont pas présentés car disponibles dans de nombreux ouvrages de géostatistique (Matheron, 1970 ; Chilès and Delfiner, 1999 ; Emery and Arnaud, 2000).

La première étape des méthodes géostatistiques consiste à inférer puis à modéliser les corrélations temporelles. On utilise pour cela des outils dits variographiques. Soit $Z(t)$ la concentration en nutriment pour une station. La corrélation temporelle décrit le degré de liaison entre deux variables $Z(t)$ et $Z(t+\tau)$, par exemple les concentrations à quinze jours d'intervalle τ. A la notion de corrélation, on préfère en géostatistique celle plus générale de variabilité, mesurée par l'amplitude de l'écart entre valeurs.

Généralement, les concentrations mesurées en deux dates proches se ressemblent plus que celles mesurées en deux dates éloignées, du moins en l'absence de fluctuations saisonnières introduisant une périodicité annuelle. L'écart entre les concentrations mesurées en deux dates espacées de τ est $Z(t+\tau)-Z(t)$. Pour quantifier l'amplitude de l'écart, il est plus commode de considérer l'écart quadratique $(Z(t+\tau)-Z(t))^2$. Le nuage de corrélation entre l'intervalle de temps τ et le demi-écart quadratique $\frac{1}{2}(Z(t+\tau)-Z(t))^2$ constitue la nuée variographique (Figure 1).

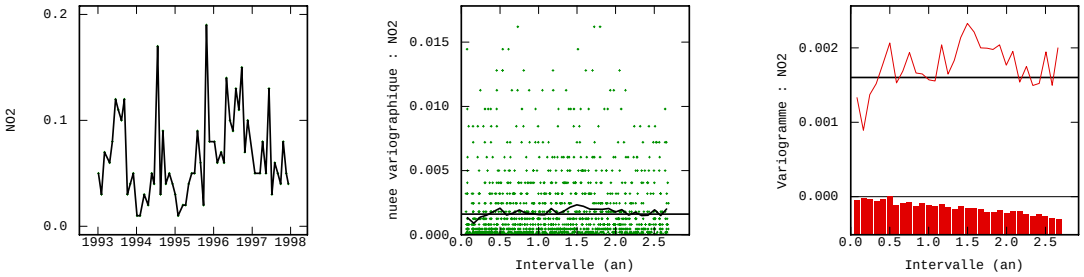


Figure 1: Mesure de la concentration en nitrites à la station 313000. A gauche : chronique des mesures, au milieu : nuée variographique avec report du variogramme, à droite : variogramme, avec en dessous, report de l'histogramme du nombre de couples par pas.

En l'absence de périodicité annuelle, une nuée variographique se présente comme un nuage dont l'enveloppe supérieure croît avec l'intervalle de temps τ. Le moyen le plus simple pour résumer la nuée variographique, consiste à en calculer la moyenne par classe d'intervalle de temps (voir la courbe tracée sur la nuée, (Figure 1). Cette courbe, généralement croissante, se stabilise parfois à partir d'un certain intervalle de temps. Elle fournit beaucoup d'information sur la variabilité temporelle de la variable étudiée. C'est le variogramme.

Le variogramme est la moyenne de la nuée variographique en fonction de l'intervalle de temps τ séparant deux dates de mesures.

Le variogramme représentant, au facteur 1/2 près, l'amplitude moyenne de l'écart quadratique, permet de quantifier la variabilité de la concentration, en fonction de l'intervalle de temps τ. Pour un même intervalle de temps τ, une variabilité plus forte se traduit par une plus forte valeur du variogramme.

A intervalle de temps nul, si l'on ne dispose que d'un seul échantillon par prélèvement, on a $\gamma(0) = 0$.

Comme un variogramme est toujours positif, sa croissance plus ou moins rapide aux petits intervalles de temps traduit la détérioration de la corrélation entre concentrations lorsque l'intervalle de temps séparant les dates augmente. Une périodicité annuelle de la concentration se traduit sur la nuée variographique et sur le variogramme par une composante périodique annuelle (voir par exemple figure de droite, Figure 1).

Le variogramme expérimental, calculé à partir des données, apporte beaucoup d'information sur la variabilité temporelle des concentrations :

- Stationnarité (Figure 2, gauche) : lorsque la concentration fluctue autour d'une moyenne constante, le variogramme oscille au-delà d'un certain intervalle de temps τ, autour d'une valeur constante. La portée est l'intervalle de temps correspondant à la zone d'influence d'un échantillon. Au-delà le variogramme est constant, et se cale au palier **;** il peut aussi présenter une composante périodique. Un variogramme toujours croissant pour la période considérée est symptomatique d'une non stationnarité à cette échelle d'observation.
- Superposition de différentes échelles temporelles de variation (Figure 2, droite): le plus souvent, plusieurs échelles de variabilité se superposent ; le variogramme se décompose en plusieurs structures gigognes, chaque composante stationnaire ayant sa propre portée et son propre palier.

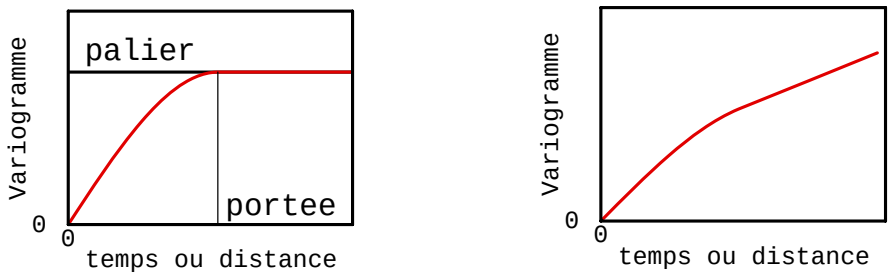


Figure 2 : Quelques modèles de variogramme. A gauche, palier et portée pour un modèle élémentaire, à droite : exemple de deux structures gigognes.

Le variogramme est calculé expérimentalement à partir des données disponibles aux dates de mesures t_i, comme la moitié de l'écart quadratique moyen en fonction de l'intervalle de temps τ :

$$\gamma(\tau) = \frac{1}{2}\frac{1}{n(\tau)}\sum_i (z(t_i+\tau) - z(t_i))^2$$

avec : - $n(\tau)$ nombre de couples de dates séparées de τ ;

- $(z(t_i+\tau) - z(t_i))^2$ écart quadratique des concentrations aux dates t_i et $t_i+\tau$
- $\frac{1}{n(\tau)}\sum_i (z(t_i+\tau) - z(t_i))^2$ moyenne de ces écarts quadratiques pour tous les couples de dates distantes de τ disponibles

Le coefficient 1/2 simplifie les calculs ultérieurs.

Un variogramme expérimental n'ayant pas les propriétés mathématiques requises, il est nécessaire de l'ajuster par un modèle. C'est une « fonction variogramme » fournissant les valeurs du variogramme pour tous les intervalle de temps τ, dans le domaine de validité de l'ajustement. Cet ajustement conclut l'analyse variographique. Il peut si nécessaire être automatisé, ce que nous avons fait laissant tout de même le choix à l'utilisateur de spécifier le modèle à ajuster, l'intervalle de temps à considérer.

Remarque : la fonction variogramme ou « variogramme théorique », est la version probabiliste du variogramme expérimental, en supposant ce dernier connu parfaitement. La moyenne temporelle (ou spatiale) des écarts quadratiques $(z(t+\tau)-z(t))^2$ de la variable régionalisée z est remplacée par la moyenne probabiliste, c'est-à-dire l'espérance des écarts quadratiques $(Z(t+\tau)-Z(t))^2$ de la Fonction Aléatoire Z. Le variogramme théorique est donc défini par

$$\gamma(\tau)=\frac{1}{2}E\left[(Z(t+\tau)-Z(t))^2\right]$$

On montre que le variogramme modélise de façon synthétique la structure temporelle ou spatiale de la Fonction aléatoire Z. Ainsi, le comportement du variogramme théorique au voisinage de l'origine $\tau=0$ modélise la régularité de Z, et son comportement aux grandes valeurs de τ considérées la stationnarité de Z pour le domaine temporel (ou spatial) considéré (voir Matheron, 1970).

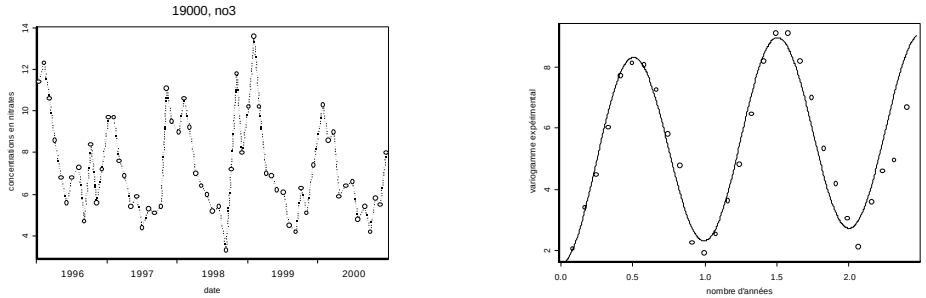


Figure 3: Mesures des concentrations en nitrates à la station 19000 entre 1996 et 2000, à gauche, et à droite, variogramme temporel expérimental associé (points) et ajustement par une fonction variogramme (courbe continue)

Le variogramme est un concept très général, définissable à plusieurs dimensions d'espace. Différents modèles spatio-temporels ont été développés.

Le krigeage

Pour estimer la concentration *Z* à une date t à partir des mesures effectuées aux dates $t_i, i=1,...,n$ le plus simple est de construire un estimateur linéaire, c'est-à-dire de la forme

$$Z^*(t)=\lambda_1 Z(t_1)+\lambda_2 Z(t_2)+...+\lambda_n Z(t_n)$$

où pour éviter les biais d'estimation, les pondérateurs λ_i sont de somme unité :

$$\lambda_1 + \lambda_2 + ... + \lambda_n = 1$$

Deux estimateurs peuvent être comparés expérimentalement de la façon suivante : on construit l'estimateur $Z^*(t)$ pour chacun d'eux, puis on réalise la mesure de $Z(t)$. On compare alors les erreurs d'estimation $Z^*(t) - Z(t)$ auxquelles conduisent ces deux estimateurs. Il est difficile de conclure sur la base d'une seule expérience. Mais si on la répète un nombre suffisant de fois pour des configurations similaires de *t* et des dates d'échantillonnage $t_1, t_2, \ldots, t_n$, on pourra mesurer lequel des deux estimateurs est le plus efficace, du moins pour ces configurations. On pourra retenir par exemple celui qui conduit le plus souvent à l'erreur la plus faible (en valeur absolue), ou celui pour lequel la valeur absolue de l'erreur est en moyenne la plus faible. En géostatistique on retient celui pour lequel le carré de l'erreur est en moyenne le plus faible. Pourquoi ? Parce que, une fois le variogramme modélisé, on est capable de calculer théoriquement la moyenne du carré de l'erreur correspondant à une infinité de situations analogues, sans avoir besoin de réaliser de nouvelles mesures. Si on adopte ce critère, le krigeage est l'estimateur optimum parmi tous les estimateurs linéaires : en moyenne il conduit à une erreur quadratique plus faible que n'importe quel autre estimateur linéaire.

Les *n* pondérateurs de krigeage s'obtiennent en résolvant un système linéaire de *n* + 1 équations à *n* + 1 inconnues, les *n* pondérateurs, plus un paramètre de Lagrange (μ) qui s'introduit à cause de la condition $\lambda_1 + \lambda_2 + ... + \lambda_n = 1$ (2-18). Ce système a la forme suivante :

$$\begin{cases} \lambda_1\gamma(0) + \lambda_2\gamma(t_2 - t_1) + ... + \lambda_n\gamma(t_n - t_1) + \mu = \gamma(t - t_1) \\ \lambda_1\gamma(t_1 - t_2) + \lambda_2\gamma(0) + ... + \lambda_n\gamma(t_n - t_2) + \mu = \gamma(t - t_2) \\ \qquad\qquad\qquad\qquad ... \qquad\qquad\qquad\qquad . \\ \lambda_1\gamma(t_1 - t_n) + \lambda_2\gamma(t_2 - t_n) + ... + \lambda_n\gamma(0) + \mu = \gamma(t - t_n) \\ \lambda_1 + \lambda_2 + ... + \lambda_n = 1 \end{cases}$$

La partie gauche du système de krigeage fait intervenir les termes $\gamma(t_j - t_i)$ et la partie droite les termes $\gamma(t - t_i)$. Autrement dit les poids de krigeage dépendent de la configuration relative des dates de mesure et de la date *t* pour laquelle on effectue l'estimation. Ils dépendent aussi de la variabilité temporelle des concentrations synthétisée par le variogramme.

Le krigeage donne également la moyenne quadratique de l'erreur ou variance de krigeage $\sigma_k^2 = Var\left(Z^*(t) - Z(t)\right)$, dont la racine carrée est l'écart-type de krigeage σ_k . Sous certaines conditions, cet écart-type peut être utilisé pour définir des intervalles de confiance.

Au lieu de la concentration $z(t)$ à un instant fixé, on peut estimer directement une moyenne temporelle $z^{an} = \frac{1}{T}\int_t^{t+\tau} z(u)du$. On montre qu'à condition de conserver les mêmes données pour les différentes estimations, le krigeage d'une moyenne sur *N* points est égal à la moyenne des *N* estimations par krigeage ponctuel. De même, le krigeage d'une intégrale temporelle est égal à l'intégrale des valeurs krigées. *K* désignant l'estimation par krigeage,

$$\begin{aligned} (z^{an})^K &= \left(\frac{1}{T}\int_t^{t+\tau} z(u)du\right)^K \\ &= \frac{1}{T}\int_t^{t+\tau} z^K(u)du \end{aligned}$$

Lorsque l'hypothèse de stationnarité de l'espérance mathématique est pertinente, on peut également estimer cette espérance par krigeage. On verrait alors que l'estimation de l'espérance diffère de celle de la moyenne temporelle annuelle. Le détail des équations de krigeage de la moyenne annuelle et des différences interannuelles ainsi que les différents abaques utiles à la programmation dans le cadre de cette étude sont donnés en fin d'annexe.

Signification de la variance de krigeage

La variance de krigeage doit être comprise comme la moyenne des erreurs quadratiques qui peuvent être observées sur des configurations $t_1, t_2, ... t_n$ similaires à la configuration étudiée ; par similaires, il faut comprendre identiques à une translation près. Elle n'est pas conditionnelle aux valeurs observées $z(t_1), z(t_2), ..., z(t_n)$, ce qui peut être vu comme une limitation du krigeage, qui pour être levée demanderait de recourir à l'espérance conditionnelle. Ce dernier estimateur n'est généralement pas aisé à calculer et suppose une information considérablement plus riche que le variogramme sur la variabilité temporelle du phénomène étudié.

Krigeage de la moyenne annuelle

On désigne par $Z(t)$ la fonction aléatoire associée à la variable temporelle $z(t)$, concentration en fonction du temps.

Soit T un intervalle de temps (on prend T= 1 an dans les exemples du rapport)

On note $Z(T)$ la variable aléatoire obtenue en intégrant sur T :

$$Z(T) = \frac{1}{T}\int_T Z(t)dt$$

On veut estimer $Z(T)$ par krigeage.

On note $Z^*(T)$ cet estimateur.

L'estimateur par krigeage est linéaire, il s'écrit donc $Z^*(T) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(t_\alpha)$

L'estimateur par krigeage est sans biais :

$$\begin{aligned} E(Z^*(T) - Z(T)) &= E(\sum_{\alpha=1}^{n} \lambda_\alpha Z(t_\alpha) - Z(T)) \\ &= m(\sum_{\alpha=1}^{n} \lambda_\alpha - 1) \end{aligned}$$

Ce qui implique la condition suivante : $\sum_{\alpha=1}^{n} \lambda_\alpha = 1$

La variance d'estimation est minimale :

$$\begin{aligned} Var(Z^*(T) - Z(T)) &= Var(Z^*(T)) + Var(Z(T)) - 2Cov(Z^*(T), Z(T)) \\ &= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(t_\alpha - t_\beta) + \overline{C}(T,T) - 2\sum_{\alpha=1}^{n} \lambda_\alpha \frac{1}{T}\int_T C(t_\alpha - t)dt \end{aligned}$$

On introduit un paramètre de Lagrange μ pour minimiser la variance d'estimation.

Il faut donc minimiser :

$$\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta C(t_\alpha - t_\beta) + \bar{C}(T,T) - 2\sum_{\alpha=1}^{n}\lambda_\alpha \frac{1}{T}\int_T C(t_\alpha - t)dt + 2\mu(\sum_{\alpha=1}^{n}\lambda_\alpha - 1)Z(T)$$

On dérive par rapport à $\lambda_1, \lambda_2, ...\lambda_n$, on obtient alors le système suivant :

$$\forall\alpha \in \{1,...n\}, \sum_{\beta=1}^{n}\lambda_\beta C(t_\alpha - t_\beta) + \mu = \frac{1}{T}\int_T C(t_\alpha - t)dt$$

Dans le cas de fonctions aléatoires stationnaires ou intrinsèques, La covariance C peut être remplacée par le variogramme γ dans les équations de krigeage. Ainsi on a par exemple :

$$\forall\alpha \in \{1,...n\}, \sum_{\beta=1}^{n}\lambda_\beta \gamma(t_\alpha - t_\beta) + \mu' = \frac{1}{T}\int_T \gamma(t_\alpha - t)dt \text{ avec } \mu' = -\mu$$

La résolution du système fournit les poids et donc l'estimation de la moyenne sur T et sa précision.

<u>Modèles de variogrammes et abaques associés</u>

Pour simplifier la programmation, on utilise des fonctions auxiliaires χ et F. Elles sont définies de la manière suivante :

$$\chi(L) = \frac{1}{L}\int_0^L \gamma(r)dr$$

$$\begin{aligned} F(L) &= \bar{\gamma}(L,L) \\ &= \frac{1}{L^2}\int_0^L\int_0^L \gamma(u-u')dudu' \\ &= \frac{2}{L^2}\int_0^L u\chi(u)du \end{aligned}$$

χ et F sont des fonctions paires.

<u>Schéma sphérique à une dimension de portée a et de palier C</u>

$$\gamma(h) = \begin{cases} C\left(\frac{3r}{2a} - \frac{r^3}{2a^3}\right) & \forall r \in [0,a] \\ C \quad r \geq \text{a} \end{cases}$$

Si $0 \leq L \leq a$, $\chi(L) = C \times \left(\frac{3L}{4a} - \frac{L^4}{8a^3}\right)$ et $F(L) = \frac{1}{2}\frac{L}{a} - \frac{1}{20}\frac{L^3}{a^3}$

Si $L \geq a$, $\chi(L) = C \times \left(1 - \frac{3a}{8L}\right)$ et $F(L)1 - \frac{3a}{4L} + \frac{a^2}{5L^2}$

Schéma cosinus de période a et d'amplitude 2C

$$\gamma(r) = C\left(1 - \cos\left(\frac{2\pi r}{a}\right)\right)$$

Si $L = 0$, $\chi(L) = F(L) = 0$

Si $L > 0$, $\chi(L) = C - \frac{Ca}{2\pi L}\sin\left(\frac{2\pi L}{a}\right)$ et $F(L) = C + \frac{2Ca^2}{(2\pi)^2 L^2}\left(\cos\frac{2\pi L}{a} - 1\right)$

Schéma pépitique de palier C

$$\gamma(h) = \begin{cases} 0 \text{ si } h = 0 \\ C \text{ si } h > 0 \end{cases}$$

$$\chi(L) = F(L) = C$$

Schéma linéaire de paramètre C

$$\gamma(h) = C|h|$$

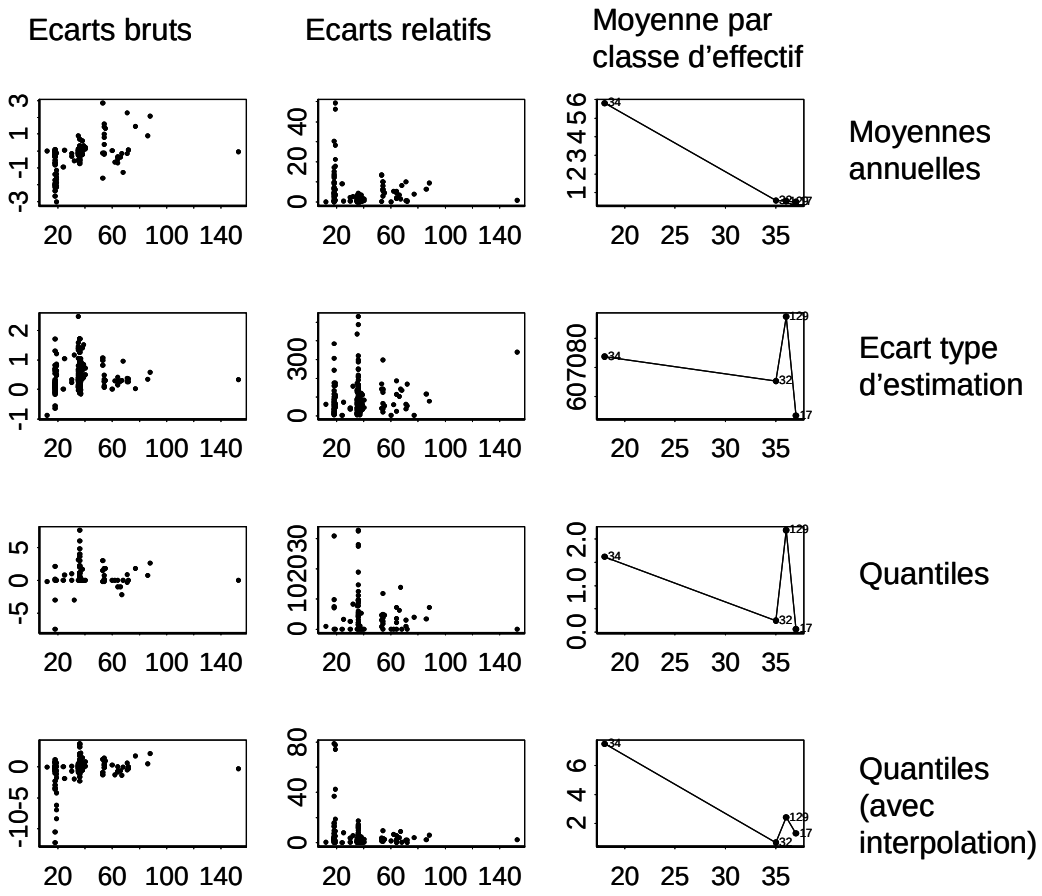
$$F(L) = C\left|\frac{L}{3}\right| \text{ et } \chi(L) = C\left|\frac{L}{2}\right|$$

Annexe B :
Calculs des indicateurs sur trois années et entre 1985 et 2005

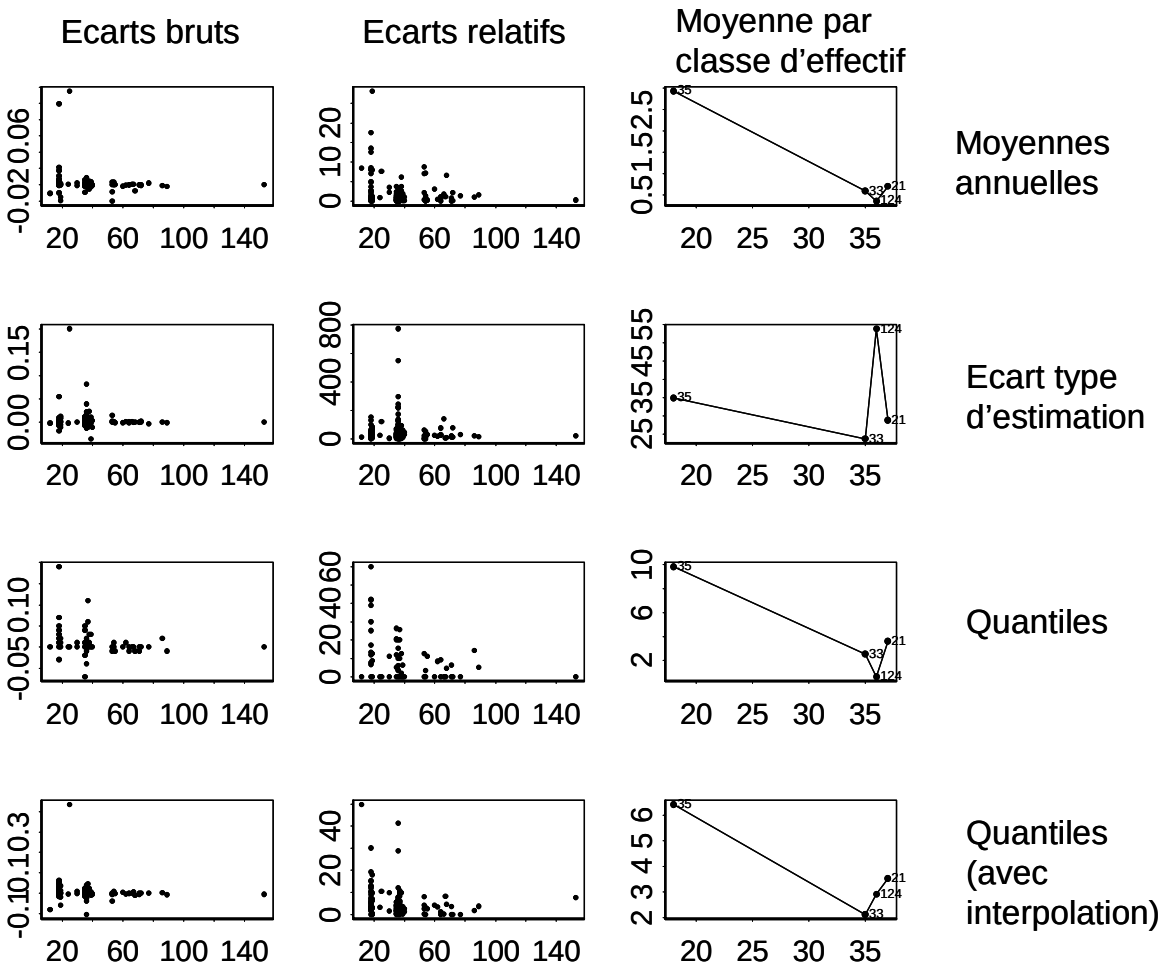
Cette annexe présente les comparaisons entre les estimations géostatistiques et statistiques des indicateurs pour un calcul de moyenne et quantile entre 2002 et 2004 (le nombre de mesures est ainsi triplé pour les estimations) et pour un calcul de moyenne et quantile par année entre 1985 et 2005.

Calcul sur trois ans (2002-2004)

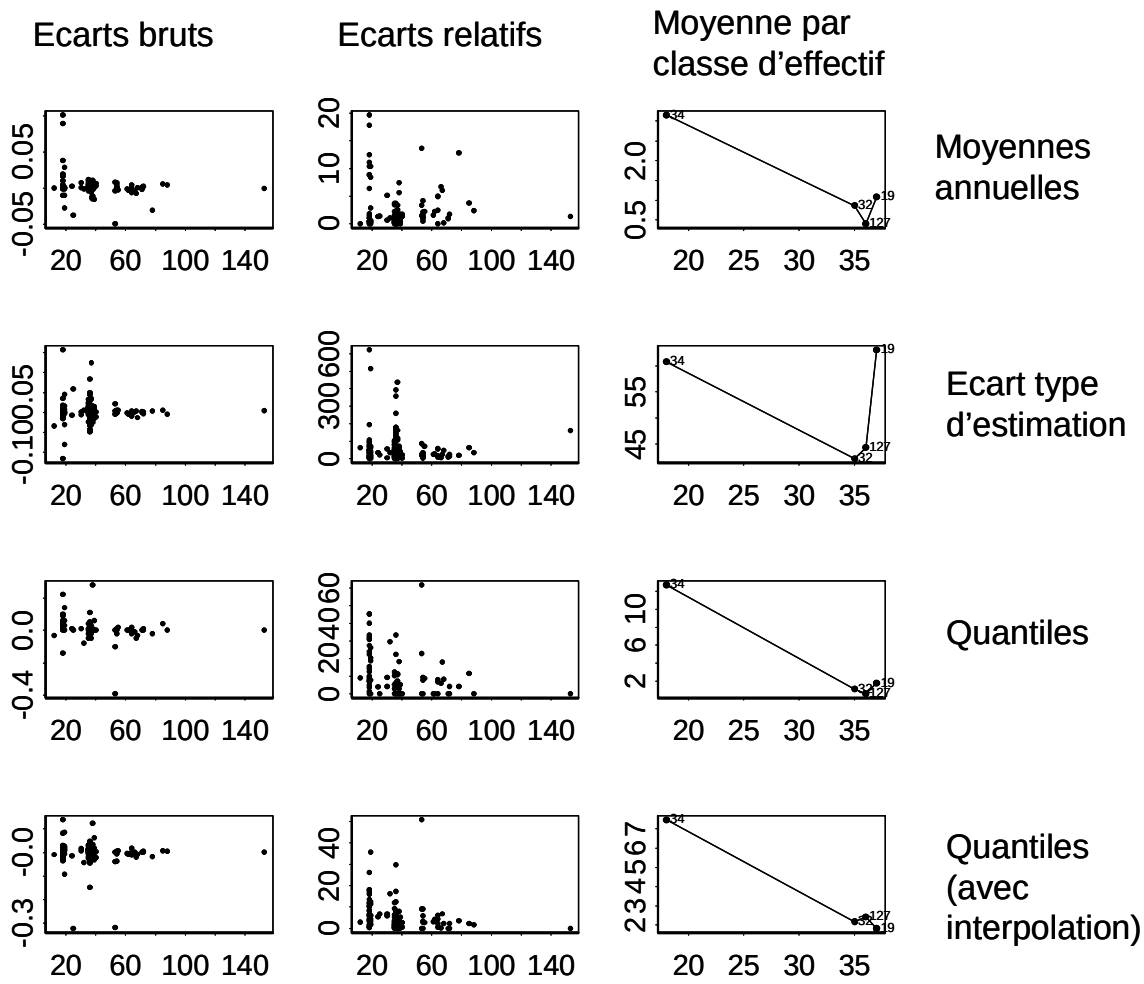
Nitrates



Nitrites

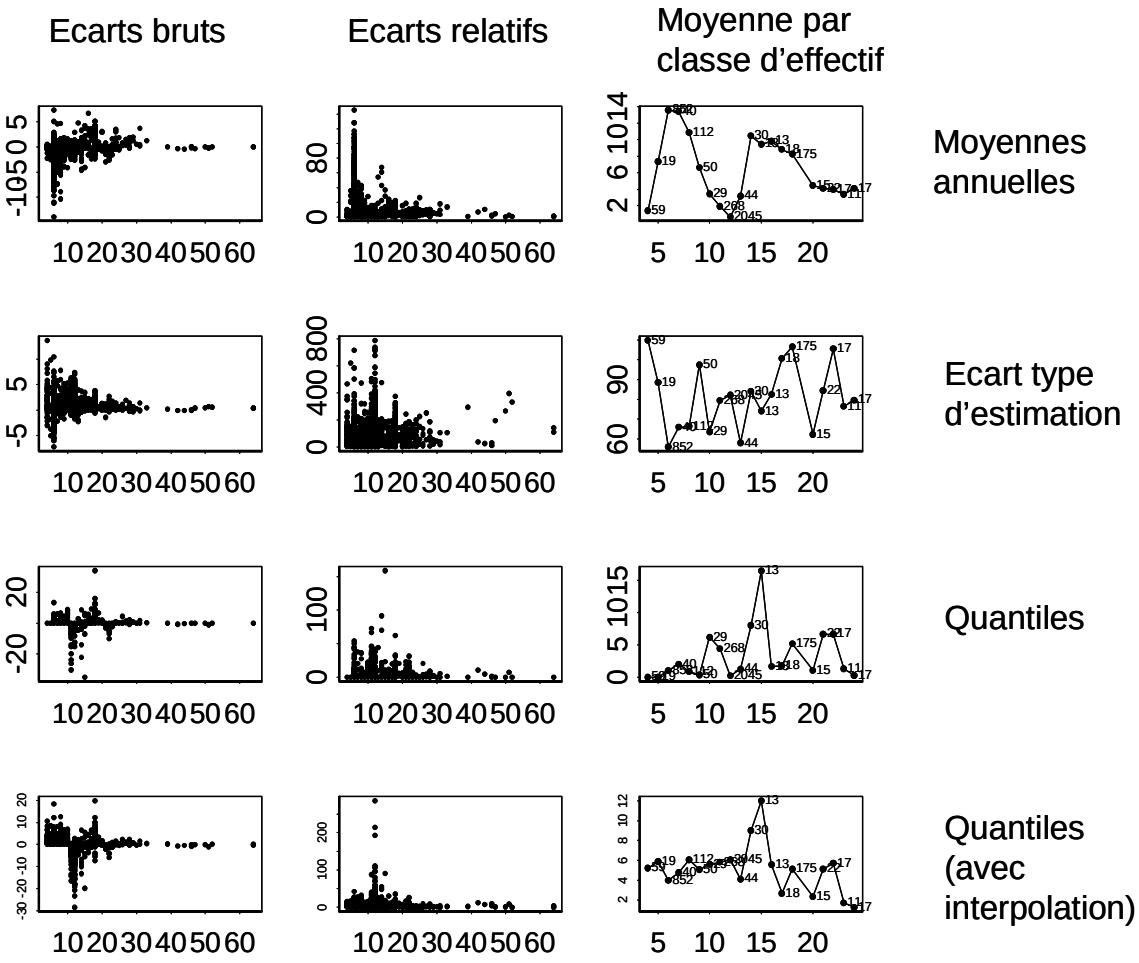

Ecarts bruts
Ecarts relatifs
Moyenne par classe d'effectif
Moyennes annuelles
Ecart type d'estimation
Quantiles
Quantiles (avec interpolation)

Orthophosphates

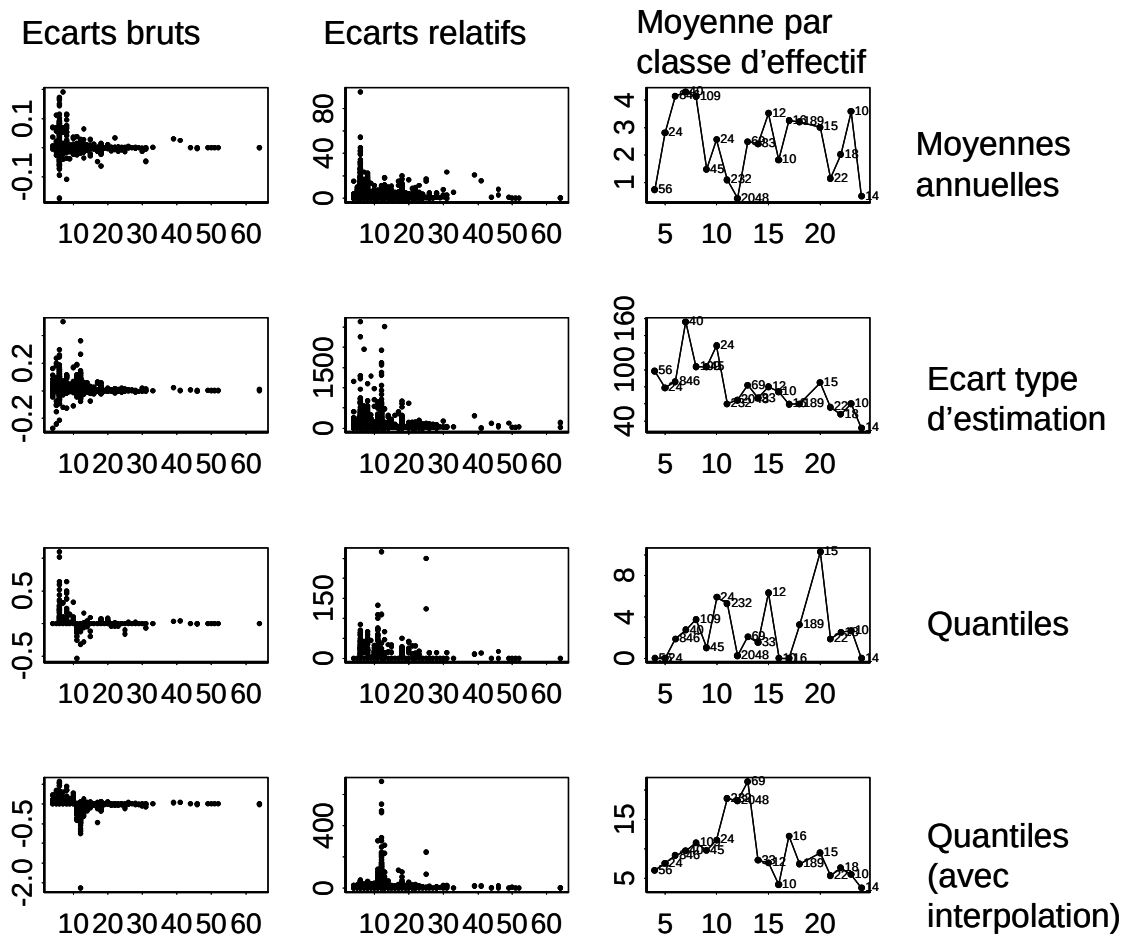


Par année entre 1985 et 2005

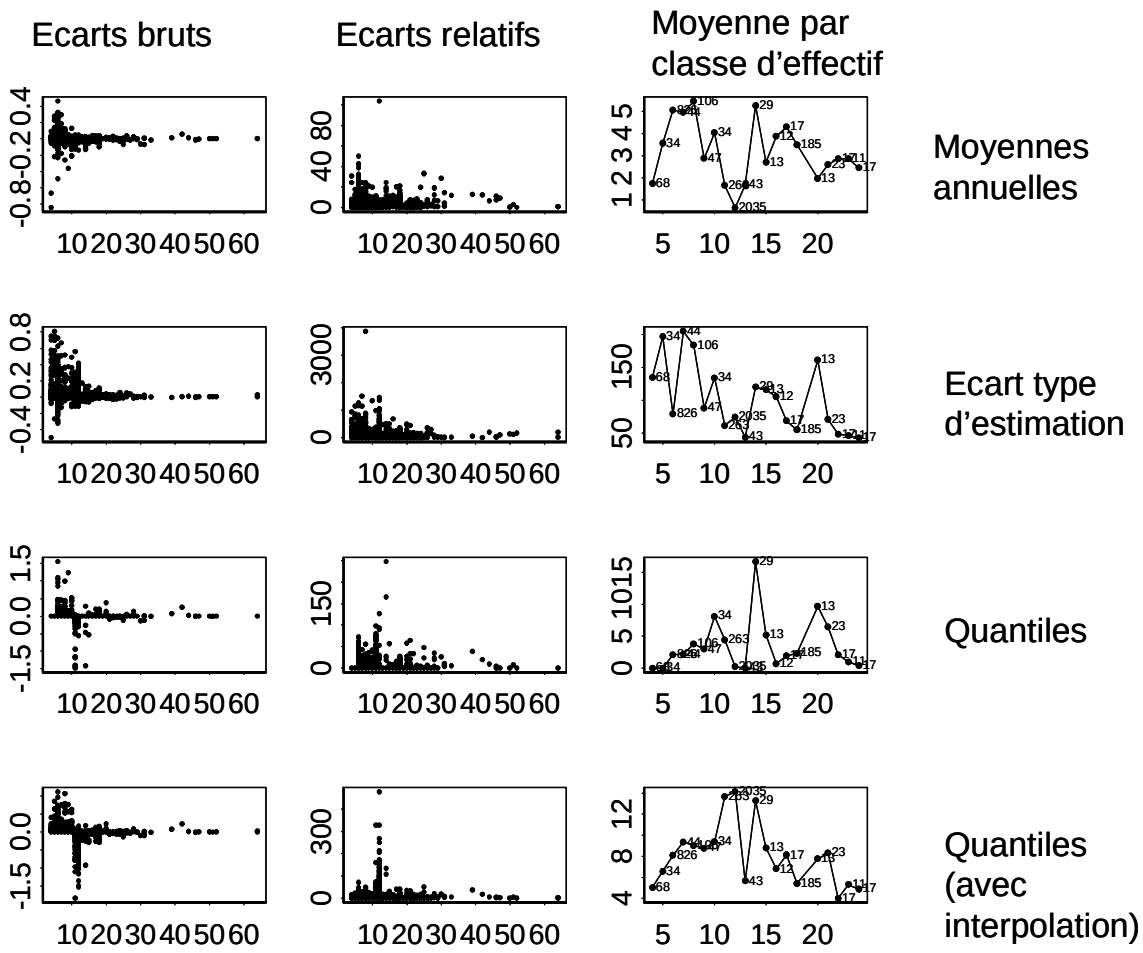
Nitrates



Nitrites


Ecarts bruts
Ecarts relatifs
Moyenne par classe d'effectif
Moyennes annuelles
Ecart type d'estimation
Quantiles
Quantiles (avec interpolation)

Orthophosphates



Annexe C :
Publication Comptes Rendus Géosciences

Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Geoscience 338 (2006) 307–318
COMPTES RENDUS GEOSCIENCE
http://france.elsevier.com/direct/CRAS2A/

Géosciences de surface (Hydrologie–Hydrogéologie)

Modèles géostatistiques de concentrations ou de débits le long des cours d'eau

Chantal de Fouquet *, Caroline Bernard-Michel

Centre de géostatistique, École des mines de Paris, 35, rue Saint-Honoré, 77305 Fontainebleau, France

Reçu le 11 février 2005 ; accepté après révision le 2 février 2006

Disponible sur Internet le 3 mai 2006

Présenté par Ghislain de Marsily

Résumé

Estimer les concentrations ou les débits le long d'un réseau hydrographique nécessite des covariances ou des variogrammes valides sur un support arborescent. Nous généralisons deux modèles récemment proposés dans la littérature, en les étendant notamment au cas intrinsèque. Nous proposons une construction par « rivière », c'est-à-dire sur les chemins reliant les sources à l'exutoire. La combinaison de fonctions aléatoires (FA) monodimensionnelles stationnaires ou intrinsèques fournit, sur le réseau, des modèles stationnaires ou intrinsèques par arête, avec des discontinuités aux confluences. La construction de l'exutoire vers les sources fournit des modèles stationnaires ou intrinsèques par rivière, sans discontinuité aux confluences, avec le variogramme linéaire comme cas particulier. L'extension au modèle linéaire de corégionalisation est immédiate. ***Pour citer cet article : C. de Fouquet, C. Bernard-Michel, C. R. Geoscience 338 (2006).***

© 2006 Académie des sciences. Publié par Elsevier SAS. Tous droits réservés.

Abstract

Geostatistical models for concentrations or flow rates in streams. Estimating concentrations or flow rates along a stream network requires specific models. Two classes of models, recently proposed in the literature, are generalized, to the intrinsic case in particular. We present a global construction by 'streams', i.e. on the whole set of paths between sources and outlet. Combining stationary or intrinsic one-dimensional random functions leads to stationary or intrinsic models on segments, with discontinuities at the forks. A construction from outlet to sources, leads to stationary or intrinsic models on each stream, without any discontinuity at the forks. The linear variogram is found as a particular case. The extension to the linear model of coregionalization is immediate, allowing a multivariate modelling of concentrations. ***To cite this article: C. de Fouquet, C. Bernard-Michel, C. R. Geoscience 338 (2006).***

© 2006 Académie des sciences. Publié par Elsevier SAS. Tous droits réservés.

Mots-clés : Géostatistique ; Support arborescent ; Réseau hydrographique ; Concentrations ; Débits ; Fonction aléatoire

Keywords: Geostatistics; Tree graph; Stream network; Concentrations; Rate of flow; Random function

Abridged English version

Estimating concentrations of pollutants along a stream network makes it necessary to take into ac-

* Auteur correspondant.
Adresse e-mail : chantal.de_fouquet@ensmp.fr (C. de Fouquet).

1631-0713/$ – see front matter © 2006 Académie des sciences. Publié par Elsevier SAS. Tous droits réservés.
doi:10.1016/j.crte.2006.02.002

count the geometry of the network. Usual geostatistical models as the spherical covariance were developed for Euclidean space, and are no more valid on tree graphs. Generalizing recent models [2,11,12] built on stream segments, we present a construction combining one-dimensional Random Functions (RF) defined on each path between sources and the outlet.

Ver Hoef et al. [12] developed a class of valid models derived from the one-dimensional moving average method. Integrating a kernel function upstream from a location (Eq. (11bis), Annex 2), they obtain a random function Z whose values upstream of any fork are independent on the different parts of the network. The moving average is calculated by distributing the kernel function on the upstream segments with a proper weighting, to ensure that the variance is constant (Fig. 2 and Eq. (12)). On each segment, the covariance C_1, derived from the kernel function, is stationary. Between two 'stream connected' points s_i and t_j, the covariance is equal to C_1 up to a weighting term, depending on the weights attached to the segments ending at the nodes lying between s_i and t_j (Eq. (6)).

This model can be simplified and generalized, considering the one-dimensional random functions Y_I defined on each path linking one source to the outlet. When different paths join at a node, the resulting random function Z downstream is a linear combination of the corresponding Y_I (Eq. (1)) using their respective weights. When the kernel function is defined on the half-line, this model is equivalent to the Ver Hoef one. But any one-dimensional random function model can now be used for the Y_I: for example, the kernel function can be symmetric. This model is easily extended when the different Y_I are spatially correlated: the values of Z on two segments upstream from a node are now correlated. Intrinsic RF Y_I can be considered too. This can be useful, for example, to describe flow rates that usually increase along the stream.

In the previous models, the RF Z is discontinuous at each node, in the mean square sense (or equivalently, its variogram or covariance is discontinuous at the nodes) and stationary or intrinsic on each segment between nodes. For points belonging to different segments, the variogram or the covariance depends on the weights.

To obtain RF without discontinuities at the nodes along each stream, Monestiez et al. [11] and Bailly et al. [2] constructed a model running in the opposite sense, from the outlet to the sources. It is based on a conditional independence between parts of the network upstream a node, knowing Z downstream, between the node and the outlet. The covariance of Z along each stream can be any one-dimensional covariance. The covariance between points on two different streams is given in the present paper (Eq. (10)). Extending the model to variograms and thus to intrinsic random functions on each stream, we obtain the linear variogram as a particular case.

We prove that among the covariances or variograms, which only depend on the distance along the tree, only the exponential and the linear scheme are consistent with the conditional independence hypothesis, and can then be constructed this way.

All the previous models can be incorporated in a linear coregionalization model defined on a stream network for a multivariate modeling of concentrations.

1. Introduction

Estimer des flux, des concentrations ou la quantité de matières en suspension [6] dans un réseau hydrographique à partir des mesures aux stations nécessite de tenir compte de la géométrie du réseau. Or les modèles géostatistiques usuels, développés dans des espaces euclidiens de dimensions deux ou trois, voire quatre pour les phénomènes spatiotemporels, ne sont généralement plus valides pour des variables définies sur un « support arborescent ». En effet, les théorèmes de Bochner et de Schoenberg font explicitement intervenir la distance euclidienne dans la caractérisation spectrale des covariances ou des variogrammes (voir par exemple [3]).

Utilisant la distance curviligne le long d'un réseau hydrographique, Ver Hoef et al. [12] montrent ainsi que, pour le schéma sphérique, certaines valeurs propres de la matrice de covariance peuvent être négatives. Comme conséquence pratique, les variances calculées avec ce « modèle » peuvent également devenir négatives. D'où la nécessité de modèles de covariances ou de variogrammes adaptés à la topologie des graphes.

Les fondements théoriques ont été récemment établis pour des fonctions aléatoires (FA) définies sur les sommets d'un graphe [7]. Le problème se pose différemment pour une FA définie sur un support continu, c'est-à-dire en tout point des arêtes du graphe. Seuls les graphes à structure « d'arbre » sont considérés dans la suite, ce qui exclut le cas de canaux reliant différents cours d'eau ou réseaux hydrographiques. Deux classes de FA ont été proposées récemment, fondées sur une construction des sources vers l'exutoire [12] ou, en sens inverse, de l'exutoire vers les sources [2,11].

Nous proposons ici une construction par combinaison de FA monodimensionnelles quelconques. Les modèles ainsi obtenus sur l'arbre présentent des discontinuités aux confluences ; ils généralisent le modèle Ver Hoef [12], qui correspond au cas particulier où la co-

variance est l'autoconvoluée d'une fonction définie sur la demi-droite. Des résultats complémentaires sont ensuite donnés pour le modèle Bailly–Monestiez [2,11], continu aux confluences. Ces deux classes sont généralisées au cas intrinsèque, les modèles non stationnaires permettant notamment de représenter des débits, généralement croissants dans le sens de l'écoulement.

2. Éléments bibliographiques

2.1. Différents modèles de FA sur un arbre

À la suite de Monestiez [10], pour modéliser le résidu sec des fruits sur un pêcher, Audergon et al. [1] calculent les variogrammes expérimentaux selon quatre distances définies sur l'arbre. Retenant comme distance le nombre d'embranchements entre deux points, les auteurs ajustent une covariance exponentielle, et effectuent un krigeage. Lorsque plusieurs distances apparaissent pertinentes, le choix de la plus appropriée est une étape importante de la modélisation.

Pour modéliser la largeur du fluvisol dans la partie aval du réseau hydrographique de l'Hérault [11] ou les fossés de drainage du bassin de Roujan [2], ces auteurs construisent une FA sur un arbre, de l'exutoire vers les sources, en posant une hypothèse d'indépendance conditionnelle entre points situés sur des cours d'eau différents, connaissant l'ensemble des valeurs à l'aval de leur confluence. Par cours d'eau, tous les modèles de covariance monodimensionnelle sont admissibles ; entre points situés sur des cours d'eau différents, la covariance n'est pas stationnaire. Des simulations conditionnelles sont présentées.

Ver Hoef et al. [12] proposent une construction en sens inverse, des sources vers l'exutoire. Posant une hypothèse d'indépendance entre rivières à l'amont de leur confluence, ces auteurs adaptent le procédé classique des moyennes mobiles, en répartissant sur les arêtes, à l'amont des confluences, le « noyau » défini sur la demi-droite. Ils estiment ensuite par krigeage (ponctuel ou de bloc) la concentration en métaux lourds le long d'un réseau hydrographique. Cressie et al. [5] reprennent ce modèle en combinant distance euclidienne et distance curviligne.

Cressie et al. [4] présentent une modélisation spatio-temporelle de la concentration en nitrates le long d'une rivière drainant un bassin versant. La structure du réseau hydrographique n'intervient pas dans la modélisation effectuée en deux dimensions, le temps et une dimension d'espace, via l'abscisse curviligne le long de la rivière. Une dérive multiple permet de prendre en compte la saisonnalité ainsi que de nombreuses variables décrivant le « milieu ».

Le développement de modèles de FA définies sur un support arborescent répond notamment à la nécessité d'une quantification précise des concentrations en différents polluants le long des cours d'eau. Seul l'aspect spatial est ici examiné.

2.2. Stationnarité

Sauf mention contraire, on examine les propriétés d'ordre deux des FA, c'est-à-dire leur covariance ou leur variogramme.

Une covariance (respectivement un variogramme) est dite stationnaire si elle ne dépend que de la distance curviligne entre deux points du graphe, et non stationnaire lorsqu'elle dépend des deux points séparément. Si la covariance sur les arêtes dépend de leur orientation, celle sur l'ensemble du graphe est considérée comme non stationnaire.

3. Combinaison de FA monodimensionnelles

Sur tout arbre, nous construisons une classe de FA par combinaison de « processus » définis sur des segments.

3.1. Principe de la construction

Soient deux affluents, de débits respectifs d_1, d_2 et de concentrations c_1 et c_2 immédiatement à l'amont de leur confluence ; juste à l'aval, le débit $d = d_1 + d_2$ et la concentration $c = \frac{d_1}{d_1+d_2}c_1 + \frac{d_2}{d_1+d_2}c_2$ s'obtiennent par combinaison linéaire des variables définies sur les affluents. À la confluence, le débit et la concentration présentent une discontinuité.

Prolongeons les affluents en les considérant comme des « filets d'eau » distincts, dont la réunion forme les « rivières ». Pour chaque filet, nous définissons, de la source à l'exutoire, un débit et une concentration, fonctions de l'abscisse curviligne comptée depuis l'exutoire. À l'aval des confluences successives depuis les sources, les débits des filets se cumulent pour former le débit total, et la concentration dans la rivière se ramène à une combinaison des concentrations des différents filets, en proportion de leur débit relatif. On en déduit un procédé général de construction de FA sur un arbre.

Lorsque le rapport des débits des « filets » reste constant entre deux confluences, les coefficients de la combinaison linéaire des concentrations des « filets » sont constants par arête, et changent aux confluences. Nous examinerons d'abord ce modèle simplifié.

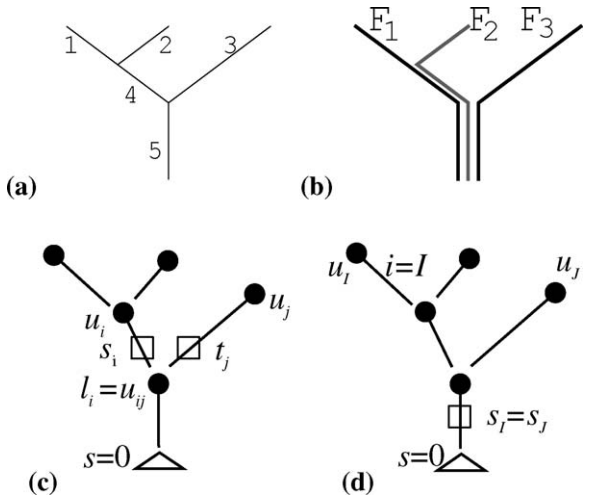


Fig. 1. Description de l'arbre. (**a**) Définition des arêtes, et (**b**) des « rivières ». (**c**) Notation par arête ou (**d**) par rivière.

Fig. 1. Description of the tree. (**a**) Definition of segments, and (**b**) of streams. (**c**) Notation from segments and (**d**) from streams.

3.2. Définitions et notations

Nous utilisons la terminologie géographique usuelle, complétée par celle relative aux graphes. Certaines notations sont reprises de [12].

Le réseau hydrographique est représenté par un arbre dont les sommets sont les sources, les confluences ou l'exutoire, supposé unique. Une « rivière » désigne un chemin d'une source vers l'exutoire ; le nombre de rivières est égal à celui des sources. Toute arête est caractérisable par les indices des rivières qui la traversent (Fig. 1a et b).

Deux points sont reliés (au fil de l'eau), si l'un est à l'aval de l'autre ; ils appartiennent alors à une ou à plusieurs rivières communes. Deux points non reliés sont situés sur des rivières différentes, à l'amont de leur confluence. Par analogie, deux arêtes sont « reliées » si l'une est à l'aval de l'autre, et non reliées dans le cas contraire.

Les rivières sont indicées en majuscules par leur source ; les arêtes sont notées en minuscules. L'arête à l'aval immédiat d'une source est aussi indicée par la source. Toute rivière a donc le même numéro que son arête amont. Tout point du réseau hydrographique, noté s_i ou s_J, est repérable par sa distance curviligne s, comptée positivement depuis l'exutoire (où $s = 0$), ainsi que par le numéro i de son arête ou celui J d'une des rivières passant par cette arête (Fig. 1c et d). Le sommet amont u_i appartient à l'arête i ; le sommet aval, considéré comme l'amont de l'arête suivante dans le sens du courant, n'appartient pas à l'arête i. L'ensemble des indices des arêtes à l'amont de l'arête i, excluant i, est noté U_i et l'ensemble des indices des arêtes à l'aval de l'arête i, incluant cette arête, D_i. L'ensemble B_{ij} des indices des arêtes situées entre s_i et t_j est :

- vide si les arêtes ne sont pas reliées ;
- égal à $(U_i \cup U_j) \cap (D_i \cup D_j)$ pour deux arêtes reliées, celle à l'*amont étant incluse, et celle à l'aval, exclue*.

L'ensemble des indices des *arêtes* sur la rivière J à l'amont de l'arête i (cette arête étant exclue) est noté B_{iJ}. V_i désigne l'ensemble des indices des *rivières* passant par l'arête i. L'abscisse de la confluence des rivières I et J est notée u_{IJ}, et l'abscisse de la confluence des rivières passant par les arêtes i et j, u_{ij}. La distance curviligne le long de l'arbre est alors :

- $d(s_i, t_j) = |s - t|$ sur toute rivière ;
- $d(s_i, t_j) = (s_i - u_{ij})\ (t_j - u_{ij})$ entre points non reliés.

La longueur de la rivière J coïncide avec l'abscisse curviligne u_J de sa source. En pratique, toute confluence peut être numérotée par son arête aval.

Dans la suite, la FA Y_J représente le débit ou la concentration du « filet » F_J, dont le support est un segment de même longueur u_J que la rivière J associée. Les « filets » sont en bijection avec les rivières.

La FA Z représentant le débit ou la concentration du réseau hydrographique est définie sur l'arbre indicé par ses arêtes. La covariance de Z est écrite comme une fonction de $s - t$ lorsqu'elle ne dépend que de la distance curviligne entre les points, ou comme une fonction de s_i et t_j, lorsqu'elle dépend aussi des arêtes.

3.3. Combinaison de FA stationnaires indépendantes

Pour un arbre à N sources, soient N FA ou « composantes » Y_J, $1 \leqslant J \leqslant N$, centrées et de même covariance $C_1(h)$ autorisée en dimension 1.

En toute confluence, attribuons à chaque arête amont k un poids w_k (Fig. 2(a)). Dans la combinaison linéaire (1), le coefficient des composantes Y_J des filets passant par l'arête i est égal au produit des poids des arêtes situées strictement à l'amont de i, depuis la source u_J. Sur l'arbre, la FA Z est définie par :

$$Z(s_i) = \sum_{J \in V_i} \left(\prod_{k \in B_{iJ}} w_k \right) Y_J(s) \tag{1}$$

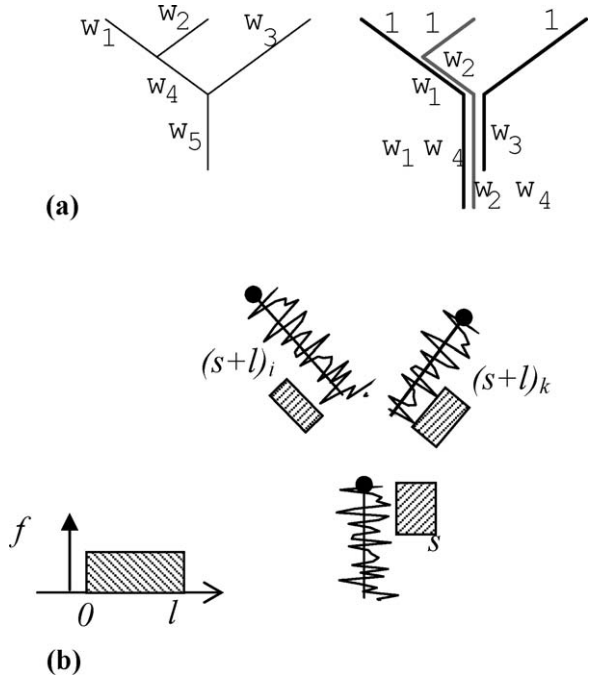


Fig. 2. Combinaison par filets (**a**) ou par arête (**b**). (**a**) À gauche, les poids attribués aux arêtes, et à droite, la pondération résultante sur les filets. (**b**) Moyennes mobiles en une confluence, modèle Ver Hoef. En grisé, le noyau f.

Fig. 2. Combining streams (**a**) or segments (**b**). (**a**) On the left, weights assigned to the segments and on the right, resulting coefficients along the streams. (**b**) Moving average at a fork, Ver Hoef model.

Les sources sont traitées comme des confluences à une seule arête amont, de coefficient unité. Sur l'arête $i = I$ à l'aval immédiat d'une source, $Z(s_i) = Y_I(s)$.

Supposons d'abord les composantes Y_J mutuellement indépendantes.

Dans la combinaison (1), des points non reliés n'ont aucune composante commune. Les Y_J étant spatialement indépendantes, la covariance de Z entre ces points est nulle.

Soient deux points reliés, situés sur des arêtes différentes, i étant à l'amont de j. Seules les composantes des filets communs, ceux passant par l'arête i, ont une contribution non nulle dans la covariance. La covariance de Z entre ces points s'écrit : $s > t$ et $V_i \subset V_j$

$$C_Z(s_i, t_j) = C_1(s-t) \sum_{J \in V_i} \Big(\prod_{k \in B_{iJ}} w_k \Big) \Big(\prod_{\ell \in B_{jJ}} w_\ell \Big) \quad (2)$$

Le long de toute rivière, la covariance de Z est proportionnelle à $C_1(s-t)$, le facteur de proportionnalité variant suivant les arêtes i et j. Tous les coefficients des arêtes situées entre les sources et chacun des points interviennent dans (2).

En deux points d'une même arête i, Z est combinaison linéaire des mêmes composantes Y_J, $J \in V_i$, et sa covariance s'écrit :

$$C_Z(s_i, t_i) = C_1(s-t) \sum_{J \in V_i} \prod_{k \in B_{iJ}} w_k^2 \quad (3)$$

En particulier,

$$Var\, Z(s_i) = C_1(0) \sum_{J \in V_i} \prod_{k \in B_{iJ}} w_k^2$$

Par arête, la variance de Z est constante et sa covariance est stationnaire. Aux confluences, Z est discontinue en moyenne quadratique, ou ce qui est équivalent, sa covariance est discontinue. Z est non stationnaire par rivière et donc sur l'arbre : sa variance est généralement modifiée à chaque confluence. À même distance curviligne, la covariance dépend des sources communes aux deux points et des confluences intermédiaires, ainsi que des poids affectés aux arêtes.

Ce modèle, construit des sources vers l'exutoire, permet de décrire les concentrations le long d'un réseau hydrographique. Lorsqu'ils sont connus (calculés par exemple à partir de l'aire du bassin versant drainé), les débits relatifs interviennent via les poids aux confluences.

Pour des débits, qui se somment aux confluences, la condition de conservation de masse revient à attribuer un coefficient unité à chaque arête. La FA Z, construite par sommation des composantes indépendantes Y_J s'écrit alors :

$$Z(s_i) = \sum_{J \in V_i} Y_J(s) \quad (4)$$

La variance $C_1(0)$ étant supposée identique pour tous les filets, la variance de Z est proportionnelle au nombre de filets en un point : constante par arête, elle croît des sources vers l'exutoire.

Supposons qu'en toute confluence d'un nombre n quelconque d'arêtes, la somme des carrés des poids affectés à ces n arêtes amont soit égale à 1 :

en toute confluence à n arêtes amont,

$$\sum_{j=1}^{n} w_j^2 = 1 \quad (5)$$

Les démonstrations données dans [12] restent valides pour une covariance C_1 quelconque. Par récurrence sur les confluences successives depuis les sources, on montre que, sur toute arête, la covariance (3) est égale à $C_1(s-t)$. La variance de Z, égale à $C_1(0)$, est alors constante sur l'arbre. Pour deux points reliés s_i

et t_j séparés par au moins une confluence, seuls interviennent dans la covariance les poids aux confluences situées entre s_i et t_j :

$$C_Z(s_i, t_j) = C_1(s - t) \prod_{k \in B_{ij}} w_k \tag{6}$$

Le modèle Ver Hoef [12] correspond au cas particulier où la covariance C_1 est l'autoconvoluée d'un noyau f défini sur la demi-droite ; les composantes Y_J sont construites par convolution d'une mesure aléatoire orthogonale,[1] par le noyau f. L'équivalence de ce modèle avec la combinaison de composantes indépendantes (1) est donnée en Annexe 2. Dans la combinaison par « filets », la covariance C_1 est quelconque ; dans le cas d'une convolution, le noyau f est par exemple symétrique.

Plus généralement, introduisons une fonction de pondération $a_J(s)$ par filet. La combinaison linéaire

$$Z(s_i) = \sum_{J \in V_i} a_J(s_i) Y_J(s) \tag{7}$$

définit une FA de variance et de covariance généralement non stationnaires, données respectivement par :

$$Var\, Z(s_i) = C_1(0) \sum_{J \in V_i} \big(a_J(s_i)\big)^2 \quad \text{et}$$

$$C_Z(s_i, t_j) = C_1(s - t) \sum_{K \in V_i \cap V_j} a_K(s_i) a_K(t_j) \tag{8}$$

Le modèle initial correspond à une fonction constante par arête, pour laquelle $a_J(s_i) = \prod_{k \in B_{iJ}} w_k$. Pour des débits $a_J(s)$ est constante et égale à 1 le long de chaque rivière (relations (4) et (7)). Lorsque la pondération est constante par arête, la covariance de Z est stationnaire par arête.

Ce modèle général (7), (8) s'applique également, lorsque le support arborescent est discrétisé.

Remarques :

- (1) pour construire une FA Z d'espérance m constante, on somme m à la combinaison des composantes Y_J centrées ;
- (2) une concentration ou un débit étant des variables positives, les composantes Y_J sont par exemple les transformées par une anamorphose positive de FA de loi spatiale gaussienne [3] ;
- (3) d'autres modèles de FA sur un arbre sont obtenus en modifiant l'opérateur agissant sur les Y_J. La combinaison linéaire (7) peut être remplacée par une moyenne d'ordre quelconque, par le produit, le maximum ou le minimum ;
- (4) les modèles précédents s'étendent à d'autres graphes que les arbres, par combinaison de composantes définies sur tous les chemins reliant une « origine » à une « extrémité » du graphe.

[1] Une mesure aléatoire orthogonale ou « bruit blanc » W est telle que pour tous v, v' « mesurables », $E[(\int_v W(\mathrm{d}x))^2] = \alpha\, Mes\, v$ et si $Mes(v \cap v')$, alors $E[\int_v W(\mathrm{d}x) \int_{v'} W(\mathrm{d}t)] = 0$.

3.4. Combinaison de FA corrélées

Dans un bassin versant, les concentrations dépendent du milieu : nature des sols, type d'agriculture... Les concentrations des affluents drainant des milieux analogues sont alors corrélées. Les modèles multivariables de type dérive externe ou à résidus permettent d'introduire le contexte environnemental dans l'estimation [3,4]. Cependant, l'information correspondante n'étant pas toujours disponible ou n'expliquant pas systématiquement les liaisons observées, il est utile de disposer de modèles tels, que les valeurs entre affluents à l'amont des confluences soient corrélées.

Il suffit pour cela que les composantes Y_J soient spatialement corrélées. Tous les modèles de corégionalisation admissibles à une dimension sont utilisables dans la construction suivante.

La longueur d'une rivière étant variable et égale à l'abscisse curviligne u_I de sa source, plusieurs « calages » sont possibles. Considérons, en dimension un, N FA X_I, de covariances simples et croisées $C_{IJ}(s, t)$. Nous examinons deux cas :

- $Y_I(s) = X_I(s)$, les FA Y_I étant « calées » depuis un exutoire commun. Les Y_I ont alors mêmes covariances croisées que les X_I.
- $Y_I(s) = X_I(u_I - s)$, pour un calage des Y_I depuis les sources. Ce calage induit une *corrélation croisée différée* de $|u_I - u_J|$ entre les composantes définies sur des rivières de longueurs différentes. Par exemple, avec un modèle multivariable stationnaire et symétrique $C_{IJ}(h) = C_{IJ}(-h)$, comme le classique modèle linéaire de corégionalisation, on obtient :

$$Cov\big(Y_I(s), Y_J(t)\big) = C_{IJ}(u_I - u_J + t - s)$$

Lorsque les composantes Y_I sont corrélées, la FA Z définie par (7) admet la covariance :

$$C_Z(s_i, t_{i'}) = \sum_{J \in V_i} \sum_{J' \in V_{i'}} a_J(s_i) a_{J'}(t_{i'})\, Cov\big(Y_J(s), Y_{J'}(t)\big)$$

Toutes les covariances simples et croisées des composantes Y_J associées aux filets passant par s_i ou $t_{i'}$ interviennent dans la covariance de Z.

3.5. Combinaison de FA intrinsèques

En dimension 1, soit R une FA stationnaire d'espérance m et de covariance $k(h)$. L'intégrale $S(x) = \int_0^x R(t)\,\mathrm{d}t$ est une FA intrinsèque, de dérive linéaire de pente m et de variogramme $\gamma(h) = \int_0^h (h-t)K(t)\,\mathrm{d}t$ [8]. La variance de S, définie en tout point, est non stationnaire :

$$Var\,S(x) = \int_0^x \int_0^x K(t-t')\,\mathrm{d}t\,\mathrm{d}t'$$

Dans le cas limite où R est pépitique (c'est-à-dire est une mesure aléatoire orthogonale), le variogramme de S est linéaire (voir par exemple [3]).

Lorsque R est positive, la FA S, positive et croissante, est un modèle admissible pour décrire un débit à accroissements stationnaires par unité de longueur de la rivière. $Y_J\ (u_J)$ représentant le débit à la source, le débit du J^{e} filet est $Y_J(s) = Y_J(u_J) + \int_s^{u_J} R_J(t)\,\mathrm{d}t$ et le débit sur le réseau est défini par (7). À l'aval d'une confluence, la somme des RJ, $J \in V_i$, sur les filets passant par s_i représente l'accroissement élémentaire de débit, croissant avec l'ordre hydrologique de la rivière.

Plus généralement, soient N FA intrinsèques Y_J, indépendantes (pour simplifier), supposées ici de variogramme quelconque $\gamma_J(h)$ associé à la covariance non stationnaire $C_J(s, s+h)$. La FA Z construite suivant (4) en sommant ces composantes à l'aval des confluences est intrinsèque par arête et de variogramme

$$\gamma_Z\big(s_i, (s+h)_i\big) = \sum_{K \in V_i} \gamma_K(h)$$

somme des variogrammes pour les « filets » passant par l'arête. La covariance associée est la somme des covariances non stationnaires de chacun des filets :

$$C_Z\big(s_i, (s+h)_i\big) = \sum_{J \in V_i} C_J(s, s+h)$$

La variance de Z est finie et dépend du point s_i.

Entre deux points reliés situés de part et d'autre d'une confluence, la covariance non stationnaire s'écrit comme une somme portant sur les seuls filets communs aux deux points, c'est-à-dire ceux passant par le point le plus en amont $(s+h)_j$:

$$h > 0, \quad C_Z\big(s_i, (s+h)_j\big) = \sum_{K \in V_j} C_K(s, s+h)$$

Deux points non reliés n'ont aucune composante commune. La covariance de Z en ces points est nulle et le variogramme est alors la moyenne des variances :

$$\begin{aligned}\gamma_Z(s_i, t_j) &= \frac{1}{2} Var\big(Z(s_i) - Z(t_j)\big) \\ &= \frac{1}{2} Var\,Z(s_i) + \frac{1}{2} Var\,Z(t_j)\end{aligned}$$

Ce variogramme dépend séparément de s_i et t_j.

La FA Z présente des discontinuités aux confluences. Ce modèle se généralise aux combinaisons linéaires de composantes intrinsèques, ainsi qu'au cas où les composantes Y_I sont spatialement corrélées.

4. FA stationnaires ou intrinsèques par rivière

Le modèle Bailly–Monestiez [2,11] est construit en sens inverse, de « l'exutoire » vers les « sources ». Évitant les discontinuités aux confluences, ce modèle peut décrire des phénomènes variés, comme des caractéristiques végétales. Ces auteurs l'utilisent pour modéliser la largeur du fluvisol ou celle des fossés d'un réseau de drainage. En inversant l'orientation de l'arbre, on peut aussi l'appliquer aux concentrations dans un delta. Enfin, lorsque la dérive prend en charge les discontinuités aux confluences, ce modèle s'applique aux résidus [2,11].

4.1. FA stationnaire par rivière

Dans ce modèle, la covariance est stationnaire par rivière et identique pour toutes les rivières. Pour deux rivières quelconques, les composantes Y_I coïncident de leur confluence à l'exutoire, et elles évoluent indépendamment de la confluence vers les sources (Fig. 3). Tandis que dans le modèle Ver Hoef, les *valeurs* de Z sur les différentes rivières sont *indépendantes*, dans le modèle Bailly–Monestiez cette indépendance est *conditionnelle* aux valeurs de Z sur le chemin commun aux rivières.

Pour simplifier, et puisqu'il en est ainsi en pratique, le graphe est désormais discrétisé. On se ramène à un arbre un peu particulier, dont la majorité des sommets comporte une seule arête « amont » et une seule arête « aval ».

Le modèle Bailly–Monestiez est le suivant. À une anamorphose près, Z est supposée de loi spatiale gaussienne. La numérotation étant arbitraire, $Z(s_1) = Y_1(s)$ de covariance $C_1(h)$, est d'abord construite sur la « première » rivière F_1, par exemple la plus longue. Z étant supposée construite sur les $J-1$ premières rivières, soit x_J la confluence d'abscisse maximum raccordant F_J aux rivières précédemment construites :

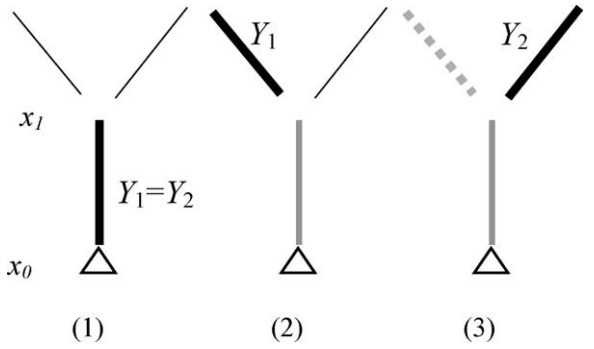


Fig. 3. Indépendance conditionnelle de Y_1 et Y_2 entre la confluence u_{12} et les sources. Trait noir épais, le segment en construction ; trait gris épais, les « données » utilisées pour le conditionnement ; tireté : poids de krigeage nuls.

Fig. 3. Conditional independence between Y_1 and Y_2 from the fork u_{12} to the sources. Black bold: segment in construction, grey bold: data used for conditioning, dotted line: null kriging weights.

$x_J = \max\{u_{JK}, K < J\}$. Entre l'exutoire $s = 0$ et x_J, on pose $Y_J(s) = Z(s)$; entre la confluence x_J et la source u_J, Y_J de covariance a priori C_1 est construite conditionnellement aux $Y_J(s)$, $0 \leqslant s \leqslant x_J$. Ceci est possible par simulation séquentielle, ou par toute méthode de simulation non conditionnelle à 1D, en conditionnant ensuite par les valeurs entre l'exutoire et x_J [3]. En tout point, on pose :

$$Z(s_J) = Y_J(s) \tag{9}$$

La covariance de Z entre rivières se calcule à l'aide de la loi conditionnelle dans le cas gaussien : l'espérance conditionnelle coïncide alors avec le krigeage à moyenne connue et la variance résiduelle est égale à la variance de krigeage. Pour deux rivières distinctes d'indices I, J de confluence u_{IJ}, posons $Y_I(s) = Y_I^K(s) + R_I(s)$ et $Y_J(t) = Y_J^K(t) + R_J(t)$, $Y_I^K(s)$ et $Y_J^K(t)$ désignant le krigeage à moyenne connue de $Y_I(s)$ et $Y_J(t)$ *par les seules valeurs communes comprises entre l'exutoire et la confluence*, désignées désormais comme « données ». Les résidus $R_I(s)$ et $R_J(t)$ du krigeage à moyenne connue sont sans corrélation avec les « données » $Y_I(s) = Y_J(s)$, $0 \leqslant s \leqslant u_{IJ}$. Dans le modèle Bailly–Monestiez, ces résidus sont, de plus, spatialement indépendants pour $I \neq J$, et la covariance entre $Z(s_I)$ et $Z(t_J)$ s'écrit

$$E\big[Z(s_I)Z(t_J)\big] = E\big[Y_I^K(s)Y_J^K(t)\big] = \sum_{\alpha,\beta} \lambda_s^\alpha C_1(s_\alpha - s_\beta)\lambda_t^\beta \tag{10}$$

λ_s^α désignant le poids de s_α dans le krigeage à moyenne connue, au point d'abscisse s. Cette covariance admet une expression matricielle synthétique : K désignant la matrice de covariance $C_1(s_\alpha - s_\beta)$ des « données » sur le tronçon commun $[0, u_{IJ}]$ et K_s (respectivement K_t) la matrice des covariances $C_1(s_\alpha - s)$ (resp. $C_1(s\beta - t)$) entre « données » et $Y_I(s)$, $u_I < s \leqslant u_{IJ}$ (resp. $Y_J(t)$, $u_{IJ} < t \leqslant u_J$), un calcul simple montre que :

$$E\big[Y_I^K(s)Y_J^K(t)\big] = {}^{\mathrm{t}}K_s K^{-1} K_t$$

La covariance de Z, généralement non stationnaire entre points non reliés, dépend de la distance de s_I et t_J à la confluence u_{IJ}.

Le modèle Bailly–Monestiez comporte la covariance exponentielle comme cas particulier stationnaire sur tout l'arbre. Dans le cas de simulations non conditionnelles, le krigeage à moyenne connue, effectué suivant les abscisses curvilignes croissantes, ne fait plus intervenir que le point aval précédemment construit le plus proche.

Remarque : en l'absence de discrétisation de l'arbre, il convient de considérer le krigeage sur un support continu. Sous réserve de l'existence d'une mesure appropriée (qui n'existe pas nécessairement pour des covariances très régulières, telles que l'exponentielle de Gauss [9]), ce krigeage s'écrit : $s > x_I$, $Y_I^K(s) = \int_0^{x_I} Y_I(t)\lambda_I(\mathrm{d}t)$. La covariance entre rivières distinctes admet alors une expression intégrale, généralisation immédiate des sommes finies (10) :

$$E\big[Y_I^K(s)Y_J^K(t)\big] = \int_0^{u_{IJ}} \int_0^{u_{IJ}} \lambda_I(\mathrm{d}s)\, C_1(s-t)\lambda_J(\mathrm{d}t)$$

Dans la suite, sauf mention contraire, le support arborescent est supposé discrétisé.

4.2. FA intrinsèque par rivière

La construction précédente s'étend aux FA intrinsèques, en posant de façon analogue une hypothèse d'indépendance des résidus du krigeage *intrinsèque* sur les rivières distinctes.

Soit toujours Z défini par (9), les Y_I étant ici des FA intrinsèques en dimension un, de variogramme $\gamma_1(h)$ quelconque. Entre deux points reliés, $\gamma_Z(s_I, t_I) = \gamma_1(s - t)$. Le variogramme de Z est stationnaire le long de toute rivière.

Soient s_I et t_J non reliés. Notons $Y_I^*(s)$ le krigeage intrinsèque de $Y_I(s)$ par les « données » $Y_I(s) = Y_J(s)$, $0 \leqslant s \leqslant u_{IJ}$, et de même pour J. Alors :

$$\gamma_Z(s_I, t_J) = \frac{1}{2} Var\big(Y_I(s) - Y_J(t)\big)$$

avec :

$$Y_I(s) - Y_J(t) = Y_I(s) - Y_I^*(s) + Y_I^*(s) - Y_J^*(t) + Y_J^*(t) - Y_J(t)$$

Dans le krigeage intrinsèque, les résidus $Y_I(s) - Y_I^*(s)$, $Y_J(t) - Y_J^*(t)$ sont non corrélés aux combinaisons linéaires autorisées des « données », donc en particulier à $Y_I^*(s) - Y_J^*(t)$. Les résidus sur des rivières distinctes étant supposés indépendants, la variance de la somme est la somme des variances :

$$\gamma_Z(s_I, t_J) = \frac{1}{2} Var\big(Y_I(s) - Y_I^*(s)\big) + \frac{1}{2} Var\big(Y_I^*(s) - Y_J^*(t)\big) + \frac{1}{2} Var\big(Y_J(t) - Y_J^*(t)\big)$$

Le premier et le dernier terme correspondent aux variances de krigeage $\sigma^2(s_I)$, $\sigma^2(t_J)$ de $Y_I(s)$ ou $Y_J(t)$ par les $Y_I(s) = Y_J(s)$, $0 \leqslant s \leqslant u_{IJ}$; ils dépendent de la position relative de s_I et t_J par rapport à ces points, en particulier par rapport à la confluence u_{IJ}. Le calcul du troisième terme est classique : notant λ_s^α (resp. λ_t^α) le poids de s_α dans le krigeage intrinsèque de $Y_I(s)$ (resp. $Y_J(t)$),

$$\begin{aligned} &Var\big(Y_I^*(s) - Y_J^*(t)\big) \\ &= Var\left(\sum_\alpha \big(\lambda_s^\alpha - \lambda_t^\alpha\big) Y_I(s_\alpha)\right) \\ &= -\sum_{\alpha,\beta} \big(\lambda_s^\alpha - \lambda_t^\alpha\big)\gamma_1(s_\alpha - s_\beta)\big(\lambda_s^\beta - \lambda_t^\beta\big) \end{aligned}$$

Via les poids de krigeage, ce terme dépend, là encore, de la position de s_I et t_J par rapport à la confluence. Le variogramme de Z, généralement non stationnaire entre points non reliés, s'écrit finalement :

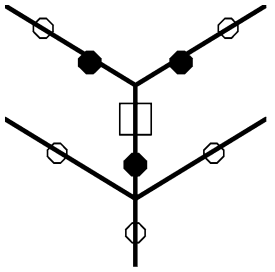
$$\gamma_Z(s_I, t_J) = \frac{1}{2}\sigma^2(s_I) + \frac{1}{2}\sigma^2(t_J) - \sum_{\alpha<\beta} \big(\lambda_s^\alpha - \lambda_t^\alpha\big)\gamma_1(s_\alpha - s_\beta)\big(\lambda_s^\beta - \lambda_t^\beta\big)$$

qui admet également une expression matricielle.

Ce résultat s'étend au cas où le support n'est pas discrétisé, sous réserve de l'existence d'une mesure appropriée pour l'écriture du krigeage.

Considérons le cas particulier d'un processus de Wiener–Lévy (ou mouvement brownien). W désignant une mesure aléatoire orthogonale, l'intégrale $Y(s) = \int_0^s W\,\mathrm{d}t$ est intrinsèque et de variogramme linéaire (cf. partie 3.5).

On vérifie sans difficulté que « l'effet d'écran » du variogramme linéaire en dimension un se retrouve sur tout support arborescent, discrétisé ou non : quelle que soit la configuration de krigeage, *le long de tout chemin passant par le point à estimer*, seules les « données » les plus proches de part et d'autre de ce point admettent un poids non nul (Fig. 4).



□ point cible , ○ poids nuls, ● poids non nuls

Fig. 4. Effet d'écran pour le variogramme linéaire.
Fig. 4. Screen effect of the linear variogram for intrinsic kriging.

Sur F_1, construisons alors Z de variogramme linéaire. Z étant ensuite supposé de variogramme linéaire sur les $I - 1 \geqslant 1$ premières rivières, soit $x_1 = \max\{u_{IJ}, J < I\}$ la confluence d'abscisse maximum entre F_1 et ces rivières. Raccordant Z par continuité en cette confluence, on pose : $x_1 < s \leqslant u_I$, $Z(s_I) = Z_I(x_I) + \int_{x_I}^s W_I\,\mathrm{d}t$. Ceci revient à conditionner par l'ensemble des valeurs précédemment construites sur F_I.

Les *accroissements* de Z sur F_I à l'amont de x_1 étant choisis indépendants de ceux sur les $I - 1$ premières rivières, la variance de la somme de deux accroissements, de part et d'autre de la confluence, est égale à la somme des variances, et le variogramme de Z reste linéaire entre points non reliés. Sur l'arbre, Z est intrinsèque et de variogramme linéaire.

4.3. Indépendance conditionnelle et stationnarité

La covariance exponentielle (voir par exemple [11]), et le variogramme linéaire sont admissibles sur tout support arborescent, discrétisé ou non. Pour ces deux schémas, la construction d'une FA de loi spatiale de type mosaïque est donnée en annexe, sans discrétisation du support.

Dans le cas discret, montrons que ces deux modèles sont les seuls *compatibles avec l'hypothèse d'indépendance conditionnelle et stationnaire sur l'arbre* (i.e. ne dépendant que de la distance curviligne).

Dans la suite, l'écriture en covariance (et pour une variance unité) correspond au krigeage à moyenne connue, et celle en variogramme au krigeage intrinsèque. Considérons le krigeage en x^2 à partir de x_0 et x_1 séparés par la confluence u_{12} (Fig. 5), $h_i = |u_{12} - x_i|$ désignant la distance de x_i, $0 \leqslant i \leqslant 2$, à la confluence. Pour $(i, j) = (0, 1)$ ou $(1, 0)$ respectivement, les poids

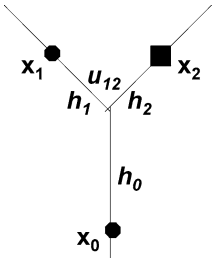


Fig. 5. Configuration de krigeage.
Fig. 5. Kriging configuration on the node.

de krigeage s'écrivent selon le cas

$$\lambda_i = \frac{C(h_2+h_i) - C(h_2+h_j)C(h_i+h_j)}{1-C^2(h_0+h_1)} \quad \text{et}$$

$$\lambda_i = \frac{1}{2} + \frac{\gamma(h_2+h_j) - \gamma(h_2+h_i)}{2\gamma(h_i+h_j)}$$

D'après la condition d'indépendance conditionnelle, $\lambda_1 = 0$, et par suite :

$$C_Z(h_0+h_1)C_Z(h_0+h_2) = C_Z(h_1+h_2) \quad \text{et}$$

$$\gamma_Z(h_0+h_1) + \gamma_Z(h_0+h_2) = \gamma_Z(h_1+h_2)$$

Lorsque x_0 est situé sur la confluence, $h_0 = 0$ et on retrouve l'effet d'écran en dimension 1. Les relations nécessaires $C_Z(h_1)C_Z(h_2) = C_Z(h_1+h_2)$ ou $\gamma_Z(h_1) + \gamma_Z(h_2) = \gamma_Z(h_1+h_2)$, quels que soient h_1 et h_2 bornés, admettent comme solutions la covariance exponentielle et le variogramme linéaire. Lorsque $h_0 > 0$, il n'y a pas d'effet d'écran.

5. Conclusion

Les quelques modèles présentés s'adaptent aisément à une modélisation multivariable. Pour représenter les concentrations en différentes substances, par exemple des nutriments (nitrates, phosphates...), on pourra chercher un ajustement en modèle linéaire de corégionalisation. Il est immédiat de vérifier que sur tout arbre, ce modèle s'obtient classiquement par combinaison de facteurs spatiaux mutuellement indépendants [3].

Dans les applications pratiques, une modélisation *bivariable débits–concentrations* reste à développer, pour les cas où des mesures de débits sont disponibles en certaines stations. Cette modélisation devra tenir compte des valeurs approchées de ces débits, déduites par exemple de la superficie des bassins versants.

L'application aux données réelles pose l'importante question de l'inférence des modèles. Dans les exemples cités, il apparaît une « dérive » marquée des sources vers l'exutoire. Une modélisation de type « dérive + résidu » est alors recherchée pour incorporer au modèle diverses informations sur le milieu et se ramener à une variable stationnaire. Modélisant la largeur de fossés de drainage ou de fluvisols, Bailly et al. et Monestiez et al. calent des dérives qui sont notamment fonction de la longueur cumulée des drains à l'amont d'un point, et ajustent des variogrammes stationnaires pour les résidus.

Ver Hoef et al. montrent que les variogrammes expérimentaux calculés le long des rivières sont structurés et nettement inférieurs à ceux, pépitiques, calculés sur tout l'arbre. Ceci est compatible avec l'hypothèse d'indépendance entre rivières à l'amont des confluences. L'étude des variogrammes expérimentaux sur le réseau devrait ainsi guider la modélisation.

Remerciement

Ce travail a été effectué grâce à la subvention CV02000187 du ministère français en charge de l'Environnement. Les auteurs remercient H. Beucher, J.-P. Chilès, D. Renard et J.-P. Vert pour leur relecture attentive ou leur aide graphique ou linguistique.

Annexe 1. Variogramme linéaire et covariance exponentielle sur un arbre

Nous donnons une méthode de construction, sur tout graphe sans cycle, d'une FA de variogramme linéaire ou de covariance exponentielle, fondée sur le processus de Poisson en dimension 1.

Soient T_n les abscisses des points d'un processus de Poisson de densité θ sur la droite, et A_n une suite de variables aléatoires identiques en loi, d'espérance m et de variance σ^2, mutuellement indépendantes et indépendantes du processus. La FA Y, définie à une constante près comme

- constante entre deux points poissoniens ;
- présentant un saut d'amplitude A_n en T_n,

est intrinsèque, de dérive θmh et de variogramme (défini comme la demi-variance des accroissements) linéaire $\frac{1}{2}\theta(m^2+\sigma^2)|h|$. Si l'on introduit un point d'abscisse x_0 fixée, les propriétés du processus avant et après x_0 restent inchangées.

Il en résulte un procédé de construction d'une FA de variogramme linéaire sur un arbre. On construit un processus de Poisson composé sur tous les chemins du graphe, et on « raccorde » par continuité la FA sur toute nouvelle arête aux chemins précédemment simulés (Fig. 6). En une confluence, la FA admet la même valeur sur toutes les arêtes.

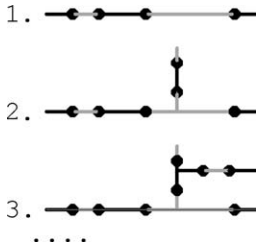


Fig. 6. FA de covariance exponentielle ou de variogramme linéaire, construites sur un processus ponctuel de Poisson. Les valeurs sur les segments contigus sont constantes.

Fig. 6. RF with exponential covariance or linear variogram on a tree, built on a Poisson point process. Values are constant on contiguous segments.

Plus précisément, la méthode est la suivante : (1) construire sur les arêtes, des processus de Poisson indépendants, de densité θ ; (2) poser $Y(x) = A_0$ en une arête, par exemple celle issue de la racine ; entre deux points du processus, $Y(x)$ est constante, et présente un saut d'amplitude A_n aux points de discontinuité T_n.

Pour construire sur la droite une FA Y de covariance exponentielle, on implante sur chaque segment, compris entre deux points du processus (T_n, T_{n+1}), une valeur constante $Y(x) = A_i$, la FA conservant la même valeur sur toutes les arêtes en une confluence. Comme pour le variogramme linéaire, ce procédé est admissible pour tout graphe sans cycle.

D'après le théorème central limite, la sommation d'un grand nombre de FA indépendantes ainsi construites fournit, au facteur usuel de normation près, une FA de loi spatiale (ou d'incréments) multigaussiens.

Annexe 2. Équivalence de la combinaison de composantes indépendantes et du modèle Ver Hoef

Le modèle Ver Hoef utilise la construction classique par moyennes mobiles. Soit W un « bruit blanc » de variance unité et f une fonction de carré sommable ou « noyau ». En dimension un, la FA définie par

$$Z(s) = \int f(t-s)W\,\mathrm{d}t \tag{11}$$

admet la covariance stationnaire

$$C_1(h) = \int f(t)f(t-h)\,\mathrm{d}t$$

produit de convolution de f par son symétrisé :

$$C_1(h) = f * \check{f}, \quad \text{où } \check{f}(t) = f(-t)$$

Dans ce modèle, l'indépendance de Z sur les arêtes à l'amont d'une confluence est obtenue en utilisant un noyau défini sur la demi-droite $\mathbb{R}^+$. Dans la relation (11), la moyenne mobile est calculée par pondération de W à l'amont du point courant, la borne supérieure étant conventionnellement mise à $+\infty$, que le support de f soit ou non borné :

$$Z(s_i) = \int_{s_i}^{\infty} f(t-s_i)W\,\mathrm{d}t \tag{11bis}$$

Que devient la moyenne mobile, lorsque t atteint une confluence à l'amont de s_i ? Afin d'obtenir une variance stationnaire, Ver Hoef et al. répartissent le noyau f sur les n arêtes à l'amont de la confluence (Fig. 2b), en les pondérant par des poids w_j dont la somme des carrés est égale à 1 (relation (5)).

La construction est la suivante : (1) construire un mesure aléatoire orthogonale W en tout point du graphe, indicé par son arête ; (2) effectuer la moyenne mobile en la répartissant sur les arêtes amont, l'intégrale étant conventionnellement étendue à l'infini au-delà des sources :

$$Z(s_i) = \int_{s_i}^{u_i} f(t-s_i)W_i\,\mathrm{d}t + \sum_{j\in U_i}\left(\prod_{k\in B_{ij}} w_k\right)\int_{l_j}^{u_j} f(t-s_i)W_j\,\mathrm{d}t$$

La covariance C de la FA Z ainsi construite est :

- nulle entre rivières différentes :
- $C(s_i, t_j) = C_1(s-t)\prod_{k\in B_{ij}} w_k$ sur deux arêtes reliées, le deuxième facteur dépendant des confluences entre les arêtes ;
- $C(s_i, t_i) = C_1(s-t)$ sur toute arête.

La covariance C est stationnaire entre arêtes, mais non stationnaire sur l'arbre.

Pour montrer l'équivalence des deux modèles, considérons pour simplifier un arbre comportant n arêtes amont convergeant toutes en une seule confluence d'abscisse u. À l'aval de la confluence, le modèle Ver Hoef s'écrit : $0 \leqslant s \leqslant u$.

$$Z(s) = \int_{s}^{u} f(t-s)W\,\mathrm{d}t + \sum_{j=1}^{n} w_j \int_{u}^{u_j} f(t-s)W_j\,\mathrm{d}t \tag{12}$$

Soient alors n FA indépendantes Y_j, définies respectivement sur chaque *rivière*, de la source à l'exutoire, par $Y_j(s) = \int_s^{u_j} f(t-s)W_j\,\mathrm{d}t$. Pour s à l'aval de la confluence, posons :

$$Y(s) = \sum_{j=1}^{n} w_j Y_j(s) \tag{13}$$

soit, en introduisant le point de confluence :

$$Y(s) = \sum_{j=1}^{n} w_j \int_s^u f(t-s)W_j\,\mathrm{d}t + \sum_{j=1}^{n} w_j \int_u^{u_j} f(t-s)W_j\,\mathrm{d}t$$

Par linéarité de la convolution, le premier terme du second membre s'écrit :

$$\int_s^u f(t-s)\left(\sum_{j=1}^{n} w_j W_j\,\mathrm{d}t\right)$$

Les n bruits blancs W_j étant mutuellement indépendants et de variance unité :

$$Var\left(\sum_{j=1}^{n} w_j W_j\,(\mathrm{d}t)\right) = \left(\sum_{j=1}^{n} w_j^2\right) Var\,W\,\mathrm{d}t = 1$$

Les FA Y et Z admettent donc même espérance et même covariance. Lorsque les « bruits blancs » ne sont pas gaussiens, l'histogramme de Z ou de Y n'est plus nécessairement stationnaire, mais dépend de la loi des W_j, la covariance C restant inchangée.

L'équivalence entre (12) et (13) se généralise à un arbre quelconque, en considérant successivement toutes les confluences depuis les sources. La méthode Ver Hoef revient donc à effectuer la moyenne pondérée suivante, dans laquelle la sommation porte sur les *rivières passant par* s_i :

$$Z(s_i) = \sum_{J\in V_i}\left(\prod_{k\in B_{iJ}} w_k\right)\int_{s_i}^{u_J} f(t-s_i)W_J\,\mathrm{d}t$$

la pondération attribuée en s_i à la rivière F_J dépendant des confluences depuis sa source. L'écriture de Z comme combinaison linéaire, à coefficients variables, de FA indépendantes définies sur les rivières, permet de généraliser ce modèle à la combinaison de FA monodimensionnelles de covariance quelconque.

Références

[1] J.-M. Audergon, P. Monestiez, R. Habib, Spatial dependences and sampling in a fruit tree: A new concept for spatial prediction in fruit studies, J. Hortic. Sci. 68 (1) (1993) 99–112.

[2] J.-S. Bailly, P. Monestiez, P. Lagacherie, Exploring spatial variability along drainage networks with geostatistics, Math. Geol. 38 (5) (2006), in press.

[3] J.P. Chilès, P. Delfiner, Geostatistics: Modeling Spatial Uncertainty, Wiley, New York, 1999.

[4] N. Cressie, J.J. Majure, Spatio-temporal statistical modelling of livestock waste in streams, J. Agric. Biol. Environ. Stat. 2 (1) (1997) 24–47.

[5] N. Cressie, J. Frey, B. Harch, M. Smith, Spatial prediction on a River Network, J. Agric. Biol. Environ. Stat., in press.

[6] D. Dumas, Optimisation de la quantification des flux de matière en suspension d'une rivière alpine : l'Isère à Grenoble, C. R. Geoscience 336 (2004) 1149–1159.

[7] R. Kondor, J.-P. Vert, Diffusion kernels, in: B. Schoelkopf, K. Tsuda, J.-P. Vert (Eds.), Kernel Methods in Computational Biology, MIT Press, Cambridge, MA, USA, 2004, pp. 171–192.

[8] G. Matheron, Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux sciences de la Nature, Masson, Paris, 1965.

[9] G. Matheron, La théorie des variables régionalisées, et ses applications. Les cahiers du centre de morphologie mathématique de Fontainebleau, fasc. 5, ENSMP, Fontainebleau, France, 1970.

[10] P. Monestiez, R. Habib, J.-M. Audergon, Estimation de la covariance et du variogramme pour une fonction aléatoire à support arborescent : Application à l'étude des arbres fruitiers, in : M. Armstrong (Ed.), Geostatistics, Kluwer Academic Publishers, Dordrecht, Pays-Bas, 1989.

[11] P. Monestiez, J.-S. Bailly, P. Lagacherie, M. Voltz, Geostatistical modelling of spatial processes on directed trees: Application to fluvisol extent, Geoderma 128 (3–4) (2005) 179–191.

[12] J.M. Ver Hoef, E. Peterson, Theobald D. Spatial statistical models that use flow and stream distance. Environ. Ecol. Stat (2006), in press.