Record Number:
Author, Monographic: Cluis, D.//Laberge, C.
Author Role:
Title, Monographic: Variance et erreur d'estimation de l'interpolation linéaire dans un monde markovien
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 1985
Original Publication Date: Mai 1985
Volume Identification:
Extent of Work: 18
Packaging Method: pages
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, Rapport de recherche
Series Volume ID: 186
Location/URL:
ISBN: 2-89146-184-3
Notes: Rapport annuel 1984-1985
Abstract: 10.00$
Call Number: R000186
Keywords: rapport/ ok/ dl

Variance et erreur d'estimation de l'interpolation linéaire dans un monde markovien

par

D. Cluis et C. Laberge

Rapport scientifique no 186

mai 1985

INRS-Eau
Université du Québec
C.P. 7500, Sainte-Foy
Québec 10, G1V 4C7

Résumé

La plupart des réseaux d'acquisition de données et de surveillage générale fournissent des informations équidistantes et instantanées; la fréquence des mesures nécessaire pour l'obtention efficiente de cette information est reliée à la variabilité intrinsèque du phénomène mesuré. En conséquence, les phénomènes hydrologiques et météorologiques doivent être échantillonnés plus intensivement que les variables plus stables relatives aux eaux souterraines. Une fois les données acquises, une estimation de valeurs prises par le processus à des pas de temps intermédiaires de même que la combinaison de deux ou plusieurs séries de temps prélevées à des fréquences différentes constituent un problème fréquent, mais non résolu. Dans le domaine de l'acquisition des données de qualité des eaux courantes, par exemple, une estimation des débits massiques constitue un préalable essentiel à l'interprétation des phénomènes de transport et des relations sources-conséquences et à la détection des tendances éventuelles. Pour évaluer cette variable secondaire importante, il faut combiner les débits, variables de haute variabilité acquises à hautes fréquences avec des concentrations, variables de faible variabilité acquise à basses fréquences. Ceci peut être effectué en utilisant une combinaison d'aggrégation et d'interpolation des données. L'aggrégation de données hautes-fréquences n'a que des effets mineurs sur la structure des données: une réduction de la variance et une modification de la structure de persistance des données transformées par rapport à la série mesurée. Par contre, l'étalement de l'information résultant d'une interpolation linéaire crée une hétéroscédasticité, mais surtout introduit une erreur d'estimation dont la variance croît avec le nombre des partitions.

Nous nous intéressons ici aux phénomènes présentant une persistance Markovienne positive; dans ce cadre, nous avons développé l'expression analytique de la variance globale de l'erreur d'estimation pour la série des valeurs mesurées de p en p unités de temps, puis interpolée; ensuite nous avons transposé ce résultat en utilisant la caractéristique d'invariance de la structure des persistances, typique aux séries Markoviennes pour établir l'erreur quadratique moyenne d'une série de pas de temps unitaire

interpolée. De cette façon, la variance de l'erreur introduite dans le signal par interpolation linéaire est reliée à la persistance de la série mesurée et au nombre des subdivisions.

Mots clés: Acquisition de données / fréquence d'échantillonnage / processus Markovien / interpolation / erreur quadratique moyenne.

Introduction

Il arrive que l'on désire combiner deux ou plusieurs séries de temps géophysiques qui ont été échantillonnées systématiquement à des intervalles de temps multiples entiers ou non les uns des autres. Si on se limite à considérer seulement les évènements échantillonnés simultanément, il résulterait une perte importante de l'information existante; ceci est spécialement vrai si les données originales sont constitués d'un échantillonnage équidistant comme c'est le cas pour la plupart des réseaux d'acquisition de données à vocation de surveillance générale. La fréquence des mesures nécessaire pour faire une acquisition efficace de cette information est alors, reliée à la variabilité temporelle intrinsèque des différents phénomènes. Selon les phénomènes échantillonnés (météorologie, hydrométrie de surface, eau souterraine), ces fréquences peuvent différer de plusieurs ordres de grandeur.

Dans le domaine de la surveillance de la qualité de l'eau, par exemple, l'estimation des débits massiques ou flux de matière constitue un prérequis nécessaire à l'interprétation des phénomènes de transport, des relations sources-effets et à l'évaluation de tendances éventuelles. Pour obtenir cette variable secondaire importante, on doit combiner les données de débits (haute fréquence / haute variabilité) avec des données de concentrations (basse fréquence / faible variabilité). Ainsi dans la Province de Québec, les niveaux d'eau enregistrés tous les 15 minutes conduisent à des données de débits publiées sur une base journalière. Le réseau d'acquisition des données de qualité des eaux courantes fournit, de son côté, les caractéristiques instantanées d'échantillons prélevés toutes les 3 ou 4 semaines. L'estimation des débits massiques, c'est-à-dire des flux de polluants peut s'effectuer en combinant une interpolation et une aggrégation, mais il est important d'évaluer combien d'information fictive a été introduit dans les calculs par cette manipulation des données; cette connaissance est essentielle, par exemple, dans le cadre de la modélisation qualitative, quand ces débits massiques "mesurés" servent à la calibration. Négliger l'imprécision des valeurs de référence peut, en effet, conduire à des tentatives futiles de perfectionner un modèle alors que les données d'entrée ont été totalement exploitées.

La définition d'un pas de calcul "approprié" pour la variable combinée ainsi que l'estimation des erreurs induites constituent un problème courant, mais apparemment non résolu. En fait, l'aggrégation de données à hautes fréquences a des conséquences structurales: une réduction de variance et une modification de la persistence des données transformées, mais il n'y a pas d'introduction d'information externe à la série temporelle. Par contre, l'étalement de l'information résultant d'une interpolation linéaire crée un certain niveau hétéroscédasticité, modifie la structure de persistence, mais surtout induit une erreur d'estimation dont la variance croît avec le nombre de subdivisions, c'est de ce sujet que nous traiterons ici.

Hypothèses

Supposons d'abord que la série chronologique Z_i suive un modèle additif classique:

$$Z_i = T_i + S_i + X_i \qquad [1]$$

où T_i et S_i représentent respectivement les composantes de tendance et de saisonnalité du phénomène observé, X_i représente les fluctuations temporelles à court-terme et finalement i représente un indice de temps d'occurrence. Nous supposerons, de plus, que l'on dispose d'un ensemble de données suffisamment grand pour identifier clairement et enlever *a priori* T, la tendance à long terme, et S, les variations saisonnières.

Les développements qui suivent traitent essentiellement la composante stationnaire X.

Le modèle statistique utilisé pour cette composante est cohérent avec le comportement à court-terme de plusieurs séries de temps géophysiques: l'échantillon est constitué de variables aléatoires identiquement distribuées et autocorrelées, avec moyenne zéro et variance σ^2; sa structure d'autocorrélation étant uniquement fonction de l'intervalle de temps entre les valeurs échantillonnées. De plus, on suppose que l'autocorrélogramme décroît exponentiellement selon la relation: $r_k = r_1^k$ où r_1 est le coefficient

d'autocorrélation correspondant à un unité d'intervalle de temps. Nous porterons un intérêt spécial au cas pratique d'une dépendance fortement positive ($r_1 > 0$).

Comme nous traiterons ici des propriétés intrinsèques et, pour éviter l'utilisation de notations multiples, nous n'essayerons pas de distinguer formellement les statistiques du processus, de leurs estimés tirés de grands échantillons.

Interpolation linéaire

L'interpolation linéaire itérative (ILI) est une technique d'estimation très souvent utilisée lorsque l'on désire générer des valeurs à une fréquence plus grande que celle des variables hydrologiques mesurées. Il est évident qu'aucune nouvelle information n'est créée par l'interpolation linéaire, le contenu original n'étant que réparti dans le temps, on ne considère simultanément que deux valeurs situées aux extrémités de chaque intervalle de temps étudié. L'interpolation linéaire itérative ne peut pas être considérée comme un estimateur optimal; cependant, selon la variabilité du processus, l'ILI constitue une méthode de génération de valeurs intermédiaires plus efficace que l'utilisation d'une espérance mathématique générale ou saisonnière; mais il existe des méthodes d'approximation numérique plus raffinées comme l'interpolation à l'aide de polynômes orthogonaux ou l'utilisation de fonctions "spline" pour obtenir des estimateurs plus puissants prenant avantage des tendances temporelles à court terme des valeurs mesurées. Néanmoins, à cause de sa simplicité, l'ILI demeure une des méthodes d'estimation, des plus usuelles et des plus pratiques, utilisée en hydrologie pour générer des valeurs à une fréquence plus élevée que celle de la série observée.

Soit X_i ($i = 1 \ldots k$) une série de mesures de longueur k. Chaque intervalle entre deux valeurs consécutives de X_i est divisé par p intervalles égaux pour donner la série interpolée Y_j ($j = 1 \ldots N$) de longueur $N = (k-1)\, p + 1$.

Chaque terme de Y_j est interpolé linéairement à partir de la série X_i de la façon suivante:

$$Y_j = Y_{(i-1)\ p+p'} = \frac{(p'-1)\ X_{i+1} + (p-p'+1)\ X_i}{p} \qquad [2]$$

où i est un entier variant de 1 à k;

p' est un entier variant de 1 à p, sauf pour $i = k$ où $p' = 1$;

p est un entier fixé qui définit le nombre de subdivisions utilisées dans l'interpolation

Moyennes

Par définition on écrit $\bar{X} = \frac{1}{k} \sum_{i=1}^{k} X_i$ et $\bar{Y} = \frac{1}{(k-1)p+1} \sum_{1}^{N} Y_j$; en remplaçant Y_j par sa définition donnée par l'équation [2] et en effectuant la sommation, on obtient:

$$\bar{Y} = \frac{2\ p\ k\ \bar{X} - (p-1)\ (X_1 + X_k)}{2\ [(k-1)\ p + 1]}$$

On s'aperçoit alors que les effets de boût (X_1 et X_k) amènent un biais de $\bar{Y}$ par rapport à $\bar{X}$. Cependant, si la série X_i est suffisamment longue alors $\bar{Y}$ devient un estimateur asymptotiquement sans biais pour $\bar{X}$. En utilisant cette propriété et afin de garder les développements mathématiques les plus concis possible, nous supposerons que les séries X_i sont suffisamment longues, centrées et stationnaires. Avec ces restrictions, on a $\bar{X} = \bar{Y} = 0$ et l'influence des effets de bout devient négligeable.

L'ILI conserve donc la stationnarité du premier ordre; cependant, ce n'est pas le cas, au sens strict pour la stationnarité du deuxième ordre. Comme on le verra plus loin, la variance et les autocovariances dépendent de p' qui représente la position du point interpolé entre deux points de mesure X_i.

Pour contourner ce problème, nous avons retenu une définition moins stricte de la stationnarité donnée par Pankratz (1983, p. 16):

> "if a data-series is stationary, then the variance of any major subset will differ from the variance of any other subset only by chance".

Pour ne pas perdre de vue cette restriction, nous allons utiliser le terme de variance globale pour une sommation effectuée avec un nombre entier de partitions.

Variance globale

Notons σ^2 et r_1 la variance et le coefficient d'autocorrélation d'ordre 1 de la série des valeurs X_i, une expression non-biaisée de la variance S^2 de la série Y_j est donnée par:

$$S^2 = \frac{1}{(k-1)p} \sum_{j=1}^{N} (Y_j - \bar{Y})^2 = \frac{\sum_{j=1}^{N} Y_j^2 - [(k-1)\,p+1]\,\bar{Y}^2}{(k-1)p}$$

En appliquant l'opérateur de la variance au point interpolé $Y(p,p') = Y_{(i-1)p+p'}$ exprimé par l'équation [2], on obtient:

$$S^2(p,p') = \sigma^2 \left[\frac{(p'-1)^2 + (p-p'+1)^2 + 2r_1\,(p'-1)\,(p-p'+1)}{p^2}\right]$$

$$= \sigma^2 \left[1-2\,\frac{(p'-1)\,(p-p'+1)\,(1-r_1)}{p^2}\right]$$

Cette équation montre l'hétéroscédasticité de la série interpolée Y_j, c'est-à-dire que la variance est dépendante de la position p' du point interpolé. La variance minimale apparaît au milieu du segment interpolé:

$$S^2\left(p' = \frac{p}{2}\right) = \sigma^2 \left[\frac{1 + r_1}{2}\right] + \frac{2}{p^2}\,(1 - r_1)] \to \sigma^2 \left(\frac{1 + r_1}{2}\right)$$

Cependant, dans le cas d'une série à pertistance positive, cette non-stationnarité de la variance n'est pas très importante. Pour des valeurs de r_1 fortement positive (proches de 1), cet effet peut être négligé. Pour un nombre entier de partitions, la série Y_j considérée dans son entier possède une variance globale stable et finie qui peut être obtenue en sommant sur p' entre 1 et p; en négligeant les effets de bout, on obtient alors:

$$S2 = \sigma^2 \left[\frac{2p^2+1}{3p^2} + r_1 \frac{p^2-1}{3p^2}\right] = \sigma^2 \left[1-\frac{(1-r_1)\ (p^2-1)}{3p^2}\right] \quad [3]$$

La variance globale asymptotique d'une série interpolée Y_j est toujours plus petite que la variance de la série orignale X_i. L'influence du coefficient d'autocorrélation d'ordre 1 apparaissant généralement plus grande que l'influence du niveau de partition p dans l'établissement de ce phénomène; le tableau 1 met ces résultats en évidence.

Erreur d'estimation introduite par l'ILI

Nous allons maintenant développer analytiquement l'espérance de l'erreur quadratique moyenne (RMSE) résultant d'une interpolation linéaire comportant p points intermédiaires; pour cela, dans une première étape nous considérerons les erreurs présentes dans les p interpolations possibles construites à partir de la série originale dont on enlève p-1 valeurs sur p. Dans une seconde étape, nous pouvons alors grâce à la structure de persistance des processus markoviens qui reste markovienne indépendamment du pas de temps utilisé, déduire l'erreur comprise par les interpolations au pas de temps p.

cas p = 2

Soit une série originale X_i mesurée à des temps unitaires, considérons les deux séries de réalisations possibles X_i' et X_i'' obtenues en interpolant pour remplacer chaque valeur intermédiaire de la série X_i:

$$X'_i = X_1 \quad X'_2 \quad X_3 \quad X'_4 \quad X_5 \ \ldots\ldots$$
$$X''_i = X_2 \quad X''_3 \quad X_4 \quad X''_5 \quad X_6 \ \ldots\ldots$$

Un point interpolé X_i' ou X_i'' est estimé par $(X_{i-1} + X_{i+1})/2$. Les deux séries d'erreurs d'estimation X-X, et X-X" sont alors constituées de zéros alternés avec des termes $X_i - ([X_{i-1} + X_{i+1}]/2)$. En négligeant les effets de bout, l'espérance de la variance des erreurs d'estimation pour les deux séries (de longueur totale 2 k) est alors:

Tableau 1: Réduction de la variance globale introduite par une interpolation linéaire à p partitions, construite sur un processus Markovien de paramètre r_1.

P \ r_1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.750	0.775	0.800	0.825	0.850	0.875	0.900	0.925	0.950	0.975
5	0.680	0.712	0.744	0.776	0.808	0.840	0.872	0.904	0.936	0.968
10	0.670	0.703	0.736	0.769	0.802	0.835	0.868	0.901	0.934	0.967
∞	0.667	0.700	0.733	0.767	0.800	0.833	0.867	0.900	0.933	0.967

$$E^2 = \frac{1}{2k} \sum_{1}^{k} (X_1^2 + \frac{X_{i+1}^2}{4} + \frac{X_{i-1}^2}{4} - X_i X_{i+1} - X_i X_{i-1} + \frac{X_{i+1} X_{i-1}}{2})$$

En introduisant la variance et les covariances de la série markovienne X_i, on obtient:

$$\frac{E^2}{\sigma^2} = \frac{3}{4} - r_1 + \frac{1}{4} r_2 = \frac{3}{4} - r_1 + \frac{1}{4} r_1^2 \qquad [4]$$

Cette expression nous permet de relier la variance des erreurs d'estimation à la variance et à la structure de persistance des valeurs échantillonnées.

Les résultat précédant s'applique à une interpolation de pas p = 2 à partir d'une série mesurée au pas de temps unitaire. Si l'on désire maintenant interpoler une série à un pas de temps inférieur au pas unitaire (créer des intervalles égaux à la moitié du pas de temps unitaire, par exemple), les erreurs d'estimation sont alors inconnues car les valeurs intermédiaires ($\frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ dans cet exemple) n'ont pas été échantillonnées, cependant, il est possible d'estimer la variance de cette erreur grâce à la structure de persistance spéciale des processus markoviens; il suffit de remplacer le coefficient d'autocorrélation d'ordre 1, $r_{½}$, de la série non échantillonnée (de pas de temps ½) par $r_1^{½}$; car dans un monde markovien, r_1 est le coefficient d'autocorrélation d'ordre 2 de cette série. On obtient alors, dans ce cas particulier et pour un processus markovien, échantillonné au pas de temps unitaire, de moyenne 0 et de variance σ^2, une variance globale de la série de temps interpolée au pas de temps ½ égale à $E^2 = \sigma^2 [\frac{3}{4} + \frac{r_1}{4}]$, de plus, l'espérance de la variance de l'erreur d'estimation sera:

$$E^2 = \sigma^2 [\frac{3}{4} - \sqrt{r_1} + \frac{r_1}{4}].$$

Généralisation à d'autres niveaux de partitions p

Avec les mêmes principes décrits dans le cas précédent, considérons maintenant les p réalisations possibles de séries "skippées" où les valeurs intermédiaires ont été interpolées linéairement:

$$X' = \underset{P}{X_1 , X'_2 , X'_3 , \ldots\ldots , X_{p+1}} , \underset{P}{X'_{p+2}, X'_{p+3} , \ldots\ldots , X_{2Xp+1}} \ldots\ldots$$

$$X'' = X_2 , X''_3 , X''_4 , \ldots\ldots , X_{p+2} , X''_{p+3}, X''_{p+4} , \ldots\ldots , X_{2Xp+2} \ldots\ldots$$

P .....

$$X^{(p)} = X_p , X^{(p)}_{p+1} , X^{(p)}_{p+2} , \ldots\ldots , X_{2p} , X^{(p)}_{2Xp+2} , X^{(p)}_{2Xp+2} , \ldots\ldots , X_{3Xp} \ldots\ldots$$

Les valeurs interpolées X'_i , X''_i, ..., $X^{(p)}_i$ sont estimées par:

$$X^{(p)} = \frac{(p-p'+1)\ X_{j+kp-p'+1} + (p'-1)\ X_{j+(k+1)\ p+p'+1}}{p} \quad [5]$$

Dans cette expression; l'indice j varie de 1 à p et représente les lignes, l'indice p' varie de 1 à p et représente les colonnes alors que k est un entier variant de 0 à N/P identifiant la partition.

Considérons maintenant l'ensemble de p séries d'erreurs d'interpolation: (X-X'), (X-X"), ... $(X-X^{(p)})$ formées de 0 et de termes:

$$X_{j+kp} - \frac{(p-p'+1)\ X_{j+kp-p'+1} + (p'-1)\ X_{j+(k+1)p-p'+1}}{p}$$

La variance de l'erreur d'estimation pour les p séries (de longueur totale pN) peut alors être exprimée par 3 termes (en négligeant les effets de bout):

$$S^2 = \frac{1}{PN} \left\{ \sum_{p'=1}^{P} \sum_{j=1}^{P} \sum_{k=0}^{N/P} \left(X^2_{j+kp} + \left[\frac{(p-p'+1)\, X_{j+kp-p'+1} + (p'-1)\, X_{J+(k+1)p-p'+1}}{p} \right]^2 \right. \right.$$

$$\left. \left. - 2\, X_{j+kp} \frac{(p-p'+1)\, X_{J+kp-p'+1} + (p'-1)\, X_{J+(k+1)p-p'+1)}}{p} \right\} \right.$$

Les deux premières sommations représentent les variances de la série originale et des p séries interpolées:

$$\sigma^2 \left[1 + \frac{(2p^2 + 1) + r_p\, (p^2-1)}{3p^2} \right]$$

La troisième sommation donne:

$$\frac{\sigma^2}{p} \left[\frac{2}{p} - \frac{4}{p^2} \sum_{p'=1}^{p} (p-p'+1)\, r_{p'-1} \right]$$

Si la structure de persistance du processus est markovienne, alors:

$$\sum_{p'=1}^{p} (p-p'+1)\, r_{p'-1} = \frac{p\, r_1}{1-r_1} - r_1 \frac{1-r_1^{\,p}}{(1-r_1)^2}$$

La variance des erreurs d'estimation des séries "p-skippées" dont les valeurs intermédiaires ont été interpolées est alors:

$$S^2 = \sigma^2 \left[1 + \frac{(2p^2+1) + r_1^{\,p}\, (p^2-1)}{3p^2} - \frac{2}{p} \frac{1+r_1}{1-r_1} + \frac{4\, r_1}{p^2} \frac{1-r_1^{\,p}}{(1-r_1)^2} \right] \quad [6]$$

En appliquant le même raisonnement que pour le cas p = 2 nous pouvons, grâce à la structure spéciale des processus markoviens, estimer l'espérance de la variance de l'erreur d'estimation pour une série échantillonnée au pas de temps unitaire et interpolée aux temps intermédiaires $\frac{1}{p}, \frac{2}{p}, \ldots \frac{p-1}{p}$. Il suffit alors de remplacer dans l'équation [6] r_1 par $r_1^{1/p}$, car dans ces séries, r_1 constituent le coefficient d'autocorrélation d'ordre p:

$$E^2 = \sigma^2 \left[1 + \frac{(2\,p^2+1) + r_1\,(p^2-1)}{3p^2} - \frac{2}{p}\,\frac{1+r_1^{1/p}}{1-r_1^{1/p}} + \frac{4r_1^{1/p}}{p^2}\,\frac{1-r_1}{(1-r_1^{1/p})^2}\right] \quad [7]$$

Cette équation détermine l'espérance de l'erreur quadratique moyenne (RMSE) de l'interpolation, relatif à la variabilité σ du processus, et fonction du coefficient d'autocorrélation d'ordre 1 ainsi que du niveau p des partitions.

Si p tend vers l'infini, l'équation [7] peut être développée selon la Règle de l'Hospital:

$$P\left[1-r_1^{1/P}\right] \xrightarrow[P\to\infty]{} -\operatorname{Ln} r_1 \quad \text{et} \quad P^2\left[1-r_1^{1/P}\right]^2 \xrightarrow[P\to\infty]{} \operatorname{Ln}^2 r_1$$

Ainsi l'expansion asymptotique de l'équation [7] devient, pour des grandes valeurs de p:

$$E^2 = \sigma^2 \left[\frac{5+r_1}{3} + \frac{4}{\operatorname{Ln} r_1} + \frac{4(1-r_1)}{\operatorname{Ln}^2 r_1}\right]$$

Le tableau 2 illustre l'erreur quadratique moyenne relative E/σ résultant d'une interpolation avec p partitions, construite à partir d'un processus markovien de paramètre r_1. Ce tableau montre que le facteur prédominant est le paramètre de persistance r_1, ainsi, pour des grandes valeurs de r_1, l'interpolation n'introduit que très peu de variance d'erreur; de plus une augmentation dans le niveau de partition p n'est pas un facteur important. La figure 1 met graphiquement en évidence ces faits. Pour des valeurs faibles de r_1, le processus d'interpolation, comme toutes les méthodes d'estimation, est moins efficace. De plus pour r_1=0 (le cas des valeurs échantillonnées indépendantes), on obtient même un erreur quadratique d'estimation

Tableau 2: Espérance de l'erreur quadratique moyenne relative résultant d'une interpolation linéaire à p partitions, construite sur un processus Markovien de paramètre r_1.

P \ r_1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.866	0.677	0.594	0.527	0.466	0.410	0.354	0.297	0.236	0.162
5	1.13	0.782	0.680	0.600	0.530	0.465	0.401	0.337	0.267	0.184
10	1.21	0.796	0.691	0.610	0.539	0.472	0.408	0.342	0.271	0.186
∞	1.29	0.801	0.695	0.613	0.541	0.475	0.410	0.344	0.272	0.187

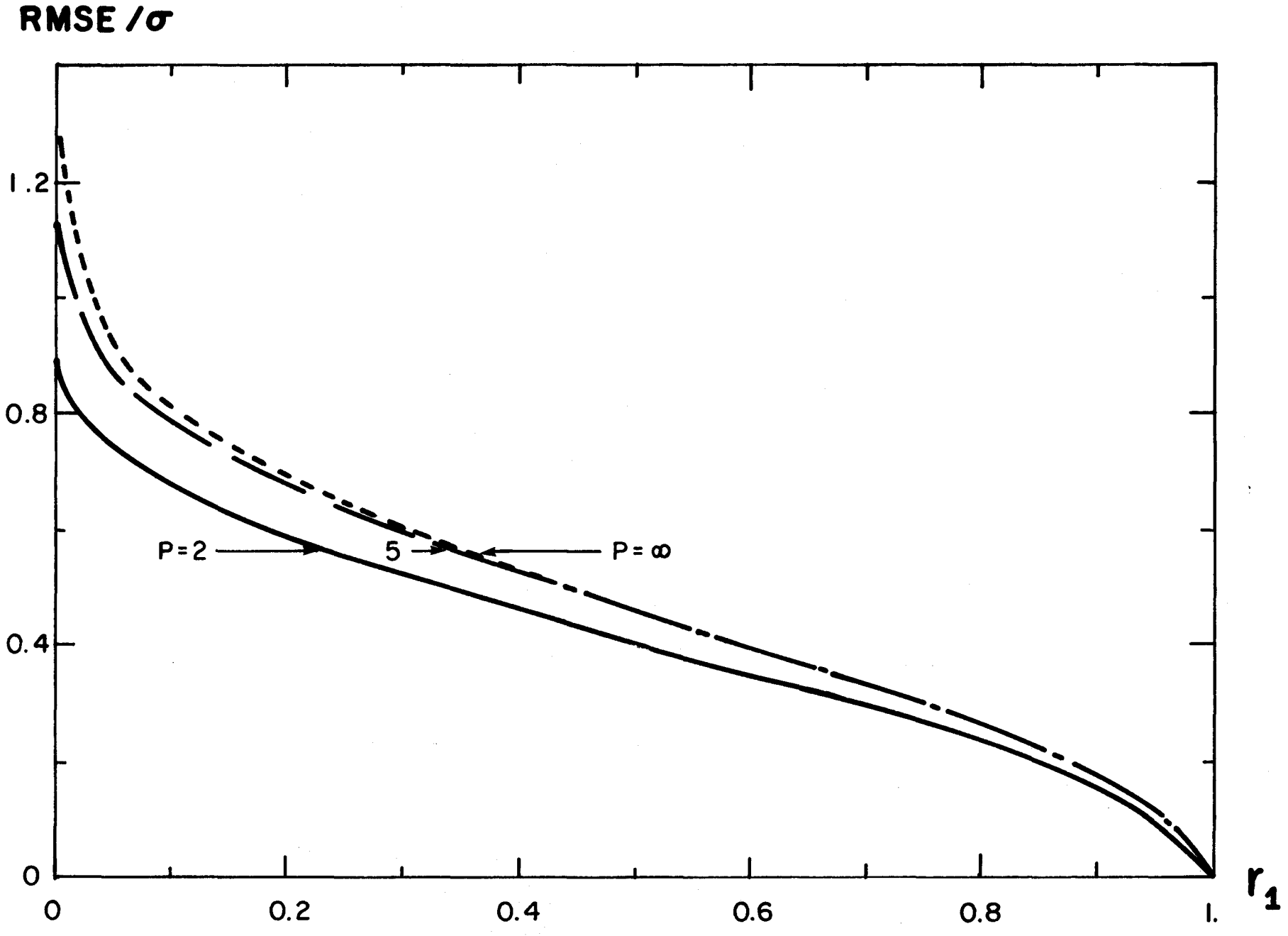

Figure 1: Variance d'estimation introduite selon r_1 et p.

plus grande que l'estimateur "global" qui remplace les valeurs intermédiaires (non-échantillonnées) par la moyenne générale; cette méthode d'estimation amenant une espérance d'erreur quadratique moyenne relative de: $E'/\sigma = [\frac{p-1}{p}]^{\frac{1}{2}}$.

Les résultats analytiques obtenus pour la variance et l'espérance de la variance de l'erreur introduite par l'interpolation linéaire d'une série Markovienne X_i $[0; \sigma^2; r_k=r_1^k]$ sont présentés dans le tableau 3.

Application à l'erreur introduite par des valeurs manquantes occasionnelles

On considère une série de temps de longueur N dont n valeurs non-consécutives sont manquantes. Selon les résultats précédents, une série de longueur $\frac{N}{2}$ interpolée au pas de temps 2 comporterait pour chaque valeur crée, une espérance de l'erreur quadratique moyenne de $\frac{E}{\sigma} = [\frac{3}{2} - 2\sqrt{r_1} + \frac{r_1}{2}]^{\frac{1}{2}}$ donc pour n valeurs reconstituées, l'espérance de l'erreur quadratique moyenne vaut: $\frac{E}{\sigma} = [\frac{n}{N}(\frac{3}{2} - 2\sqrt{r_1} + \frac{r_1}{2}]^{\frac{1}{2}}$ ce qui donne un contrôle quantitatif sur l'erreur globalement introduite dans la série.

Application au calcul des débits massiques

Dans l'interprétation des phénomènes de transport chimique en rivière, l'estimation des débits massiques constitue une étape essentielle, soit $c_{(t)}$ et $q_{(t)}$ les processus continus représentant les évolutions des concentrations et des débits dans le temps, le produit $c_{(t)} \cdot q_{(t)}$ représente le flux instantanné de matière exporté et $\int_{t_1}^{t_2} c_{(t)} \cdot q_{(t)} \cdot dt$ représente le débit massique écoulé entre les deux instants t_1 et t_2. Si on dispose de séries temporelles équidistantes et simultanées, le débit massique est évalué par une sommation de type $\sum_{i=1}^{N} c_i q_i \cdot \Delta t$ en utilisant la méthode des trapèzes pour l'intégration numérique. Si les fréquences d'échantillonnage ne sont pas les mêmes, certaines manipulations de données (aggrégation, interpolation) sont nécessaires afin de générer des séries synchrones et équidistantes.

Tableau 3: Synthèse des résultats analytiques obtenus pour l'interpolation d'une série $X_i[0, \sigma^2, r_k=r_1^k]$.

	Variance relative de l'interpolation S^2/σ^2	Espérance relative de la variance de l'erreur E^2/σ^2
a) cas général		
valeur locale	$1-2\,\frac{(p'-1)(p-p'+1)(1-r_1)}{p^2}$	$2\left[1-\frac{(p'-1)(p-p'+1)(1-r_1)}{p^2}-\frac{p-p'+1}{p}\,r_1^{(p'-1)/p}-\frac{p'-1}{p}\,r_1^{(p-p'+1)/p}\right]$
valeur globale	$1-\frac{(p^2-1)(1-r_1)}{3\,p^2}$	$1+\frac{(2p^2+1)+r_1\,(p^2-1)}{3\,p^2}-\frac{2}{p}\,\frac{1+r_1^{1/p}}{1-r_1^{1/p}}+\frac{4}{p^2}\,\frac{1-r_1}{(1-r_1^{1/p})^2}$
b) cas $p=2$		
$p'=1$	1	0
valeur locale		
$p'=2$	$(1+r_1)/2$	$(1-r_1^{\frac{1}{2}})\,(3-r_1^{\frac{1}{2}})/2$
valeur globale	$(3+r_1)/4$	$(1-r_1^{\frac{1}{2}})\,(3-r_1^{\frac{1}{2}})/4$

Comme il existe plusieurs combinaisons possibles permettant d'obtenir une fréquence commune aux séries transformées, nous nous proposons d'examiner la conséquence de ces manipulations de données.

Si les données transformées du processus $c_{(t)}$ sont destinées à être intégrées dans le temps par la méthode des trapèzes, les erreurs introduites par l'interpolation linéaire et l'élimination de valeurs intermédiaires ont une formulation similaire: la seule différence provient du nombre de points contenus dans la série originale et du coefficient d'autocorrélation associé à ce pas de temps.

Ces considérations doivent servir de base à la stratégie à adopter dans le cas de débits massiques. La variance de l'erreur sur un débit massique est un problème plus complexe encore, impliquant éventuellement les covariances ou les relations fonctionnelles entre $c_{(t)}$ et $q_{(t)}$.

Conclusion

Dans un contexte de grands échantillons Markoviens, nous avons développé la variance globale et la variance de l'erreur d'estimation résultant de l'interpolation linéaire itérative. Ces résultats montrent que l'information fictive introduite par l'interpolation linéaire est beaucoup plus sensible au paramètre de persistance r_1 qu'au nombre p des partitions.

D'un point de vue pratique, les développements précédents permettront aux utilisateurs de contrôler la quantité d'information fictive externe qu'ils sont prêts à accepter d'introduire afin de pouvoir utiliser l'information échantillonnée à une fréquence plus grande que celle des prélèvements.