Record Number: 16720
Author, Monographic: Ouarda, T. B. M. J.//Gaudet, J.//Tremblay, M. E.//Bobée, B.
Author Role:
Title, Monographic: Revue bibliographique des techniques statistiques de conception et de consolidation des réseaux hydrométéorologiques
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 2000
Original Publication Date: Février 2000
Volume Identification:
Extent of Work: 37
Packaging Method: pages
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, rapport de recherche
Series Volume ID: 557
Location/URL:
ISBN: 2-89146-333-1
Notes: Rapport annuel 2000-2001
Abstract: Rapport préparé pour Hydro-Québec, Unité prévisions et ressources hydriques
Call Number: R000557
Keywords: rapport/ ok/ dl

Revue bibliographique des techniques statistiques de conception et de consolidation des réseaux hydro-météorologiques

Février 2000 *Rapport de recherche No R-557*

REVUE BIBLIOGRAPHIQUE DES TECHNIQUES STATISTIQUES DE CONCEPTION ET DE CONSOLIDATION DES RÉSEAUX HYDRO-MÉTÉOROLOGIQUES

Rapport préparé pour

Hydro-Québec
Unité Prévisions et Ressources Hydriques
75 Boul. René-Lévesque Ouest, 9^{e} étage
Montréal (Québec) H2Z 1A4

À l'attention de M. Claude Gignac

par

Taha B.M.J. Ouarda
Jocelyn Gaudet
Marie-Eve Tremblay
Bernard Bobée

Chaire en Hydrologie Statistique
Institut National de la Recherche Scientifique, INRS-Eau
2800, rue Einstein, C.P. 7500, Sainte-Foy (Québec) G1V 4C7

Rapport de recherche no R-557

Février 2000

Table des matières

Sommaire

Le présent rapport présente une revue bibliographique des techniques de conception et de consolidation des réseaux hydro-météorologiques. Pour la conception et la gestion des ouvrages hydrauliques, il est essentiel de bien connaître les propriétés des variables hydro-météorologiques. Cette connaissance est basée sur l'utilisation de réseaux hydro-météorologiques. Le présent travail porte sur les critères statistiques intervenants dans le choix de la configuration des réseaux.

Entropie et information de Shannon (conception et rationalisation de réseaux)

L'entropie est une mesure de l'information contenue dans les observations d'une station de jaugeage. Il s'agit de maximiser l'information communiquée par les stations choisies pour faire partie du réseau. Dans le cas d'un réseau hydrographique, l'information communiquée dépend en grande partie de l'incertitude sur les intrants, c'est-à-dire sur la mesure des débits (sur la mesure des débits lors de l'établissement de la courbe de tarage, et lors de l'utilisation subséquente de la courbe de tarage à partir d'une mesure de niveau d'eau). Pour maximiser l'information, il faut alors minimiser l'incertitude sur les débits. Les principaux problèmes de l'approche sont liés à la grande difficulté de représenter les fonctions économiques, ainsi qu'aux difficultés de définir les distributions de probabilité utiles à l'estimation de la quantité d'information.

Analyse de classification (rationalisation)

Avec cette approche, on analyse premièrement le réseau pour identifier les stations véhiculant essentiellement la même information hydrologique, donc la redondance. Cette étape est réalisée à l'aide d'indices de similitude basée sur une corrélation pondérée. Par la suite, on élimine les stations qui sont bien prédites par d'autres stations du réseau. Le résultat est une diminution du nombre total de stations utilisées, et une perte minimisée de la quantité d'information totale. Il est aussi possible de choisir des stations spécifiques, préférant celles qui transmettent une information unique. Il est cependant difficile de considérer certains facteurs comme la durée et la qualité d'enregistrement, les préférences des utilisateurs quant au choix des stations, ainsi que les relations spatiales dans le choix des stations à conserver.

Approche bayésienne (conception, augmentation de réseaux)

L'approche bayésienne permet de modéliser de façon probabiliste les différentes sources d'incertitude associées aux paramètres hydrologiques, ainsi que les paramètres économiques reliés aux coûts et aux bénéfices découlant de l'opération de réseaux hydro-météorologiques. L'approche bayésienne a été appliquée à la conception et l'augmentation des réseaux de mesure, en se basant sur les données observées ou des données générées. L'approche bayésienne devra être poursuivie dans le cadre des travaux visant à estimer la valeur économique des données hydro-météorologiques.

Covariance spatiale et krigeage (conception, rationalisation, extension)

La covariance spatiale permet d'évaluer quelles stations peuvent être retirées tout en conservant le plus d'information spatiale possible à l'aide du variogramme qui permet d'estimer la structure de la dispersion spatiale du réseau. Le principal problème est que cette technique montre souvent moins de variabilité que le monde réel, entraînant une sous-estimation des besoins en données. Le krigeage est une technique permettant d'estimer de façon optimale le minimum de variance spatiale reliée à un réseau, et ce pour tout point de la région à l'étude, à l'aide d'un variogramme pondéré dans l'espace. Pour éliminer une station avec cette technique, on choisit celle qui cause la plus petite augmentation dans la variance spatiale de l'estimation de la variable d'intérêt. Le krigeage est aussi affecté par une sous-estimation de la variance dans l'espace car les fonctions de covariance spatiale et les variances liées aux erreurs ne sont pas toujours connues avec certitude.

Corrélations entre les stations (rationalisation)

Dans un cadre strictement statistique, la technique vise à éliminer les stations d'un réseau qui sont fortement corrélées à d'autres stations, et à estimer les valeurs futures aux stations éliminées à l'aide de la régression et de fonctions de transfert. Dans cette démarche, il faut de plus évaluer un indice qui nous informe sur l'incertitude liée aux valeurs mesurées et celles qui seront estimées, ce jusqu'à l'horizon de temps qui nous intéresse. Les stations éliminées seront alors celles qui sont bien prédites par les stations restantes, et dont l'élimination ne causera pas une trop forte augmentation de l'indice d'incertitude. Cependant, cette technique tient difficilement compte de la structure spatiale de l'information. Il faut de plus considérer que cette

technique est appropriée pour des variables semblables aux débits annuels, mais pas pour les débits quotidiens par exemple.

Valeur économique des données: analyse des coûts

Ces études se tentent d'évaluer si les gains économiques résultants du jaugeage dépassent les coûts d'installation et d'opération de la station. Il s'agit d'analyses coûts/bénéfices, et la principale difficulté réside dans l'évaluation des bénéfices espérés de l'utilisation de la station, car les coûts sont généralement bien connus. Une autre approche est de maximiser la précision de l'estimation, c'est-à-dire de minimiser la variance, tout en minimisant les coûts. Malheureusement, ces techniques ne renseignent pas directement sur les caractéristiques spatiales de cette même variance, et ne transmettent pas d'information sur les emplacements où des stations doivent être ajoutées ou éliminées.

Valeur économique des données: Coûts d'obtention et valeur des débits

La valeur de données hydrologiques supplémentaires peut être vue comme une réduction de la pénalité économique attendue à cause de l'incertitude que la nature des processus entraîne. Il s'agit ici d'évaluer quelle sera la diminution de cette pénalité économique si on allonge la période de jaugeage par rapport à une élimination de la station. Si la pénalité économique est supérieure aux coûts d'opération, il est alors avantageux de conserver la station pour minimiser les pertes. La difficulté de l'approche consiste à évaluer la valeur future des données. En utilisant l'information connue et la théorie des prévisions inférieures cohérentes, il devient possible d'estimer un intervalle autour de la valeur économique attendue des données, dans lequel la valeur réelle des données se trouvera. Cette valeur sera la valeur d'une information jugée parfaite. Cette approche peut être appliquée à chacune des stations d'un réseau afin de juger si la valeur de l'information parfaite dépasse le coût d'opération, donc si chaque station sera rentable. Cependant, cette technique ne permet pas d'étudier la structure spatiale de l'information, ou de la variance d'estimation.

Avant-Propos

Cette étude a été réalisée dans le cadre du projet T1.2 de la Chaire en Hydrologie Statistique à l'INRS-Eau traitant de « l'évaluation des besoins en données d'Hydro-Québec et la gestion des réseaux hydro-météorologiques ». Les auteurs du rapport tiennent à exprimer leur reconnaissance à MM. Claude Gignac et René Roy pour leurs commentaires et leurs réflexions.

1 Introduction

« Data ! data ! data ! » he cried impatiently. « I can't make bricks without clay »
Sherlock Holmes

La bonne connaissance des propriétés statistiques des variables hydro-météorologiques (débits horaires, débits de crues ou d'étiages, distribution spatiale ou temporelle des précipitations, températures journalières, etc.) est importante pour la conception (par exemple, dans le cas de la construction d'un évacuateur de crue ou d'une digue) et la gestion (par exemple, dans le cas de l'opération d'une centrale hydroélectrique) des ouvrage hydrauliques. Une estimation précise et fiable est donc essentielle pour effectuer d'une façon adéquate certaines tâches telle que la prévision des apports en eau. Afin de répondre à ses besoins en données, Hydro-Québec gère un réseau hydro-météorologique comprenant de nombreuses stations et utilise également de l'information provenant d'autres réseaux privés et publics tel que le réseau du ministère de l'Environnement du Québec (MEQ) ou le réseau d'Environnement Canada (EC). Du côté météorologique, le réseau météorologique coopératif québecois (RMCQ) est utilisé. Ce dernier comprend des stations du Ministère de l'Environnement du Québec, d'Environnement Canada, d'Alcan, et de la SOPFEU (Société de protection des forêts contre le feu).

Cependant, les contraintes financières peuvent avoir un impact important sur la conception et la gestion des réseaux de mesure et peuvent, à la limite, rendre une rationalisation nécessaire. Ainsi, au cours des dernières quinze années plusieurs stations de mesure ont été éliminées à cause des restrictions budgétaires. Il est alors évidemment important d'effectuer la conception et/ou la rationalisation des réseaux de mesure de façon à acquérir le maximum d'information hydrologique possible. La conception des réseaux de mesure vise à définir la configuration de réseau la plus efficace, i.e. la configuration qui remplit toutes les contraintes de conception (incluant les contraintes de coût) tout en maximisant la quantité d'information qui est produite. Hydro-Québec s'est déjà penché sur la problématique de l'augmentation des réseaux hydro-météorologiques (voir par exemple Bisson, 1989), l'évaluation de la qualité des mesures acquises (Gauthier et Roy, 1988) et l'influence de la configuration du réseau météorologique sur la prévision des apports (Roberge et Bisson, 1982).

Dans cette optique, la Chaire en Hydrologie Statistique à l'INRS-Eau est mandatée pour évaluer l'adéquation du réseau hydro-météorologique existant pour les besoins d'Hydro-Québec et pour donner des recommandations sur sa consolidation. La contribution de l'INRS-Eau porte principalement sur la prise en compte de critères statistiques dans le choix de la configuration du réseau. Soulignons que pour effectuer une rationalisation efficace et adéquate de nombreux autres facteurs non statistiques devraient être considérés conjointement, tels que la taille du bassin, le nombre de demandes d'information dans le passé, le nombre estimé de demandes d'information dans le future, les constructions d'envergure prévues, les problèmes particuliers de crues et/ou d'étiage, les raisons historiques, le suivi du changement climatique (station à long terme), la reconstitution à l'aide des modèles déterministes, les différents coûts associés au maintien des stations et les aspects logistiques

Le présent document est destiné à présenter une brève revue bibliographique des différentes techniques de conception et de consolidation des réseaux hydro-météorologiques. Différentes classes d'approches seront considérées et pour chaque approche on identifiera si elle a été appliquée pour la rationalisation, la conception ou l'augmentation des réseaux de mesure. Dans ce document, nous utilisons les définitions suivantes: la conception d'un réseau consiste à choisir l'emplacement idéal pour ses stations; la consolidation réfère à la modification d'un réseau afin d'augmenter son efficacité; l'augmentation d'un réseau est simplement l'ajout de stations; et finalement la rationalisation correspond à une diminution de la taille d'un réseau. Il est important de noter qu'un déplacement d'une station de mesure peut être considéré comme une élimination de station suivie d'un ajout d'une autre station.

2 Entropie et information de Shannon

L'entropie est définie comme l'incertitude associée à un événement aléatoire ou l'information contenue dans cet événement. Dans le cas des réseaux, il s'agit d'une mesure de l'information contenue dans les observations d'une station de jaugeage ou météorologique.

Voici certains aspects de la théorie de l'information dans le cas discret. Un poste de communication reçoit un message aléatoire unique X, qui peut prendre les valeurs x_i : $i = 1, ..., N$ et produit un signal de sortie aléatoire Y, qui peut prendre les valeurs y_j : $j = 1, ..., M$. On connaît les distributions de X et Y, soit $P[x_i]$ et $P[y_j]$ ainsi que leur densité conjointe $P[x_i, y_j]$. La quantité d'information $T(x_i; y_j)$ fournie par y_j sur la valeur de l'input x_i a été définie de la façon suivante par Shannon, :

$$T(x_i; y_j) = \log P[x_i | y_j] - \log P[x_i]$$

Si y_j et x_i sont complètement indépendants, alors

$$P[x_i | y_j] = P[x_i]$$

et

$$T(x_i; y_j) = 0$$

Par contre, si la sortie y_j dépend totalement de l'intrant x_i , alors

$$P[x_i | y_j] = 1$$

et

$$T(x_i; y_j) = -\log P[x_i]$$

Cette valeur est la quantité d'information communiquée lorsqu'il n'y a plus d'incertitude concernant l'événement x_i. Dit autrement, c'est une mesure de l'incertitude associée à x_i. L'incertitude moyenne associée à l'intrant aléatoire X est :

$$-\sum_i P[x_i] \log P[x_i] = H(X)$$

où $H(X)$ est l'entropie de X. On a aussi que $T(X;Y)$ est la somme des informations communiquées par chaque paire d'intrants et de sorties possibles, pondérée par la probabilité d'obtenir cette paire.

$$T(X;Y) = \sum_i \sum_j P[x_i, y_j]\left(\log P[x_i|y_j] - \log P[x_i]\right)$$

$T(X;Y)$ s'appelle aussi information mutuelle et peut s'exprimer en fonction des entropies $H(X)$, $H(Y)$ et de l'entropie conjointe $H(X,Y)$:

$$T(X;Y) = H(X) + H(Y) - H(X,Y)$$

ou bien

$$T(X;Y) = H(X) - H(X|Y)$$

où

$$H(X|Y) = -\sum_i \sum_j P[x_i, y_j] \log P[x_i|y_j]$$

On a aussi que $T(X;Y) = T(Y;X)$

Cette théorie peut être généralisée à v intrants et à w sorties. Dans le cas continu, les sommations sont remplacées par des intégrales. On a alors :

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log|f(x)| dx$$

et

$$T(X;Y) = -\int_{-\infty}^{\infty} f(x)\log|f(x)|dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y)f(x|y)\log|f(x|y)|dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

Plusieurs méthodes de design ou de rationalisation des réseaux utilisent les concepts d'entropie et d'information, en voici quelques-unes :

Critère de maximisation de la communication de l'information (Caselton et Husain, 1980) (conception)

Considérons deux cas, le premier concerne l'information transmise par les mesures et les estimations provenant d'un réseau permanent de n stations devant être révélatrices des conditions aux m stations d'un réseau temporaire dispersées à différents endroits dans la région. Le second cas concerne l'information transmise par un réseau permanent de n stations en considérant que ces stations ne représentent qu'une petite partie des k sites de la région. On a donc les notations suivantes :

m = le nombre de sites d'un réseau temporaire dense d'où proviennent les données a priori
n = le nombre de stations dans un réseau permanent, n < m

S = ensemble des mesures hydrologiques provenant des m sites du réseau

S_i^n = ensemble des n variables représentant les mesures pour la i^e combinaison de n sites

Ψ = ensemble de variables qui décrivent les vraies mesures pour un très grand nombre k d'endroits dans la région, k >> m.

Le problème de conception de réseaux est défini comme la sélection de n stations permanentes provenant de m sites donnés. Pour résoudre ce problème, on doit maximiser la diminution d'incertitude liée aux m sites, ce qui revient à maximiser les informations transmises entre les m sites et les n stations du réseau permanent. La problématique se résume donc au choix des n stations parmi les m, qui maximisent l'information transmise. Pour ce faire, on maximise l'entropie des groupes de i stations parmi les m:

$$\mathrm{Max}\left[H\left(S_i^n\right)\right]$$

Pour calculer l'entropie, il faut connaître la distribution des données. Dans le travail de Caselton et Husain (1980), on dispose des précipitations journalières pour une période de deux ans. Une loi discrète est ajustée à ces données en subdivisant le rang entre zéro et la plus grande précipitation journalière en huit intervalles de même longueur. Les densités marginales et conjointes ont été calculées par les fréquences relatives.

Dans la deuxième situation, il apparaît que l'entropie des données du réseau permanent est une borne supérieure pour la réduction de l'incertitude moyenne concernant le vrai bassin hydrographique. Ceci est décrit comme suit :

$$T\left(\hat{\Psi},\Psi\right) \le H\left(S_1^n\right)$$

où $\hat{\Psi}$ est la combinaison des sorties des stations permanentes et des estimations.

Husain, 1987 (conception)

Il est souvent plus avantageux de considérer des distributions continues pour les calculs d'entropie. Husain (1987) a donc calculé des entropies univariées, bivariées et multivariées pour des lois continues. Pour les entropies univariées, il a considéré les distributions normale, log-normale, gamma, bêta, GEV et exponentielle. Dans le cas bivarié, il a fourni les entropies marginales, les entropies conjointes et l'information de Shannon pour les mêmes distributions, excepté la distribution bêta. Enfin, dans le cas multivarié, il a considéré seulement les formes multivariées de la normale et de la log-normale. Pour connaître l'ensemble des entropies et des informations de Shannon dérivées de ces densités, il faut se référer au travail original (Husain, 1987).

Ici encore, le but est de choisir les n stations parmi m (n < m), pour que le maximum d'information concernant les m - n stations qui ne sont plus jaugées soit communiqué. La fonction qui permet de faire le choix (Matalas, 1973) est :

$$Z = \text{Max} \sum_{i=1}^{m} \sum_{j=1}^{m} T\left(X_i; Y_j\right) \delta_{i,j}$$

où Z est appelé «fonction objective» et $\delta_{i,j}$ est une variable de décision, qui prend la valeur 0 si la station $i = j$ doit être supprimée. $\delta_{i,j}$ vaut aussi 0 quand l'information ne peut être communiquée de j à i. Par contre, lorsque l'information est communiquée de j à i, $\delta_{i,j}$ est égal à 1. Deux contraintes sont utilisées avec cette fonction :

1) L'information provenant de seulement une station peut être communiquée à une autre station, ou

$$\sum_{j=1}^{m} \delta_{i,j} = 1$$

2) Si le nombre maximal de stations pouvant être retenues est limité à n à cause des contraintes budgétaires, alors

$$\sum_{i=1}^{m} \delta_{i,j} = n$$

Au niveau multivarié, l'information communiquée par un ensemble de stations est calculée dans l'entropie multivariée. Si $S = (X_1, X_2, ..., X_m)$ contient les variables représentant les conditions hydrologiques des m sites et $S_0 = (Y_1, Y_2, ..., Y_n)$ contient l'information communiquée aux n stations à propos des données hydrologiques aux m sites alors

$$Z = \text{Max } T\left(X_1, X_2, ..., X_n; Y_i, Y_j, ..., Y_n\right)$$

Plusieurs combinaisons de $S_0 = (Y_1, Y_2, ..., Y_n)$ sont choisies et la combinaison donnant la communication d'information maximale est la combinaison de n stations optimales.

Utilisation du DIT et estimation de la densité par l'estimation non-paramétrique (Yang et Burn, 1994) (rationalisation)

L'information mutuelle permet de mesurer la communication d'information. L'entropie et l'information mutuelle fournissent des mesures quantitatives de :

1) l'information à une station
2) l'information transférée et perdue pendant la communication
3) la description de la relation entre les stations selon leurs caractéristiques de communication d'information

Le problème lié à l'utilisation de l'entropie dans la conception de réseaux est de représenter les données hydrométriques à l'aide de fonctions de densités multivariées. Dans la majorité des cas, dont celui décrit précédemment (Husain, 1987), on suppose que les données proviennent d'une certaine distribution continue et l'entropie est calculée à partir de la densité correspondante. Cette façon de procéder a plusieurs désavantages. Par exemple, la distribution peut être mal choisie.

Une solution à ce problème pourrait être l'application d'une méthode non-paramétrique d'estimation de la densité.

La méthode décrite par Yang et Burn (1994) contient deux étapes. La première étant la régionalisation du réseau à l'aide de l'entropie. Ceci implique l'estimation de fonctions de densités à deux dimensions pour les paires de stations de jaugeage, au moyen de l'estimation non-paramétrique. Ensuite, on définit un «directional information transfer index (DIT)» (ou indice directionnel de transfert de l'information) et on le calcule pour chaque paire. La régionalisation se fait à partir des valeurs de DIT. La deuxième étape consiste à sélectionner les stations les plus représentatives de chaque sous-groupe homogène obtenu durant la première étape.

On sait que, lorsque deux stations sont complètement dépendantes, l'information à un site peut parfaitement être communiquée à l'autre sans aucune perte. Dans ce cas, $T=H$. Toutes les situations intermédiaires entre la totale dépendance et la totale indépendance conduisent à des valeurs de T entre zéro et H. T est donc un indicateur de la capacité à communiquer l'information et du degré de dépendance de deux stations. On définit le DIT comme suit :

$$\mathrm{DIT} = T/H = \left(H - H_{\mathrm{lost}}\right)/H = 1 - H_{\mathrm{lost}}/H$$

où

$$H_{\mathrm{lost}} = H(X|Y)$$

La signification physique du DIT est la fraction de l'information transférée d'un site à l'autre. Le DIT varie entre zéro et un. La valeur zéro correspond à une totale indépendance entre les stations entre lesquelles aucune information n'est communiquée. La valeur un, à l'opposé, correspond à une dépendance complète alors qu'aucune information n'est perdue. Notons que $\mathrm{DIT}_{xy} \neq \mathrm{DIT}_{yx}$. DIT_{xy} est la fraction de l'information prédite par X à propos de la station Y. Entre les deux stations d'une paire, celle dont la valeur de DIT est supérieure devrait être gardée en priorité, à cause de sa grande capacité à prédire l'information à l'autre site.

Une application directe du DIT est représentée par la régionalisation de réseaux. Si DIT_{xy} et DIT_{yx} sont élevés, les deux stations devraient être groupées ensembles. Si les deux DIT sont faibles, les stations devraient rester en deux groupes séparés. Si seulement un DIT, DIT_{xy} par exemple, est élevé, alors la station Y peut joindre la station X si elle n'appartient pas à un autre groupe, Si tel est le cas, X ne peut pas entrer dans le groupe de Y. Lorsque toutes les stations d'un groupe ont de fortes connections mutuelles (DIT_{xy} et DIT_{yx} élevés), on peut poursuivre la sélection à l'aide du critère S-DIT, définit comme suit :

$$\mathrm{S-DIT}_i = \sum_{j=1, j\neq i}^{m} \mathrm{DIT}_{ij}$$

La station du groupe qui a le S-DIT le plus élevé devrait être gardée dans le réseau.

La méthode décrite par Yang et Burn (1994), à la différence des méthodes décrites précédemment, utilise l'estimation non-paramétrique plutôt que paramétrique pour estimer la densité et ainsi calculer l'entropie. Nous n'expliquerons pas les détails de l'approche d'estimation non-paramétrique dans le présent document, mais le lecteur est référé à l'article original (Yang et Burn, 1994).

D'une façon générale, les études qui reposent sur l'entropie font face à deux problèmes limitatifs. Le premier est la grande difficulté que les chercheurs ont à représenter les fonctions économiques utilent au calcul des fonctions objectives servant à évaluer l'entropie; le second problème concerne la difficulté de définir correctement les distributions de probabilité qui sont utilisées pour évaluer la quantité d'information disponible.

3 Analyse de classification

Burn et Goulter, 1991 (rationalisation)

Tout comme la dernière méthode de rationalisation de réseaux, la procédure proposée par Burn et Goulter (1991) se fait en deux étapes. La première consiste à réaliser une analyse de classification hiérarchique pour identifier les regroupements de stations similaires. Durant la deuxième étape, une station dans chaque groupe formé précédemment est sélectionnée pour être conservée dans le réseau rationnalisé. La philosophie de la méthode est de rechercher l'information redondante, c'est-à-dire rechercher les groupes de stations de jaugeage qui fournissent essentiellement la même information à propos des caractéristiques des débits dans une région. L'approche proposée par Burn et Goulter (1991) permet d'utiliser le jugement personnel de l'utilisateur pour influencer le processus de sélection des stations. Il est alors possible de choisir plus d'une station dans un groupe donné. La méthodologie a été appliquée pour la rationalisation du réseau hydrométrique situé dans la région du bassin de la rivière Pembina au sud de la province du Manitoba, Canada.

La phase initiale permet de regrouper les stations grâce à un indice de similitude indiquant jusqu'à quel point l'information obtenue aux différentes stations est équivalente. La procédure utilisée pour l'analyse de classification est la méthode « average-linkage clustering » (classification basée sur la moyenne des groupes). Les auteurs considèrent qu'il y a au moins trois composantes importantes pour définir la similitude entre deux stations : les caractéristiques de débits moyens, de débits élevés et de débits faibles. L'indice de similitude utilisé est basé sur une pondération des corrélations calculées à partir de ces différents débits (différentes caractéristiques du régime hydrologique) :

$$r_{ij} = \frac{1}{K}\sum_{k=1}^{K} \omega^k r_{ij}^k$$

où r_{ij}^k est le coefficient de corrélation entre les stations i et j pour la composante de similitude k, ω^k est le poids accordé cette composante, ce qui reflète son importance, et K est le nombre de composantes incluses dans l'indice de similitude pour la paire de stations concernée. En général, les poids seront choisis de sorte que $\sum_{k=1}^{K} \omega^k = 1$. Par exemple, r_{ij}^1 peut représenter la corrélation entre les mesures de débits annuels des stations i et j, r_{ij}^2 peut représenter la corrélation entre les maximums annuels des débits quotidiens aux stations i et j et r_{ij}^3 peut représenter la corrélation entre des mesures de faibles débits annuels aux stations i et j. les r_{ij}^k représentent donc les éléments de la matrice de similarité entre les stations.

Burn et Goulter (1991) choisissent ensuite d'identifier une station pour chaque groupe formé précédemment, qui soit substantiellement différente des autres stations du réseau et qui

représente une source d'information unique ou des caractéristiques uniques. Lorsqu'une station est seule dans sa classe, la décision est triviale. Par contre, la sélection d'une station dans un groupe oblige à tenir compte de plusieurs facteurs. Par exemple, la durée des données, la qualité des données saisies à la station, les utilisateurs et utilisations des données prises à cette station, une mesure de similitude avec les autres stations du groupe et la capacité de chaque station de permettre de prédire l'information aux autres stations peuvent aider à faire le bon choix. Il faut aussi considérer les caractéristiques de toutes les stations retenues comme un tout. Ainsi, il ne serait pas souhaitable d'éliminer toutes les stations dans la même région géographique.

Plusieurs facteurs restent cependant difficiles à inclure dans la méthode. Par exemple, la durée de fonctionnement d'une station et la qualité de ses données s'intègrent difficilement à l'indice de similitude. De la même façon, les préférences des utilisateurs quant au choix des stations ne sont pas facilement quantifiables. Finalement, les relations spatiales entre les stations ne peuvent être considérées dans cette approche. Cela peut mener à l'élimination de toutes les stations d'une région par exemple.

4 Approche bayésienne

L'approche bayésienne permet de modéliser les différentes sources d'incertitude associées aux paramètres hydrologiques. Elle permet aussi de modéliser les paramètres économiques reliés aux coûts et aux bénéfices découlant de l'opération de réseaux hydrographiques. L'approche bayésienne peut donc être un bon outil pour quantifier les bénéfices associés à la poursuite de la collecte de données en une station. Ceci peut nous aider à prendre des décisions à propos des stations à ne plus considérer.

Davis et Dvoranchik, 1971 (conception ou augmentation)

Les grandes étapes de la théorie de la décision sont les suivantes :
1) Définir la décision à prendre et les alternatives possibles
2) Choisir une fonction d'utilité («goal function»)
    a) définir les objectifs
        i) choisir les variables d'états (arguments de la fonction d'utilité)
        ii) développer les propriétés stochastiques de ces variables d'états
    b) Établir la préférence quant au temps
    c) Inclure l'aversion au risque
3) Prendre la décision
    a) Évaluer les connaissances actuelles (calculer les sorties pour les différentes alternatives et trouver les propriétés stochastiques de ces sorties)
    b) Calculer la valeur espérée de la fonction d'utilité pour chaque alternative
    c) Choisir l'alternative qui maximise la valeur espérée de la fonction d'utilité
4) Analyser l'incertitude
    a) Déterminer la perte d'opportunité prévue ou «expected opportunity loss» : XOL (à cause de l'incertitude)
    b) Évaluer les ajouts d'informations
        i) Déterminer la réduction espérée de la perte d'opportunité
        ii) Déterminer les coûts liés à l'obtention d'informations additionnelles

La valeur des informations additionnelles est la réduction espérée de la valeur espérée de la perte d'opportunité, moins les coûts de l'obtention de ces informations.

Moss et Dawdy, 1973 (conception ou augmentation)

Dans le contexte de la conception de réseaux, l'utilité d'un jeu de données est définie comme la différence de bénéfices nets entre les réseaux incluant et excluant ces données. On mesure donc l'utilité des données par les bénéfices perdus si ces données sont manquantes. Chaque site de saisie de données a une utilité égale à l'amélioration du réseau en incluant ces données. Les données qui ne sont pas utilisées ont une utilité maximale de zéro. On ne peut que faire des calculs de probabilités sur l'utilité des données. La valeur espérée de l'utilité d'une série

de données peut être utilisée comme paramètre pour déterminer s'il est préférable d'ajouter cette série de données au réseau ou de le conserver tel quel.

La méthode consiste à définir la valeur espérée de l'utilité des données d'un site sous la condition des paramètres statistiques que l'on connaît par simulation. On conditionne donc en utilisant la distribution a priori des paramètres, qui est obtenue par une approche subjective.

On peut aussi définir l'utilité des données comme l'amélioration espérée de la conception et les bénéfices perdus à cause du manque d'information provenant des données serait alors une mesure de l'utilité des données.

Dawdy, Kubik et Close 1970 (conception)

Dans une étude pilote, des mesures ont été prises sur les débits à partir desquelles une base de données simulées sur une période de 500 ans a été développée. On utilise ces données simulées pour déterminer la conception optimale. Un réseau optimal devrait avoir une utilité maximale pour un budget donné ou avoir une conception pour laquelle l'utilité marginale d'une série de données est égale à son coût marginal. C'est pourquoi une fonction hypothétique de manques («shortages») a été construite pour associer la courbe de coût et les caractéristiques hydrologiques d'un bassin. On aura donc un bénéfice marginal égal au coût marginal pour la taille optimale du réservoir. Le taux de changement des bénéfices avec la taille du réservoir est égal au taux de changement du coût avec la taille du réservoir, lorsque la taille du réservoir est optimale. Le taux de changement du coût est obtenu directement de la courbe de coûts du réservoir.

Dans les travaux de Dawdy (1979) et de Davis et al. (1972), l'approche bayésienne est également adoptée. Pour ces travaux, la théorie de la décision est utilisée dans le cas de bassins montagneux.

5 Estimation non paramétrique de la covariance spatiale et krigeage

5.1. Estimation non paramétrique de la covariance spatiale

Guttorp, Sampson et Newman, 1992 (rationalisation ou extension de réseaux)

Dans cet article, les auteurs visent à déterminer quelles stations d'un réseau peuvent être enlevées tout en maintenant le plus d'information possible disponible, ou à quels endroits des stations peuvent être ajoutées pour maximiser l'ajout d'information. Pour ce faire, ils utilisent l'approche de Caselton-Zidek et une estimation de la covariance basée sur les travaux de Sampson-Guttrop.

Voici d'abord le critère de Caselton-Zidek. On dénote Z, un ensemble de quantités mesurables aux i sites différents. Z se décompose en deux, soient les G sites jaugés $G=\{Z_i, i \quad D\}$ et les U sites non jaugés. Le choix de D sera fait de façon à maximiser l'information dans les G sites à propos des U sites. Encore ici, il s'agit de l'information de Shannon : $I(U,G)=E[\log(f(U|G) / f(U))]$. Un cas particulier est le cas où Z est normale multivariée . Dans ce cas, $I(U,G)=-1/2\ [\log |I - R|]$, où I est la matrice identité et R une matrice diagonale contenant les carrés des coefficients de corrélation canonique entre U et G. Ceux-ci peuvent être obtenus par les estimés des covariances en diagonalisant la matrice :

$$\Sigma_{UU}^{-1/2}\ \Sigma_{UG}\ \Sigma_{GG}\ \Sigma_{GU}\ \Sigma_{UU}^{-1/2}$$

où Σ_{xy} est la matrice de covariances entre x et y. Deux scénarios peuvent se présenter lorsqu'on veut éliminer une station d'un réseau. Premièrement, si le réseau est un réseau pilote situé dans une région où aucune mesure n'avait été prise avant ce réseau, on enlève simplement le site pour lequel l'information sur les stations restantes est la plus élevée, c'est-à-dire, la station la mieux prédite par les autres. Deuxièmement, si on s'intéresse à faire de l'inférence sur un réseau plus dense contenant n stations, on doit alors maximiser l'information provenant des n - 1 sites du réseau actuel, soient les stations G, à propos des autres sites de la même région contenus dans un réseau élargi. Ces sites sont vus comme les stations U.

L'ajout d'une station dans un réseau se fera de façon similaire. En effet, si il n'y a pas d'autre alternative que le réseau considéré, le nouveau site à considérer devrait être à un endroit où il y a peu d'information, celle-ci étant calculée en considérant toutes les stations du réseau comme des stations G et le nouveau site comme une unique station U. Dans le cas où les données d'un réseau alternatif sont disponibles et où l'on veut remplacer un site du réseau actuel par un site du réseau alternatif, on choisit la combinaison qui permet de maximiser l'information fournie par le réseau actuel avec la nouvelle station, à propos des stations restantes dans l'autre réseau. On calcule donc l'information pour toutes les combinaisons possibles.

La méthode utilise une estimation non paramétrique de la covariance spatiale. Pour ce faire, on considère x_1, x_2, ..., x_n les emplacements des stations et (z_1, z_2, ..., z_n) les séries chronologiques observées à chaque station. On suppose que les z_i sont des observations d'un champ aléatoire $Z(x)$. On considère la dispersion spatiale, ou le variogramme:

$$V^2(x,y) = \mathrm{Var}(Z(x) - Z(y)) = c(x,x) + c(y,y) - 2c(x,y)$$

Pour estimer $V^2(x,y)$, il faut transformer la carte géographique des stations (le «G-plane») en une carte, probablement en plus de deux dimensions, où la dispersion augmente de façon monotone avec la distance (la D-image) La configuration des stations de la D-image se fait en choisissant les distances entre les points h_{ij} qui minimisent :

$$\min_{\delta} \frac{\sum_{i<j} (\delta(d_{ij}^2) - h_{ij})^2}{\sum_{i<j} h_{ij}^2}$$

où le minimum est pris sur toutes les fonctions δ monotones, de façon à ce que $\delta(d_{ij}^2)$ représente une régression des h_{ij} sur les d_{ij}^2. Un variogramme Gaussien est alors ajusté au graphique des h_{ij} par rapport aux d^2_{ij}. Ensuite, il faut utiliser des fonctions de lissage basées sur des splines, pour estimer la dispersion entre deux points. Les fonctions de lissage et le variogramme nous donnent la structure de dispersion spatiale du réseau, et permettent d'estimer la dispersion entre deux points (stations déjà existantes ou non). Il est alors possible de vérifier l'effet sur la quantité d'information de l'ajout ou du retrait d'une ou plusieurs stations du réseau, et ainsi de l'optimiser.

Switzer, 1979 (rationalisation de réseaux)

Cet article présente aussi une technique d'estimation de la variance spatiale. Le variogramme ($\gamma(x',x'')$) est défini comme:

$$\gamma(x',x'') = \frac{1}{2} E[\in(x';t) - \in(x'',t)]^2$$

où x' et x" sont deux points dans le bassin de drainage au temps t, et où E, l'opérateur d'espérance, dépend du processus générateur aléatoire. En pratique, il est possible d'estimer le variogramme à partir des données du réseau de stations. Pour ce faire, il faut travailler sur les résidus de la régression entre la variable d'intérêt et la variable de contrôle pour estimer les paramètres c_1 et c_2 de l'équation suivante:

$$\begin{aligned} 2\gamma(x',x'') &= E\{\in(x';t) - \in(x'',t)\}^2 \\ &= [c_1 + \|a' - a''\| c_2 \cdot \|x' - x''\|] \end{aligned}$$

où $\|x'-x''\|$ est la distance spatiale entre les stations et $\|a'-a''\|$ est la différence de la covariance de la variable d'intérêt. Les résidus éloignés spatialement ont une grande chance d'être différents, et les résidus qui correspondent à de grandes différences dans la variable d'intérêt risquent d'être très différents, augmentant ainsi le RMSE. La suppression d'une station d'un réseau consiste à enlever la station pour laquelle on aura la plus faible augmentation du RMSE. Par contre, l'ajout d'une station est un problème plus complexe, car en général, on n'ajoute pas une station pour faire baisser le RMSE, mais pour obtenir des informations spécifiques à propos d'un site particulier. Les auteurs utilisent une carte sur laquelle il y a une erreur et la magnitude de celle-ci pour chaque site. Ils utilisent l'une ou l'autre des mesures suivantes :

- «root mean squared interpolation error averaged over the basin ($RMSE_A$)»
- «maximum RMSE over the basin ($RMSE_M$)»

pour estimer quelle station ajouter ou éliminer. Un des problèmes de la technique réside dans le fait que les cartes interpolées montrent souvent moins de variabilité et de valeurs extrêmes que la réalité. Ce problème peut être en partie minimisé, où à tout le moins vérifiée, en interpolant les valeurs des stations à partir du modèle comme si elles étaient inconnues, à titre de contrôle, selon Switzer.

5.2. Krigeage

Le krigeage est une technique géostatistique d'un intérêt particulier pour la conception des réseaux de mesure, et pour l'estimation des valeurs des variables hydro-météorologiques dans les sites où on ne dispose pas de station de mesure. La technique est reliée aux techniques présentées en 5.1 par l'utilisation du variogramme comme outil de base de l'estimation. Le second potentiel offert par le krigeage est l'estimation de la réduction de variance qu'offre l'ajout de stations fictives dans les régions de grande variance d'estimation des variables.

Villeneuve et al., 1979 (conception ou extension de réseaux)

La technique du krigeage a été appliquée par une équipe de chercheurs de l'INRS-Eau dans l'optimisation du réseau hydrométrique du Québec. On montre que le krigeage est une technique d'estimation optimum en termes de variance minimum, et qu'elle permet d'estimer la variance dans tout point du domaine krigé. Les auteurs évaluent ultimement la variable d'écoulement q au point x_o de la façon suivante:

$$q(x_o)=\sum_{i=1}^{n}\lambda_i\left[q(x_i)+\varepsilon(x_i)+\varepsilon'(x_i)\right]$$

où les n x_i représentent les stations existantes du réseau, les λ_i sont les poids optimaux recherchés et associés à chacune des n stations i, ϵ est l'erreur de mesure, et ϵ' est l'erreur d'échantillonnage temporel. Selon la théorie classique du krigeage, la variance de l'estimation est:

$$\mathrm{Var}\left[q_o^* - q(x_o)\right] = Cov(0) + \mu - \sum_{i=1}^{n} \lambda_i \cdot Cov(x_i - x_o)$$

où Cov est une fonction de covariance (variogramme), et μ est le multiplicateur de Lagrange.

Le krigeage permet d'estimer la valeur optimale des poids λ qui minimisent la variance, avec:

$$\sum_{i=1}^{n} \lambda_i = 1$$

et

$$\sum_{j=1}^{n} \lambda_j \left[Cov(x_i - x_j) + Cov_{ij}'\right] + \lambda_i \sigma_i^2 = Cov(x_i - x_o) + \mu \qquad \text{pour i} = 1,...,\text{n}$$

où σ^2 est la variance. Il est alors possible d'identifier les stations dont l'élimination ne causerait pas une augmentation trop importante de la variance de l'estimation de $q(x_o)$ pour le reste du domaine, si le but de l'exercice est la rationalisation du réseau. Si on vise plutôt à augmenter le réseau, il est possible de trouver les endroits où l'ajout de stations causerait une baisse de cette même variance. Le fait que la fonction de covariance spatiale (Cov(h)) ainsi que les variances d'erreur (ε et ε') ne soient pas nécessairement connues peut cependant causer une sous-estimation de la variance de la fonction de krigeage, selon Villeneuve et al., et limiter en partie l'applicabilité de la méthode.

Pardo-Igúzquiza 1998 (conception, optimisation, ou rationalisation de réseaux)

Cet auteur utilise le krigeage afin d'établir un réseau optimal de pluviomètres. En tant qu'outil de minimisation, Pardo-Igúzquiza utilise l'approche de "l'annealing", par analogie avec la métallurgie. Par contre, le fait que le problème d'optimisation ne soit pas linéaire, ne possède pas de solution analytique, et se caractérise par de nombreux minima locaux complique l'analyse. L'utilisation d'un algorithme basé sur une chaîne de Markov permet de contourner ces problèmes et de faire converger les solutions vers des minima de variance spatiale, permettant ainsi de décider de l'emplacement optimal des pluviomètres. Une fonction de coûts étant également introduite dans le problème, cet article sera examiné en détail dans la section 8, qui traite de la valeur économique des données.

Huang et Yang 1998 (rationalisation potentielle)

L'utilisation du krigeage par ces auteurs diffère des exemple précédents, car il n'est pas question de gérer un réseau. Par contre, leur approche est d'intérêt pour le problème qui nous concerne. En utilisant le krigeage simple, les auteurs montrent comment il est possible d'évaluer la variation spatiale dans la contribution aux débits sur un bassin-versant. Par extension, il serait possible d'éliminer théoriquement une ou plusieurs stations de jaugeage qui ont été réellement utilisées et de vérifier comment les estimations de débits changent dans l'espace, ce qui pourrait

être utile dans l'optique de la rationalisation d'un réseau afin d'évaluer la performance potentielle du réseau rationalisé.

6 Analyse des corrélations entre les stations

Ouarda, et al., 1997 (rationalisation)

Considérons deux stations de jaugeage voisines, probablement situées sur la même rivière. Ces stations sont donc soumises aux mêmes conditions climatiques et météorologiques. On peut alors supposer qu'il y a une corrélation entre les données recueillies à ces deux stations. Avant de faire une analyse de corrélation, il faut normaliser les données. La normalisation, souvent une transformation logarithmique permet d'avoir une relation linéaire entre deux variables. Supposons que la station Y a n_1 années d'enregistrement de données et que la station X en a n_1+n_2. Ceci peut être représenté de la façon suivante :

Site X : $x_1, x_2, \ldots, x_{n_1}, x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}$

Site Y : $y_1, y_2, \ldots, y_{n_1}$

Pour estimer les débits au site Y dans les années n_2, on procède par régression linéaire :

$$y_i = \alpha + \beta x_i$$

Si on s'intéresse à la moyenne de la variable Y, alors la valeur moyenne de la série pour les n_1+n_2 années est :

$$\hat{\mu}_y = \overline{y}_1 + \frac{n_2}{n_1+n_2}\hat{\beta}\left(\overline{x}_2 - \overline{x}_1\right)$$

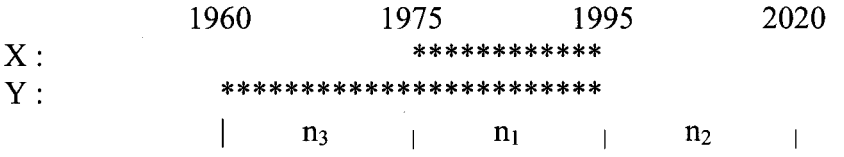
où $\overline{y}_1$ est la moyenne des y_i observés durant la période n_1 et $\overline{x}_1$ et $\overline{x}_2$ sont les moyennes des x_i observés aux périodes n_1 et n_2 respectivement. La variance de cet estimateur est :

$$\mathrm{Var}\left[\hat{\mu}\right] = \frac{\sigma_y^2}{n_1}\left[1 - \frac{n_2}{n_1+n_2}\left(\rho^2 - \frac{1-\rho^2}{n_1-3}\right)\right]$$

où σ_y^2 est la variance de Y et ρ est le coefficient de corrélation entre X et Y. On peut montrer qu'il y a amélioration de l'estimation, par rapport à celle calculée seulement à partir des années n_1, si $\rho > \left(\left|n_1 - 2\right|\right)^{-1/2}$. Donc, pour un site donné, on devrait utiliser la station du réseau qui conduit à une variance minimale pour l'estimateur de la moyenne. En général, cette station doit être fortement corrélée avec celle possédant plusieurs années d'enregistrement de données.

La rationalisation de réseaux, d'un point de vue strictement statistique, consisterait à éliminer la station la plus fortement corrélée avec les autres stations du réseau et à estimer les données des années futures à l'aide de la régression, tel que décrit précédemment. Mais, on ne

connaît pas les données futures et c'est ce qui nous intéresse. Il faut donc modifier un peu la méthode décrite précédemment. Voici un exemple de situation que l'on peut avoir :



Si la station X est abandonnée, alors on peut utiliser la formule de variance décrite précédemment en considérant n_2+n_3 comme période d'extension. Par contre, si Y est éliminée, il faut modifier la formule de variance de la façon suivante :

$$\mathrm{Var}\left[\hat{\mu}\right]=\frac{\sigma_y^2}{n_1}\left[1-\frac{n_3\left(n_1+2n_2+n_3\right)}{\left(n_1+n_2+n_3\right)^2}-\frac{n_2\left(n_1+n_2\right)}{\left(n_1+n_2+n_3\right)^2}\left(\rho^2-\frac{1-\rho^2}{n_1-3}\right)\right]$$

Supposons que, pour des raisons budgétaires, on veuille garder seulement k des m stations d'un réseau. Le nombre de combinaisons possibles de stations à abandonner est C(m,k). Pour chaque combinaison, on calcule un indice d'information et on ordonne les combinaisons. Il faut d'abord choisir un horizon, c'est-à-dire la valeur de n_2. Pour ce faire, il faudrait considérer plusieurs horizons et observer la sensibilité de la décision optimale. Ensuite, on choisit un indice de performance qui reflète la quantité d'information qui sera disponible après n_2 années. On peut, par exemple, définir un indice global d'incertitude de la façon suivante :

$$U_g(Q)=\sum_{\text{réseau}}\sqrt{\mathrm{Var}\left[\hat{\mu}_i(Q)\right]}$$

où Q est la variable de base pour la rationalisation et $\hat{\mu}$ est l'estimateur de la moyenne. Pour toutes les stations qui seront gardées, le meilleur estimateur de la moyenne sera basé sur n_2 années, tandis que pour les stations qui seront enlevées, la variance sera calculée à partir de l'équation décrite plus haut. Pour chacune des k stations enlevées, la meilleure station auxiliaire pour faire l'interpolation des données se trouve dans les m - k stations restantes du réseau. Après avoir examiné toutes les combinaisons possibles de stations à supprimer, on choisit celle pour laquelle $U_g(Q)$ est minimal.

Cette méthode est appropriée si on considère des variables comme le débit annuel, mais n'est pas très indiquée pour des variables comme des débits quotidiens, car il faut alors tenir compte des saisons. Pour faire la rationalisation d'un réseau, les paires de stations fortement corrélées entre elles doivent être identifiées et une d'elles doit être éliminée. À partir des données de l'autre station, on trouve alors une fonction de lien, qui permettra d'estimer les données à la station supprimée. Il a donc été suggéré (Ouarda et *al.* , 1997) d'utiliser les modèles de fonctions de transfert pour établir la relation entre les stations.

7 Autres approches

7.1. Optimisation de la planification de réseaux sur la base de la multi-régionalisation

Solomon, 1972 (conception ou extension de réseaux)

La multi-régionalisation conduit à une méthode pour déterminer l'allocation optimale des fonds disponibles pour un réseau entre les nouvelles stations, les coûts d'opérations, etc. En général, les trois composantes principales de l'erreur de l'estimation du débit moyen à une section de la rivière peuvent être définies comme :

E_1 : Erreur due au transfert de données
E_2 : Erreur due à la période d'échantillonnage
E_3 : Erreur due aux mesures
E : Erreur dans le réseau

On a donc : $E^2 = E_1^2 + E_2^2 + E_3^2$

Si C_0 représente les coûts annuels constants, C_1 est le coût d'installation de nouvelles stations, C_2 est le coût d'opération du réseau (incluant le coût d'opération des stations, l'inflation, et le pourcentage de stations éliminées, gr) et C_3 est le coût des mesures et des nouveaux instruments.

On a alors :

$$\Delta C = C_0 + \Delta C_1 + \Delta C_2 + \Delta C_3$$

où ΔC est le coût d'opération du réseau pour un horizon de temps donné. Avec l'ajout de nouvelles stations, l'erreur sera réduite dans le réseau à l'horizon de temps selon :

$$\Delta E = \sqrt{(H_1 - K_1 \log N_1) + \frac{H_2^2}{N_2} + \frac{H_3^2}{N_3}}$$

où K_1, H_1, H_2, et H_3 sont des paramètres hydrologiques empiriques, N_1 est le nombre de stations, N_2 est le nombre d'année jusqu'à l'horizon, et N_3 est le nombre d'observations par station. Le problème d'optimisation consiste à minimiser, à l'aide de techniques numériques, la fonction reliant ΔC et ΔE :

$$\frac{\Delta C}{\Delta E} = f(N_1, N_2, N_3, K, E_1, C_0, C_1, C_2, C_3, gr)$$

7.2. Procédures d'extension des séries de données

Plusieurs méthodes existent pour l'extension des séries de données dans une station de mesure à partir des observations dans une autre station possédant une série de données plus complète. La majorité de ces méthodes représentent différentes variantes de l'approche d'analyse de corrélations ou de la régression. Plusieurs articles traitent de ce sujet (par exemple, Matalas et Jacobs, 1964; Matalas et al., 1976; Hirsch, 1979; Alley et Burns, 1983; Vogel et Stedinger, 1985; et Grygier et Stedinger, 1989). Ces méthodes peuvent être utilisées dans le cadre de techniques de rationalisation de réseaux (tel que Ouarda et al., 1997), puisqu'elles nous permettent de prédire les observations à des sites désormais exclus du réseau. L'étude des méthodes d'extension des séries de données reste à l'extérieur de l'objectif du présent travail.

8 Valeur économique des données

La gestion efficace d'un réseau de cueillette de données ne fait pas face seulement à des problèmes techniques et statistiques: il y a aussi d'importantes contraintes économiques liées au maintien ou à l'installation des stations de jaugeage. Le problème est généralement de savoir si les gains économiques résultants de l'installation et de l'exploitation à une station dépassent les coûts d'installation et de maintient de celle-ci. La difficulté principale réside dans l'estimation de la valeur monétaire des données hydrologiques pour différentes périodes de temps. La littérature sur la valeur économique des données hydrologiques n'est pas très développée; par contre, il existe des approches prometteuses qui seront décrites ici.

8.1 Analyse de coûts/bénéfices

Griffin, 1998

Dans cet article, Griffin introduit l'analyse de coûts/bénéfices pour des secteurs reliés aux ressources hydriques. Cette approche permet d'évaluer d'une façon générale si les bénéfices reliés à une construction ou une utilisation d'une ressource hydrique seront supérieurs aux coûts envisagés, et ce pour divers horizons temporels. L'approche repose sur l'utilisation de deux indices, le rapport bénéfices/coûts (BCR), et la valeur nette actuelle (NPV). Ces indices incluent une grande variété de facteurs économiques, comme l'inflation, les taux d'intérêts, ou tout autre facteur d'intérêt pour l'étude. Les indices BCR et NPV sont de plus considérés comme équivalents, sauf en cas de circonstances claires qui indiqueraient le contraire. Le BCR est :

$$BCR = \left(\sum_{t=0}^{T} \frac{B_t}{(1+d)^t} \right) \left(\sum_{t=0}^{T} \frac{C_t}{(1+d)^t} \right)^{-1}$$

où le temps t = 0,...,T; B est le bénéfice annuel total, C est le coût annuel total, et d est le taux d'intérêt en format décimal. D'une façon similaire, le NPV se calcule selon :

$$NPV = \sum_{t=0}^{T} \frac{B_t - C_t}{(1+d)^t}$$

On estime qu'un projet (par exemple, la construction d'un barrage) sera profitable si BCR>1, ou si NPV>0. Un désavantage de la méthode, outre sa grande généralité, est la difficulté d'estimer B et C avec la moindre précision. Si C est plus facilement prévisible, B peut être très difficile à estimer. Il faut de plus respecter cinq grands principes économiques qui ne s'appliquent pas toujours aux projets d'intérêt dans le cadre de ce rapport afin d'appliquer cette technique. S'il est possible de comparer différents projets entre eux, aucun des indices ne permet

de vérifier l'adéquation des échelles spatiales des projets étudiés. Finalement, il semble que la technique soit peu adaptée aux projets à très long terme (plus de quelques décennies).

8.2 Coûts des réseaux de pluviomètres

Bien que le but de ces études soit différent de ce qui nous préoccupe dans ce rapport, leurs approches sont intéressantes et méritent que l'on s'y attarde quelque peu.

Andricevic, 1990 (conception de réseaux)

Dans la gestion d'un réseau de pluviomètres, il est utile de pouvoir maximiser la puissance statistique d'un plan d'échantillonnage pour un budget donné, et de pouvoir minimiser les coûts tout en obtenant la puissance statistique requise. Ces contraintes sont par contre très difficiles à respecter. La méthode proposée pas Andricevic vise à balancer les exigences économiques et statistiques. Sous forme mathématique, on obtient :

$$\text{minimiser} \quad T_c = F_c + \sum_{i=1}^{nw} a(i) + \sum_{i=1}^{nw} \beta(i)\, fr(i)$$

de façon à ce que :

$$Tr\left[Cov(N)\right]^{-1} \geq \sum_{i=1}^{M_{alt}} (IRT)_i$$

où

$$IRT = \frac{1}{\bar{\sigma}_{ik}^2}$$

où T_c est le coût total d'échantillonnage, F_c est un coût fixe, a est le coût d'installation d'une station, β est le coût unitaire d'échantillonnage, fr est la fréquence d'échantillonnage, Cov(N) est la covariance du niveau d'eau souterraine à la fin de l'horizon, IRT est un seuil de fiabilité de l'information, M est le nombre de nœuds dans la simulation, et $\bar{\sigma}^2$ est la variance tolérée au temps k. Il faut effectuer les calculs pour chaque alternative (alt), et pour chaque puits i jusqu'au nombre total de puits nw. Il suffit alors de tester différents nombres de puits et fréquences d'échantillonnage afin d'identifier l'optimum, tout en forçant l'algorithme à éviter nw=1 et fr=1. Un problème de l'approche est qu'il est difficile d'obtenir la localisation optimale des puits. Seul le calcul de Cov(N) permet de le faire, et ce si l'on utilise un algorithme qui minimisera cette valeur, tout en considérant le fait que les niveaux reliés à chacun des puits n'est pas connu avec certitude pour l'horizon temporel, car il s'agit d'une prévision.

Pardo-Igúzquiza 1998 (conception, optimisation, ou rationalisation de réseaux)

Tel que présenté à la section 5.2, Pardo-Igúzquiza utilise les géostatistiques afin d'optimiser un réseau de pluviomètres. Au contraire des autres techniques de krigeage ou des travaux présentés dans cette section, l'auteur utilise à la fois de l'information économique et spatiale afin d'optimiser un réseau. Le but est de trouver une distribution spatiale de pluviomètres qui minimise la variance dans l'estimation de la variable d'intérêt et le coût du réseau. Pour ce faire, il faut minimiser la fonction objective OF:

$$OF = \delta(N, x_i) + C\Delta.C(N, x_i)$$

où N est le nombre de stations à i=1,...N sites x_i, $C\Delta$ est une mesure de précision équivalente à un changement de coût unitaire, $C(N,x_i)$ la fonction de coût des pluviomètres, et $\delta(N,x_i)$ le variogramme qui dépends de la variance. $C(N,x_i)$ peut être approximée par :

$$C(N, x_i) = \sum_{i=1}^{N} M(x_i)$$

pour le coût M à chaque site. Le variogramme peut être estimé selon la théorie du krigeage par :

$$V_E^2 = \mu + \sum_{i=1}^{N} \lambda_i \bar{g}(h_{iA}) - \bar{g}(h_{AA})$$

où $\bar{g}(h_{iA})$ est le variogramme de la variable d'intérêt pour la région A, et $\bar{g}(h_{AA})$ est le variogramme quand la valeur extrême du vecteur h décrit indépendamment la région A. La démarche consiste à choisir un réseau existant, et à vérifier ce qui arrive à la fonction objective lorsqu'une station est ajoutée ou retirée, selon le but de l'étude. Le réseau optimal, qui est représenté par $_{min}OF$, comprend un nombre et une disposition des stations qui maximise la précision de l'échantillonnage tout en minimisant les coûts d'opération. La méthode permet aussi d'estimer les critères ci-haut pour tout design de réseau qui peut être souhaité, à partir d'un ensemble de stations non-existantes par exemple.

8.3 Coûts et valeur des débits

Adeloye 1996 (conception ou rationalisation de réseaux)

Un regret (Opportunity Loss, OL) est une pénalité économique pour ne pas avoir d'information parfaite sur le processus à l'étude. Par exemple, une incertitude sur le débit d'une rivière peut entraîner des coûts importants si un ouvrage est construit à partir de ces données. La technique ici proposée vise à identifier les bénéfices économiques liés à une période de jaugeage pour un site donné, en minimisant OL. Adeloye utilise la planification de réservoirs comme exemple illustratif de la technique de l'OL. Cette approche combine les bénéfices de l'approche bayésienne en considérant la nature stochastique du processus étudié, et surmonte le problème des données futures en créant des séries temporelles à l'aide d'une simulation de type Monte-Carlo. Le bénéfice brut espéré (B) d'un aménagement est, après avoir discrétisé la variable D (débit utilisé) par rapport à la capacité (Cap), la valeur maximale de la fonction suivante:

$$E[B \mid Cap] = \sum_{D_i} B(D_i) \Pr(D \mid Cap)$$

Cette information est ensuite utilisée pour évaluer OL. Le regret est évalué comme étant la différence entre la valeur présente des bénéfices attendus et la valeur probable des bénéfices attendus si un échantillonnage supplémentaire des débits est effectué jusqu'à l'horizon temporel T:

$$OL = E[B \mid Cap]_o - E[B \mid Cap]_T$$

C'est à cette étape que les simulations de séries futures sont effectuées, afin d'estimer le regret attendu (XOL) :

$$XOL = \int_e OL(e)\, f(e) de$$

où e représente l'erreur associée au jaugeage et à l'estimation des phénomènes hydrologiques étudiés. Il faut alors tester différents niveaux de données (longueurs de séries, fréquences d'échantillonnage, précision des équipements et techniques, etc.) afin de minimiser XOL, ce qui identifiera alors le design optimal. La valeur de Y données supplémentaires est alors facilement estimée comme étant :

$$V(Y) = XOL\,(0) - XOL\,(Y)$$

C'est à dire que la valeur de données supplémentaires sera une réduction de la pénalité économique attendue étant donnée l'incertitude des processus étudiés. La difficulté de l'approche réside dans l'évaluation de B qui ne peut être plus précise que les prévisions économiques, et de façon moindre, de la distribution de probabilité de D. De plus l'auteur présume que toutes les installations physiques sont présentes au temps 0 (pas de construction nécessaire), que les coûts d'opérations sont négligeables, et que les bénéfices seront constants dans le temps et toujours égaux aux valeurs attendues.

Fortin 1997 (conception, augmentation, ou rationalisation de réseaux)

Pour évaluer le bénéfice économique d'une observation supplémentaire à une station (ou d'une série d'observations), il faut prévoir sa valeur future. La théorie bayésienne de la décision peut être utilisée à cette fin (voir section 4 pour la théorie bayésienne), mais il est souvent très difficile de représenter l'information a priori. Afin de simplifier ce problème, Fortin utilise la théorie des prévisions inférieures cohérentes pour obtenir ces informations a priori. Un exemple d'utilisation de l'approche est présenté pour le cas de la rivière Ste-Marguerite (SM), au site proposé pour la construction de SM3. Tel que remarqué par l'auteur, il faut se souvenir que cet exemple est un exercice théorique, puisqu'il a été effectué après que la décision de mettre SM3 en chantier fut prise. L'approche préconisée par l'auteur est très flexible; elle sert à juger de la valeur économique future d'une station. Elle permet donc de décider si l'ajout ou le retrait d'une station serait bénéfique ou coûteux à l'intérieur d'un réseau. Nous croyons que cette approche

peut être adaptée au design d'un réseau, en traitant chaque station existante ou proposée individuellement et en voyant si elle est rentable.

Au site SM3, l'information disponible a priori permet d'estimer le débit annuel moyen naturel (x) à une valeur comprise entre 46 et 139m^3/s ($\underline{E}(x) = 46\,m^3/s$; $\overline{E}(x) = 139\,m^3/s$), et il est possible d'espérer que le module naturel type soit de 109 m^3/s (E(x) = 109 m^3/s). Le problème est de savoir s'il est rentable de construire SM3 par rapport à d'autres projets alternatifs, sachant que le module de SM3 n'est pas connu avec certitude, et que le module total va dépendre de la probabilité θ de détourner une partie du débit de la rivière Moisie. Il est aussi possible d'évaluer le coût relié à l'obtention d'une observation du module à SM3, si l'on choisissait d'attendre avant de prendre un décision, en espérant qu'une (ou plusieurs) observation du module permette de prendre une meilleure décision.

Un premier problème est de prévoir θ. Une analyse du risque a priori permet de poser les conditions suivantes, connaissant les coûts de production hydroélectrique :

$\underline{E}[\theta] = 0.6$ pour préférer SM3 au projet alternatif 1 (Grande Baleine, GB)

$\underline{E}[\theta] = 0.4$ pour préférer SM3 au projet alternatif 2 (Nottaway-Broadback-Rupert,NBR)

Les coûts pour SM3 ont été estimés à l'aide d'une régression linéaire simple entre le prix de revient prévu et le module : $g(SM3, Q, Q_M, \theta)=7.4-0.0224(Q+Q_M .\theta)$, où Q est le module naturel à SM3, et Q_M est le module détourné de la rivière Moisie. θ est estimé a priori à l'aide d'une probabilité imprécise $[E(\theta),1]$. Pour SM3, nous obtenons alors le risque imprécis $[4.0, 5.0-\underline{E}(\theta)]$, qui est utilisé pour évaluer le risque de l'action optimale a priori $R(\hat{a})=\min\{c, 5.0-p\}$, où c est le coût de revient du projet alternatif et p est la probabilité de préférer SM3.

L'auteur introduit alors le regret OL pour évaluer la rentabilité de SM3 par rapport à d'autres projets. Pour un ensemble d'actions S, le regret d'une décision $\delta_0 \in S$ est donné par :

$$OL[\delta_0, Q, p] = g(\delta_0, Q, p) - \inf_{a \in S} g(a, Q, p) = \sup_{a \in S}[g(\delta_0, Q, p) - g(a, Q, p)]$$

SM3 sera préféré à d'autres projets alternatifs si, et seulement si $g(SM3,Q,p) = 7.4-0.0224.Q-p<c$.

Cela nous donne alors :

$OL(SM3,Q,p)=\max[0; 7.4-0.0224.Q-p]$ et $OL(alt, Q, p)=\max[0; c-7.4+0.0224.Q+p]$

Si p est connue, l'espérance de ces dernières équations donne la prévision précise $XOL^P(\delta_0)$, qui sera la valeur maximale de l'information parfaite :

$$\overline{XOL} = \inf_{\delta_0 \in S} \overline{E}[OL(\delta_0, Q, p)$$

Pour le projet SM3 et un projet alternatif PA, l'auteur obtient après manipulations :

$$XOL^P\,(SM3) = (7.4\text{-}c\text{-}p)\,F(c,p) - 0.0224 \int_{46}^{Q(c,p)} Q\,dP$$

et

$$XOL^P\,(PA) = (7.4\text{-}c\text{-}p)\,(F(c,p) - 1) + 0.0224 \int_{Q(c,p)}^{139} Q\,dP$$

XOL^P sera maximale pour des mesures de probabilité sensibles à une information supplémentaire, c'est-à-dire qu'elle sera maximale lorsque recueillir une mesure de débit sera bénéfique économiquement, et fera diminuer le regret. Dans le cas de SM3, et utilisant l'information disponible, la valeur de l'information parfaite sera de :

$$\overline{XOL} = \left\{(6.4 - c - p).\frac{30}{93};\ \ (p + c - 4.3).\frac{63}{93}\right\}$$

Dans le contexte de l'étude de Fortin, cette approche permet d'estimer la valeur d'une information parfaite à une valeur du même ordre que l'écart de rentabilité entre SM3 et les projets alternatifs. Une connaissance parfaite du module à SM3, si on compare le projet à celui de NBR, aura une valeur de 0.2¢/kwh par rapport au prix de production.

Les travaux de Fortin portent aussi sur le prix de l'observation du débit annuel. En utilisant l'information disponible a priori, incluant R(â), retarder la mise en chantier de SM3 (donc obtenir une observation du débit annuel) coûterait {6.2-c, 1.2+p}¢ pour chaque kwh perdu durant l'année. En considérant la production prévue la première année, le coût équivaudrait à max{14-2.2c, 2.6+2.2p} millions de dollars. Avec les PA (NBR, GB), cela équivaut à une somme de 4 à 5 millions de dollars. Il faut cependant se rappeler qu'une connaissance parfaite du module à SM3 permettrait des économies de l'ordre de 0.2¢/kwh, soit de 6 à 8 millions de dollars annuellement. Selon Fortin, la valeur d'une observation supplémentaire du débit à SM3 (obtenue à un coût de 4 à 5 millions $) n'est pas nécessairement de 6 à 8 millions de dollars mais dépend de l'idée que l'on a a priori de sa précision.

Ces dernières approches sont cependant légèrement limitées au niveau spatial, car elles ne considèrent pas la structure et l'organisation spatiale de l'information. De plus, la variance d'estimation spatiale n'est pas incluse dans ces méthodes, ce qui limite leur application lors de l'étude de l'ensemble d'un réseau ou de régionalisation.

9 Bibliographie

Adeloye A.J., 1996, An opportunity loss model for estimating the value of streamflow data for resservoir planning, *Water Resources Management*, v. 10, p. 45-79.

Alley W.M. et Burns A.W., 1983, Mixed-Station Extension of Monthly Streamflow Records, *Journal of Hydrologic Engineering*, v. 109, no. 10, p. 1272-1284.

Andricevic R., 1990, Cost-effective network design for groundwater monitoring, *Stochastic Hydrology and Hydraulics*, p. 27-41.

Bisson J.-L., 1989, Plan d'installation de stations hydrométéorologiques, Hydro-Québec, Service Charges et Ressources, 37p.

Burn D. H. et Goulter I.C., 1991, An Approach to the Rationalization of Streamflow Data Collection Networks, *Journal of Hydrology*, v. 122, p. 71-91.

Caselton W.F. et Husain T., 1980, Hydrologic Network Information Transmission, *Journal of the Water Resources Planning and Management Division*, v. 106, no. 2, p. 503-520.

Davis D.R. et Dvoranchik W.M., 1971, Evaluation of the Worth of Additional Data, *Water Ressources Bulletin*, v. 7, no. 4, p. 700-707.

Davis D.R., Kisiel C.C. et Duckstein L., 1972, Optimum Design of Mountainous Raingauge Networks Using Baysien Decision Theory, *Int. Symp. on Distribution of Precipitation in Mountainous Areas*, Norway, Volume I, WMO- no. 326, p. 108-110.

Davis D.R., Kisiel C.C. et Duckstein L., 1972, Optimum Design of Mountainous Raingauge Networks Using Baysien Decision Theory, *Int. Symp. on Distribution of Precipitation in Mountainous Areas*, Norway, Volume II, WMO- no. 326, p. 416-420.

Dawdy D.R., 1979, The Worth of Hydrologic Data, *Water Ressources Research*, v. 15, no. 6, p. 1726-1732.

Dawdy D.R., Kubik H.E. et Close E.R., 1970, Value of Streamflow Data for Project Design-A Pilot Study, *Water Ressources Research*, v. 6, no. 4, p. 1045-1050.

Fortin V., 1997, *Estimation de la valeur de l'information hydrologique à l'aide de probabilités imprécises*, thèse de doctorat, INRS-Eau, Université du Québec, 158 pages.

Gauthier R. et Roy R., 1988, Évaluation de la qualité des mesures limnimétriques, proposition d'un schéma d'intervention, Hydro-Québec, Division Hydrométéorologie, Prévisions et Planification de l'Exploitation du Parc d'Équipement, 52p.

Griffin R.C., 1998, The fundamental principles of cost-benefit analysis, *Water Resources Research*, v. 34, no. 8, p. 2063-2071.

Grygier C.J. et Stedinger J.R., 1989, A Generalized Maintenance of Variance Extension Procedure for Extending Correlated Series, *Water Ressources research*, v. 25, no. 3, p. 345-349.

Guttorp P., Sampson P.D. et Newman K., 1992, Nonparametric Estimation of Spatial Covariance with Applications to Monitoring Network Evaluation, *Statistics in the Environmental and Earth Sciences* : London, Edward Arnold, p. 39-51.

Hirsch R.M., 1979, An Evaluation of Some Record Reconstruction Techniques, *Water Ressources research*, v. 15, no. 6, p. 1781-1790.

Huang W-C., et Yang F-T., 1998, Streamflow estimation using Kriging, *Water Resources Research*, v.34, no.6, p. 1599-1608.

Husain T., 1987, Hydrologic Network Design Formulation, *Canadian Water Resources Journal*, v. 12, no. 1, p. 44-59.

Hydrological Network Desing and Information Transfer, 1976, Proceedings of the International Seminar Organized by the University of Newcastle and Sponsored by the World Meteorological Organization and the International Association of Hydrological Sciences, Newcastle upon Tyne, U.K., August 19 1974-August 23 1974, p. 103-109.

Matalas N.C., 1973, Optimum gauging station locations, Scientific Computing Symposium on Water and Air Resources Management, Yorktown Heights, N.Y., pp. 85-94.

Matalas N.C. et Jacobs B., 1964, A Correlation Procedure for Augmenting Hydrologic Data, *U.S. Geologic Survey Professional Paper*, 434-E, 7 pp.

Matalas N.C., Todini E. et Wallis J.R., 1976, Statistics of Data Transfer, *Hydrological Network Desing and Information Transfer, 1976, Proceedings of the International Seminar Organized by the University of Newcastle and Sponsored by the World Meteorological Organization and the International Association of Hydrological Sciences*, Newcastle upon Tyne, U.K., August 19 1974-August 23 1974, p. 103-109.

Moss M.E., 1982, Concepts and Techniques in Hydrological Network Design : Geneva, World Meteorological Organization, Report no. 19, p. 12-26.

Moss M.E. et Dawdy D.R., 1973, The Worth of Data in Hydrologic Design, Highw. Res. Rec., no. 479, p. 46-51.

Ouarda T. B.M.J., Rasmussen P.F., Bobée B. et Bernier J., 1997, Towards Sustainable Hydrometeorologic Data Collection Networks, *Practising Sustainable Water Management Canadian and International Experiences*, chapitre 2, p. 26-41.

Pardo-Igúzquiza E., 1998, Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing, *Journal of Hydrology*, v.210, p. 206-220.

Roberge F. et Bisson J.L., 1982, Influence du réseau météorologique sur la prévision des apports naturels à Hydro-Québec, analyse et amélioration possible, Hydro-Québec, Direction Mouvements d'Énergie, Division météorologie, 66p.

Solomon S.I., 1972, Multi-Regionalization and Network Strategy, *Casebook on Hydrological Network Design Practice*, Genève, WMO - no. 324, p. III - 3.3 - 1 à 3.3 - 11.

Switzer P., 1979, Statistical Considerations in Network Design, *Water Ressources Research*, v. 15, no. 6, p. 1712-1715.

Villeneuve J.-P., Morin G., Bobée B. and Leblanc D., 1979, Kriging in the design of streamflow sampling networks, *Water Resources Research*, v. 15, no6, p. 1833-1840.

Vogel R.M. et Stedinger J.R., 1985, Minimum Variance Streamflow Record Augmentation Procedures, *Water Ressources research*, v. 21, no. 5, p. 715-723.

Yang Y. et Burn H., 1994, An Entropy Approach to Data Collection Network Design, *Journal of Hydrology*, v. 157, p. 307-324.