UNIVERSITÉ DU QUÉBEC
INRS-ETE

MÉTHODES STATISTIQUES POUR L'ÉVALUATION ET LA RECONFIGURATION DES RÉSEAUX DE SUIVI DE LA QUALITÉ DE L'EAU DE SURFACE

Par

BAHAA KHALIL

Thèse présentée pour l'obtention du grade de Philosophiae Doctor (Ph.D.) en sciences de l'eau

Jury d'évaluation:

Examinateur externe	René Therrien Université Laval
Examinatrice externe	Geneviève Pelletier Université Laval
Examinatrice interne	Sophie Duchesne INRS-ETE
Membre invité	Shaden Abdel-Gawad Centre National de Recherches sur l'Eau, Le Caire, Égypte
Co-directeur de recherche	André St-Hilaire INRS-ETE
Directeur de recherche	Taha B.M.J. Ouarda INRS-ETE

Thèse présentée le 10 Décembre 2010

Abstract

This study addresses the assessment and redesign of surface water quality monitoring (WQM) networks. The design of WQM networks depends primarily upon the objectives of the monitoring program and the characteristics of the monitored region. Despite several statistical approaches that have been proposed for the assessment and redesign of long-term WQM networks, several deficiencies in these approaches exist.

The main goal of this study is to propose statistical approaches for the assessment and redesign of WQM networks that overcome existent deficiencies in the currently applied approaches. In addition, this study intends to introduce an innovative approach for the estimation of the water quality characteristics at ungauged sites. Four objectives are specified: (i) to review the current applied statistical approaches for the assessment and redesign of surface WQM networks; (ii) to develop a new statistical approach for the rationalization of water quality variables; (iii) to develop a new statistical approach for the assessment and redesign of WQM locations; and (iv) to introduce a statistical methodology for the estimation of water quality characteristics at ungauged sites.

In this study, statistical approaches used for the assessment and redesign of surface water quality monitoring networks are first reviewed. In this review, various monitoring objectives and related procedures used for the assessment and redesign of surface WQM networks are discussed. For each approach, advantages and disadvantages are examined from a network design perspective.

The literature review reveals that correlation-regression is the most common approach used to assess and eventually reduce the number of water quality variables in WQM networks. However, several deficiencies in this approach are identified. Based upon these identified deficiencies, a new statistical approach is proposed for the rationalization of water quality variables. The proposed approach overcomes deficiencies in the conventional correlation-regression approach and represents a useful decision support tool for the optimized selection of water quality variables. It allows for the identification of optimal combinations of water quality variables to be continuously measured and those to be discontinued.

To reconstitute information about discontinued water quality variables, four record extension techniques are examined. Ordinary least squares regression (OLS), the line of organic correlation (LOC), the Kendall-Theil robust line (KTRL) and KTRL2, which is a modified version of the KTRL proposed in this study. The advantage of the KTRL2 is that it includes the advantage of LOC in maintaining variability in the extended records and the advantage of KTRL in being robust in the presence of extreme values. Monte-Carlo and empirical studies are conducted to examine these four techniques for bias, standard error of moment estimates and a full range of percentiles. The Monte-Carlo study showed serious deficiencies in the OLS and KTRL techniques, while the LOC and KTRL2 techniques have results that are nearly similar. Using real water quality records, the KTRL2 is shown to lead to better results than the other techniques.

The literature review also reveals that several deficiencies in the approaches proposed for the assessment of monitoring locations exist. The deficiencies vary from one approach to another, but generally include: (i) ignoring the characteristics of the basin being monitored in the design

approach; (ii) handling multivariate water quality data sequentially rather than simultaneously; (iii) focusing mainly on locations to be discontinued; and (iv) ignoring reconstitution of information at discontinued locations. A methodology that overcomes these deficiencies is proposed. In the proposed methodology, hybrid-cluster analysis is employed to identify groups of sub-basins with similar characteristics. A stratified optimum sampling strategy is then employed to identify the optimum number of monitoring locations in each of the sub-basin groups. An aggregate information index is employed to identify the optimal combination of locations to be discontinued. Results indicate that the proposed methodology allows the identification of optimal combinations of locations to be discontinued, locations to be continuously measured and sub-basins where monitoring locations should be added.

To fulfill the last objective, two models are developed for the estimation of water quality mean values at ungauged sites. An ensemble artificial neural network (EANN) model is developed to establish the functional relationship between water quality mean values and basin attributes. The second model is based on canonical correlation analysis (CCA) and EANN. CCA is used to form canonical attributes space using data from gauged sites. Then, an EANN is applied to identify the functional relationships between water quality mean values and the attributes in the CCA space. A jackknife validation procedure is used to evaluate the performance of the two models. The results show that the developed models are useful for estimating the water quality status at ungauged sites. However, the CCA-based EANN model performed better than the EANN model in terms of prediction accuracy.

Foreword

This thesis presents the research conducted during my doctoral studies. The structure of this thesis follows the standard structure of INRS-ETE theses. The first part of the thesis includes a general summary of the work performed. The summary aims to review succinctly the main results obtained and discuss their significance. The second part of the thesis contains five articles, published (2) or submitted (3).

Articles and authors contribution

[1].Khalil, B. and T.B.M.J. Ouarda (2009). Statistical approaches used to assess and redesign surface water quality monitoring networks. *Journal of Environmental Monitoring*, 11, 1915 - 1929

[2].Khalil, B., T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana (2010). A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks. *Journal of Hydrology*, 386, 173-185.

[3].Khalil, B., T.B.M.J. Ouarda, and A. St-Hilaire (submitted). Comparison of record-extension techniques for water quality variables, *Journal of Hydrology*.

[4].Khalil, B., T.B.M.J. Ouarda, and A. St-Hilaire (submitted). A statistical approach for the assessment and redesign of the Nile Delta drainage water-quality-monitoring locations, to be submitted, *Journal of Environmental Monitoring*.

[5].Khalil, B., T.B.M.J. Ouarda, and A. St-Hilaire (in press). Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis, *Journal of Hydrology*.

In the first article, B. Khalil performed a literature review on statistical approaches used to assess and redesign surface water quality monitoring networks. T.B.M.J. Ouarda assisted on the article organisation and revised the article.

In the second article, B. Khalil proposed the methodology to overcome deficiencies in the conventional correlation-regression approach. B. Khalil, T.B.M.J. Ouarda and A. St-Hilaire had discussions and meetings throughout the work. B. Khalil carried out the analysis and wrote the manuscript. T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana revised the article.

In the third article, the modified version of the Kendall-Theil robust line proposed in this article was developed by B. Khalil and T.B.M.J. Ouarda. The generation of the synthetic records and the comparison of the four record extension techniques were carried out by B. Khalil. T.B.M.J. Ouarda assisted on the design of the Monte-Carlo and empirical experiments. B. Khalil, T.B.M.J. Ouarda and A. St-Hilaire had discussions and meetings throughout the work. T.B.M.J. Ouarda and A. St-Hilaire revised the manuscript.

In the fourth article, B. Khalil proposed the methodology to overcome deficiencies in the currently used approaches. B. Khalil, T.B.M.J. Ouarda and A. St-Hilaire had discussions and meetings throughout the work. B. Khalil carried out the analysis and wrote the manuscript. T.B.M.J. Ouarda and A. St-Hilaire revised the article.

In the fifth article, the idea of estimating water quality characteristics at ungauged sites originated from B. Khalil and T.B.M.J. Ouarda. B. Khalil developed the two models and carried out the comparison. B. Khalil, T.B.M.J. Ouarda and A. St-Hilaire had discussions and meetings throughout the work. T.B.M.J. Ouarda and A. St-Hilaire revised the manuscript.

Dedication

This dissertation is dedicated to my parents, Mohamed Khalil and Nadeen El-Shakankiry, my brother Hossam, for encouraging and supporting me and redirecting me at many critical occasions in my life, and to my wife, Noura, and my daughter, Nadine, who graciously supported me and gave me the time to pursue a dream.

Acknowledgment

The completion of this work would not have been possible without the continued support, patience, and insight of my advisors Prof. Taha B.M.J. Ouarda and Prof. André St-Hilaire. A special thanks to Prof. Ouarda for his guidance and support through the development and writing of several papers and this dissertation. Prof. Ouarda was a source of unending optimism and provides a keen vision of statistics, engineering and the world in which they fit, for which I am thankful. He spent countless hours with me through this process and was very supportive and understanding. I would like to express a deep appreciation to Prof. St-Hilaire, who enthusiastically aided me through fruitful discussions at various points in my research.

I owe my deepest gratitude to Prof. Shaden Abdel-Gawad, the chairperson of the National Water Research Center (NWRC) for her support throughout my career. Without her corporation I could not have obtained relevant data and support from different research institutes within the NWRC. I am indebted to my colleagues in the Drainage Research Institute and the Strategic Research Unit for their support.

I am grateful to my colleagues in the Canada Research Chair on the Estimation of Hydrometeorological Variables for their support. Their helpful attitude was typical of my experience with the entire staff of the INRS-ETE. Many thanks are also extended to Dr. Karem Chokmani and Dr. Salaheddine El-Adlouni for their classroom instructions and personal support.

Table of content

List of Tables

List of Figures

PART A: THESIS SUMMARY

1. Introduction

1.1 Contexte

L'eau douce est une ressource essentielle pour l'agriculture, l'industrie et l'existence humaine (Bartram et de la Balance, 1996). En effet, la demande en eau destinée à l'industrie, l'irrigation et la production hydroélectrique ne cesse de s'accroître avec le développement mondial. Une bonne qualité d'eau ainsi qu'une quantité suffisante sont essentielles pour le développement durable. La qualité de l'eau est une expression utilisée pour décrire les caractéristiques chimiques, physiques et biologiques de l'eau par rapport à un usage particulier.

La qualité de l'eau est influencée par un large éventail de phénomènes naturels et anthropiques. Différents processus naturels (hydrologiques, physiques, chimiques et biologiques) peuvent nuire aux caractéristiques des éléments et des composés chimiques de l'eau douce. De plus, plusieurs impacts anthropiques peuvent dégrader la qualité de l'eau comme l'activité industrielle, l'usage agricole ou des chantiers d'ingénierie fluviale (Chapman, 1996).

L'évaluation des ressources en eau exige une connaissance et une compréhension complète des processus affectant à la fois la quantité et la qualité de l'eau (Harmancioglu et al., 1999). Afin de comprendre la dynamique des processus pouvant altérer la qualité de l'eau d'un bassin versant, un programme bien conçu de suivi est requis. En effet, les programmes de suivi aident à appréhender les différents processus qui peuvent détériorer la qualité de l'eau et peuvent ainsi fournir les informations nécessaires à la gestion de cette qualité.

Les programmes de suivi de la qualité de l'eau englobent une variété d'activités qui comprennent les éléments suivants: une définition des objectifs de suivi et d'information souhaités, la conception du réseau de suivi, la conception de protocoles d'échantillonnage, le choix des équipements d'échantillonnage, l'analyse en laboratoire selon les méthodes standard, la vérification, le stockage et l'analyse des données.

Les activités d'un programme de suivi, telles que les procédures d'échantillonnage, la manipulation des échantillons, le stockage et l'analyse en laboratoire, doivent être effectuées par un personnel dûment formé pour assurer la qualité et l'utilité des données produites. Au cours des dernières décennies, les chercheurs se sont concentrés sur la conception de réseaux de suivi. La première étape de la conception d'un réseau de suivi est la définition des objectifs. La conception passe par la mutation des objectifs en un protocole qui décrit les variables à mesurer, ainsi que leur localisation spatiale et leur fréquence d'échantillonnage.

Historiquement, la localisation des sites d'échantillonnage de la qualité de l'eau était principalement basée sur l'accessibilité des sites, mais sans appliquer une approche systématique pour le choix de ces derniers (Sanders et al., 1983, Ward et al., 1990). Récemment, le nombre de sites a augmenté afin d'inclure des sites d'intérêt tels que celles situées en amont et en aval des zones hautement industrialisées ou très peuplées, des zones avec des sources de pollution connues et des zones d'utilisations intensives des terres (Tirsch et Male, 1984, Dixon et Chiswell, 1996). Toutefois, dans bien des cas, les visites des sites et l'échantillonnage ne sont effectués que lorsque le temps et le budget le permettent, sans qu'un véritable protocole soit établi sur la fréquence de la prise de mesures.

À la fin des années soixante, les critères de fréquence d'échantillonnage pour le suivi de la qualité des eaux établis dans l'état de Californie aux États-Unis ont été basés sur le débit de la rivière d'intérêt et sur les caractéristiques des bassins hydrographiques environnants (Sanders et al., 1983). C'était l'époque de l'expansion des réseaux de suivi des projets axés sur la création de réseaux régionaux ou nationaux. Par exemple, le suivi de la qualité de l'eau aux Pays-Bas avait commencé en 1950 sur quatre sites d'échantillonnage avec un nombre limité de variables dans la rivière Meuse. Ce réseau de suivi a ensuite été élargi sur environ 400 sites d'échantillonnage en 1981 dans tous les Pays-Bas. Sur certains de ces sites, environ 100 variables de la qualité de l'eau ont été suivies (Wetering et Groot, 1986). En Australie, un projet de suivi de la qualité de l'eau a été implanté dans certaines régions à partir des 1960. Par la suite, le programme de suivi de la qualité de l'eau a été établi au niveau du pays. Le réseau de l'état du Queensland incluait environ 700 sites d'échantillonnage (McNeil et al., 1989). Depuis lors, les gestionnaires de l'eau et les décideurs s'attendent à des informations de qualité propre à les aider dans la gestion de la qualité de l'eau.

Pendant les années 1980, différentes approches ont été proposées pour l'évaluation de la performance des réseaux de suivi et leur capacité à fournir les informations souhaitées. Par exemple, il a été conclu en 1981 que la révision du programme de suivi devrait avoir lieu aux Pays-Bas. En 1983, le nombre de points d'échantillonnage a été ramené de 400 à 260 (Wetering et Groot, 1986). Quant au programme de suivi de la qualité de l'eau du Queensland, après une enquête approfondie, le nombre de sites a été ramené de 700 à 400 (McNeil et al., 1989). En 1984, à la suite d'une évaluation effectuée par Lettenmaier et al. (1984) pour la municipalité de la région métropolitaine de Seattle, le réseau de suivi de qualité de l'eau des rivières a été réduit et le nombre de sites a passé de 81 à 47. Au Mexique, après une évaluation et une reconfiguration effectuées en 1997, le nombre de sites de suivi de la qualité de l'eau a été ramené de 564 à 200 (Ongley et Ordonez, 1997).

Ainsi, le problème de conception dans le suivi de la qualité de l'eau est devenu un problème d'évaluation et de reconfiguration du réseau (Harmancioglu et al., 1999). Toutefois, dans la plupart des cas récents et actuels, il y a une absence nette de stratégie ou de méthodologie pour la reconfiguration des réseaux de suivi. Une revue de littérature révèle que les anciennes approches de reconfiguration ont souvent été arbitraires, sans stratégie logique ou cohérente de conception (Strobl et al., 2006). Une méthodologie de reconfiguration de réseau de suivi inadéquate se traduit souvent par des données avec des informations limitées (GAO, 2000, 2004).

1.2 Problématique

Étant donné que la qualité des eaux est un sujet complexe, les approches statistiques peuvent apporter une contribution significative. Plusieurs méthodes statistiques sont utilisées pour évaluer la performance des réseaux de suivi de la qualité de l'eau. Dans cette optique, de nombreuses recherches ont été orientées vers l'évaluation des procédures de conception existantes et l'étude des moyens appropriés pour améliorer l'efficacité des réseaux existants (Harmancioglu et al., 1999). Toutefois, une revue de la littérature révèle plusieurs lacunes dans ces approches. En outre, il n'existe actuellement aucune stratégie établie ni aucune méthodologie pour la reconfiguration de réseaux de suivi, en particulier en ce qui concerne l'emplacement des sites d'échantillonnage (Strobl et al., 2006). Une méthodologie de reconfiguration logique et cohérente qui permet la collecte de données plus efficace et, par conséquent, des résultats plus utiles, est donc nécessaire. Une telle approche permettrait non seulement de meilleures recommandations pour lutter contre la pollution, mais aussi une meilleure allocation des ressources financières ainsi qu'une meilleure compréhension de l'écosystème étudié (Strobl et Billard, 2008).

1.3 Objectifs

L'objectif principal de cette étude est de présenter des méthodes objectives et systématiques pour l'évaluation et la reconfiguration des réseaux de suivi de la qualité des eaux de surface. Ces méthodes devraient pouvoir surmonter les limites des approches actuellement appliquées. En outre, cette étude vise à introduire une approche novatrice pour l'estimation des caractéristiques de la qualité de l'eau sur des sites non jaugés. Plus précisément, quatre objectifs spécifiques ont été choisis afin d'optimiser les informations tirées de cette étude pour une application directe dans le processus de conception des réseaux de suivi de la qualité de l'eau:

- Revoir les méthodes statistiques appliquées actuellement pour l'évaluation et la reconfiguration des réseaux de suivi de la qualité des eaux de surface;
- Développer une approche statistique pour la rationalisation des variables de la qualité de l'eau;
- Développer une approche statistique pour l'évaluation et la refonte des sites d'échantillonnage de la qualité des eaux de surface;
- Introduire une méthodologie statistique pour l'estimation des caractéristiques de la qualité de l'eau sur des sites non jaugés.

1.4 Organisation de la synthèse

Ce résumé est divisé en sept sections principales. La section 1 traite de la définition du problème et des objectifs des travaux. La section 2 présente une revue de littérature des approches statistiques proposées pour l'évaluation et la reconfiguration des réseaux de suivi de la qualité des eaux de surface. La section 3 introduit la zone d'étude. La section 4 résume les méthodologies proposées. La section 5 détaille les principaux résultats obtenus. Les conclusions du travail sont données dans la section 6. Finalement, les recommandations pour les travaux futurs sont présentées dans la section 7.

2. Revue de littérature

Cette section présente une revue de littérature des méthodes statistiques utilisées pour l'évaluation et la reconfiguration des réseaux de suivi de la qualité de l'eau de surface. Les principaux aspects techniques de la conception du réseau sont répartis en quatre sous-sections: (i) les objectifs de suivi, (ii) les variables de la qualité de l'eau, (iii) la fréquence d'échantillonnage et (iv) la distribution spatiale des sites d'échantillonnage.

Dans cette section, les objectifs de suivi et les procédures connexes utilisées pour l'évaluation et la reconfiguration à long terme des réseaux de suivi de la qualité des eaux de surface sont discutés. La pertinence de chaque approche pour la conception, la réduction ou l'expansion des réseaux de suivi est aussi discutée. Pour chaque approche statistique, les avantages et les inconvénients sont examinés dans une perspective de reconfiguration de réseau. Finalement, des méthodes pour pallier aux lacunes des approches statistiques actuellement utilisées sont recommandées.

Un résumé de la revue de littérature est présenté en cinq sous-sections. La première section présente les objectifs du suivi. La deuxième section décrit les méthodes statistiques utilisées pour la sélection des variables de la qualité de l'eau. La troisième section traite des méthodes statistiques utilisées pour l'évaluation des fréquences d'échantillonnage. Les approches statistiques utilisées pour l'évaluation et la reconfiguration des sites de suivi sont présentées dans la quatrième section. Enfin, la cinquième section présente les conclusions. La revue de littérature détaillée est présentée dans l'article I de la partie B.

2.1 Objectifs de suivi

Les objectifs de suivi devraient définir l'information attendue à la sortie du réseau. Mal préciser l'information désirée conduit à l'échec du réseau lui-même (Harmancioglu et al., 1992). Les objectifs de suivi constituent le fondement sur lequel un programme est construit. Un obstacle majeur est que les objectifs sont souvent décrits en termes globaux plutôt que spécifiques et précis (Harmancioglu et al., 1999). Trois principaux défis peuvent survenir lors de l'identification des objectifs de suivi: (i) sélectionner les méthodes d'analyse à partir de plusieurs objectifs potentiels; (ii) préciser l'objectif; et (iii) transformer les objectifs dans les questions de statistiques.

La définition des objectifs de suivi est essentielle pour la conception et le fonctionnement du réseau. Plusieurs objectifs pour la surveillance de la qualité de l'eau sont publiés dans la littérature (Ward et al., 1979; Sanders et al., 1983; Whitfield, 1988; Zhou, 1996; et Harmancioglu et al., 1999). Les objectifs les plus courants se présentent comme suit: l'identification des tendances spatiales et temporelles, l'évaluation de la conformité avec les normes et les règles, la simplification des études et des évaluations d'impact, la détermination de la pertinence des divers usages de l'eau, l'exécution générale de la surveillance, l'évaluation de différentes stratégies de contrôle, l'estimation du transport des particules dans les rivières et la simplification de la modélisation de la qualité de l'eau ou d'autres activités de recherches spécifiques.

2.2 Variables de la qualité de l'eau

La qualité de l'eau est généralement décrite par un ensemble de variables physiques, biologiques et chimiques. Elle peut être définie par une seule variable, ou par des centaines de composés chimiques. Donc, le choix des variables appropriées pour la caractérisation de la qualité de l'eau est un problème très complexe (Sanders et al., 1983). La plupart des chercheurs reconnaissent qu'il n'est pas possible de mesurer toutes les variables environnementales et qu'il faut les choisir d'une manière objective et logique, ce choix étant une étape intégrale et cruciale dans l'établissement d'un système de suivi de la qualité de l'eau (Ward et al., 1990).

Le choix des variables à mesurer (la conception) et l'ajout de nouvelles variables à celles déjà mesurées (extension) suscitent plusieurs questions, comme le type et les objectifs du suivi, les caractéristiques du bassin et le budget disponible. Les méthodes proposées pour l'évaluation et la sélection des variables entrent dans la catégorie de la réduction. Deux approches principales fréquemment décrites dans la littérature pour choisir les variables sont: la méthode de corrélation-régression (CR) et l'analyse en composantes principales (ACP).

La méthode CR est composée de trois étapes: la première étape est l'évaluation du degré d'association entre les variables mesurées par l'analyse de corrélation. Un degré élevé de corrélation entre les variables indique que certaines informations peuvent être redondantes. La deuxième étape est le choix des variables de la qualité de l'eau qui seront mesurées ou qui cesseront d'être mesurées. Cette étape est fondée sur des paramètres subjectifs, comme l'importance de la variable, l'inclusion de la variable dans les lois et règlements locaux ou internationaux, etc. Ces paramètres peuvent inclure le coût de l'analyse en laboratoire. Finalement, la troisième étape est la reconstruction de l'information à partir des variables non mesurées en utilisant d'autres variables qui sont mesurées en continu.

L'ACP peut être appliquée à un ensemble de variables de la qualité de l'eau afin de découvrir les variables qui forment des sous-ensembles cohérents qui sont relativement indépendants les uns des autres. Les variables qui sont corrélées les unes aux autres, mais indépendantes des autres sous-ensembles de variables sont combinées en une seule composante. Les composantes sont censées refléter les processus sous-jacents qui ont créé les corrélations entre les variables. Mathématiquement, l'ACP produit plusieurs combinaisons linéaires des variables observées, et chaque combinaison linéaire est constituée d'une composante. Les composantes résument les modèles de corrélation dans une matrice de corrélations observées. La matrice de saturation des composantes, obtenue par l'ACP, reflète les caractéristiques de cette procédure d'extraction qui maximise la variance expliquée successivement dans chaque composante qui sont orthogonales entre elles.

2.3 Fréquence d'échantillonnage

La fréquence d'échantillonnage est un aspect très important de la conception du réseau du suivi de qualité de l'eau. Elle influence l'utilité des données, ainsi que le coût de l'opération. Les informations obtenues par des échantillonnages fréquents peuvent être redondantes et trop coûteuses; cependant, l'échantillonnage rare peut limiter la précision et l'interprétation des observations. Les méthodes statistiques proposées pour l'évaluation et le calcul des fréquences d'échantillonnage sont directement liées aux objectifs de suivi et aux méthodes d'analyse de données.

Lettenmaier (1976) a proposé une méthode pour déterminer les intervalles d'échantillonnage optimaux basés sur un test paramétrique de tendance, où la fréquence d'échantillonnage nécessaire correspond à une puissance spécifiée du test de tendance. Cette méthode est plus connue sous l'appellation « efficacité de l'échantillon » (''effective sample'', ES). L'approche est composée de deux étapes. La première étape définit le nombre maximal d'échantillons qui peuvent être collectés par an afin d'éviter les auto-corrélations, ou au moins, de réduire leurs effets. Quant à la deuxième étape, elle estime la durée requise pour détecter les tendances au niveau de confiance et aux puissances spécifiées. Sanders et Adrian (1978) et Sanders et al. (1983) avait recommandé l'utilisation d'intervalles de confiance basés sur la moyenne comme principal critère de sélection des fréquences d'échantillonnage. Le but principal est de choisir la fréquence d'échantillonnage qui donne une estimation de la moyenne avec un degré prescrit de précision à l'intérieur des limites de confiance. Zhou (1996) a proposé une approche visant à définir la fréquence d'échantillonnage en fonction de la périodicité, fondée sur l'analyse harmonique.

Quand l'objectif est de déterminer la conformité avec les normes ou les règles, on peut utiliser le test binomial pour évaluer la fréquence d'échantillonnage. Dans ce cas, les données aberrantes ou extrêmes sont définies comme les échantillons qui dépassent d'une manière quelconque les limites prédéfinies, telles que la norme pour l'eau potable, et une limite de conformité ou d'action (Ward et al., 1990). Mace (1964) et Ward et al. (1990) ont décrit une approche pour estimer la taille de l'échantillon nécessaire pour contrôler le risque d'erreurs de type I et II lors de l'évaluation de la proportion de temps au cours de laquelle un critère est dépassé. Cette approche est valide quand les échantillons individuels sont évalués comme étant «au-dessus» ou «en-dessous» d'un critère (échelle nominale). Le nombre de dépassements d'une limite suit toujours la loi binomiale indépendamment de la distribution de données (Ellis et Lacey, 1980).

Tirsch et Male (1984) ont abordé la reconfiguration de la fréquence d'échantillonnage à l'aide de modèles de régression linéaire multivariée. La précision de suivi, décrite par le coefficient correcteur de détermination de régression, est exprimée en fonction de la fréquence d'échantillonnage. La notion d'entropie a été introduite pour déterminer les intervalles d'échantillonnage optimaux pour le suivi de la qualité de l'eau par Harmancioglu (1984). Le principe d'entropie est appliqué pour déterminer l'information contenue dans les variables stochastiques dépendantes afin de définir les intervalles d'échantillonnage optimaux en fonction du temps.

La méthode du semi-variogramme est fondée sur les travaux de Krige (1951) et Matheron (1963) et constitue la première étape d'une analyse géostatistique. Le semi-variogramme est une représentation graphique qui illustre comment la similitude entre les valeurs varie en fonction de la distance, de la direction ou du temps qui les séparent. Khalil et al. (2004) ont introduit le semi-variogramme pour définir la portée effective de corrélation et, conséquemment, évaluer la fréquence d'échantillonnage. L'objectif est de définir la fréquence d'échantillonnage afin d'obtenir une série de données indépendantes. La fréquence d'échantillonnage dans cette approche est basée sur la portée effective de corrélation.

2.4 Sites d'échantillonnages

Le choix des sites d'échantillonnage est un aspect important dans la conception d'un réseau de suivi. Les pratiques au début de l'échantillonnage de la qualité de l'eau étaient axées sur les sites accessibles, mais sans appliquer une approche objective et systématique du choix des sites (Harmancioglu et al., 1999). Par la suite, le nombre de sites a augmenté pour inclure des stations de points d'intérêt telles que celles situées en amont et en aval des zones hautement industrialisées ou très peuplées, des zones avec des sources de pollution et des zones d'utilisation intensives des terres (Tirsch et Male, 1984). Ensuite, diverses approches ont été proposées pour la sélection du nombre et de l'emplacement des stations d'échantillonnage.

L'approche de la classification hiérarchique des cours d'eau, proposée par Sanders et al. (1983) est basée sur l'ordre des cours d'eau pour décrire un réseau de suivi. Dans la procédure de classification des cours d'eau, un tronçon source est du premier ordre. Un cours d'eau qui n'est composé que de tributaires du premier ordre est désigné avec l'ordre deux, et ainsi de suite (Horton, 1945). Cette approche systématique localise les sites d'échantillonnage afin de diviser le réseau fluvial en sections qui sont similaires en contribution des affluents, de débit ou de charges en polluantes.

Une autre approche basée sur une régression linéaire multivariée a été proposée par Tirsch et Male (1984). Dans cette approche, chaque site de suivi est considéré comme la variable dépendante, et le modèle de régression considère les combinaisons des autres sites comme des variables indépendantes. Un coefficient de détermination ajusté est alors calculé pour chaque modèle et la précision du suivi change avec l'ajout ou l'élimination de certains sites au sein du réseau. Un coefficient de détermination élevé indique qu'il existe un degré élevé de redondance et que le site sélectionné comme variable dépendante pourrait ne pas être nécessaire.

Harmancioglu et Alpaslan (1992) ont proposé une approche basée sur la notion d'entropie dans laquelle la quantité d'information mutuelle entre les sites d'échantillonnage est déterminée selon le degré d'incertitude (Harmancioglu et al., 1999). La dépendance entre les sites d'échantillonnage se traduit par moins d'entropie entre ces derniers. Si la dépendance est cohérente avec le temps, un ou plusieurs sites peuvent être abandonnés avec une perte minimale d'informations.

Différentes méthodes multivariées ont été employées pour le remaniement des sites de suivi de la qualité de l'eau. Celles-ci incluent l'analyse en composantes principales (ACP), la classification ascendante hiérarchique et l'analyse discriminante.

2.5 Conclusions de la revue de littérature

Afin que les objectifs de suivi puissent aider à la conception du réseau et produire l'information souhaitée, ceux-ci doivent être définis clairement et spécifiquement. De plus, il y aurait des objectifs spécifiques pour chaque site. L'énoncé des objectifs devrait inclure le but du suivi à cet endroit, les variables à surveiller, le type d'information souhaitée et l'outil d'analyse à utiliser pour obtenir les informations souhaitées. La revue de littérature révèle que, bien que plusieurs recherches ont été entreprises pour évaluer les performances des réseaux de suivi, plusieurs lacunes dans les approches proposées persistent.

Les deux approches principales proposées dans la littérature pour la rationalisation des variables de la qualité de l'eau sont les méthodes CR et ACP. La méthode CR a l'avantage de permettre la reconstruction de l'information concernant les variables abandonnées. Toutefois, les deux approches sont basées sur l'hypothèse d'une relation linéaire entre les variables de la qualité de l'eau. Cependant les relations entre les variables physiques, biologiques et chimiques peuvent être non linéaires. De ce fait, la mesure de l'information mutuelle et les réseaux de neurones artificiels (RNA) peuvent être utilisés au lieu des analyses de corrélation et de régression linéaires. L'information mutuelle est une mesure de la dépendance non linéaire ou de la quantité d'informations redondantes entre deux variables. La méthode des RNA est plus souple que les modèles régressifs pour capturer les relations entre les variables de la qualité de l'eau et nécessite moins de connaissances préalables du système. Cependant, une lacune de cette méthode est l'absence d'un critère pour identifier les variables à mesurer en permanence et celles qui peuvent être abandonnées. Un indice de performance, qui se base sur un critère d'information, peut aider à surmonter ce problème.

Les approches proposées pour l'évaluation de la fréquence d'échantillonnage considèrent une variable spécifique pour un site spécifique et cela conduit souvent à l'optimisation d'un seul objectif. Le suivi de la qualité de l'eau ne peut pas traiter chaque besoin d'information avec une seule procédure de collecte des données; mais le système doit tenter de répondre simultanément à plusieurs objectifs d'information. Plusieurs suggestions peuvent aider à évaluer la fréquence d'échantillonnage pour un réseau avec plusieurs d'objectifs. Par exemple, les différentes fréquences d'échantillonnage peuvent être utilisées pour différents objectifs de suivi afin d'optimiser l'obtention d'informations. L'évaluation des différentes fréquences d'échantillonnage peut être faite en fonction de leur aptitude à satisfaire des objectifs multiples, plutôt que d'optimiser en fonction d'un seul objectif.

Pour l'évaluation des sites, l'approche de classification hiérarchique des cours d'eau est l'approche la plus fréquemment utilisée pour la conception d'un réseau de suivi, quand les données sur la qualité de l'eau ne sont pas disponibles. Le principal inconvénient des méthodes d'entropie et de régression est que ces deux approches ne considèrent qu'une seule variable de la qualité de l'eau. Toutefois, l'évaluation et le remaniement des sites d'un réseau de suivi de la qualité de l'eau sont plus fiables lorsqu'ils se fondent sur plusieurs indicateurs de la qualité de l'eau. Les données multidimensionnelles doivent être traitées simultanément et non pas de manière séquentielle. Les approches basées sur l'analyse de données multivariées remédient à cet inconvénient en utilisant simultanément plusieurs variables de la qualité de l'eau. Mais les approches basées sur l'analyse de données multivariées ne considèrent pas la reconstruction de l'information sur des sites abandonnés.

Les approches proposées ont l'inconvénient commun de se concentrer uniquement sur l'identification des sites déjà suivis et qui seront finalement abandonnés. Cependant, la reconfiguration spatiale optimale peut comprendre l'élimination de sites existants et l'ajout de nouveaux sites non jaugés. Cet inconvénient résulte de l'habitude de se concentrer plus sur l'évaluation en utilisant les données sur la qualité de l'eau déjà obtenues, et ignorer les caractéristiques des bassins surveillés. L'intégration des caractéristiques du bassin dans l'évaluation des sites de suivi est censée remédier, au moins en partie, à cette lacune.

3. Réseau national Égyptien du suivi de la qualité de l'eau

L'Égypte est un pays semi-aride avec une pluviométrie qui dépasse rarement 200 mm/an le long de la côte nord. L'intensité de la pluie diminue rapidement quand on s'éloigne des zones côtières et les averses éparses ne sont guère utiles pour la production agricole (Abu-Salama, 2007). Selon une entente entre l'Égypte et le Soudan (1959), la répartition de l'eau du Nil est de 18,5 milliards de m^3 au Soudan et 55,5 milliards de m^3 pour l'Égypte (Dijkman, 1993). Environ 97 % des ressources en eau de l'Égypte proviennent du Nil.

Dans une étude mondiale axée sur l'eau douce, on a annoncé que l'Égypte faisait partie des dix pays qui seront aux prises avec des problèmes liés aux ressources en eau d'ici 2025 en raison de la croissance démographique (Engelman et Le Roy, 1993). La distribution de la portion égyptienne de l'eau du Nil par rapport à sa population tend vers le seuil de pauvreté en eau et chutera bien en-dessous de ce dernier dans les prochaines années (MWRI, 1997; Wolf, 2000). Les ressources en eau douce en Égypte sont maintenant de presque 800 m^3 par personne par an (Abdel-Gawad et al., 2004; Frenken, 2005). Une des solutions appliquées pour remédier aux limites des ressources en eau en Égypte est la réutilisation des eaux de drainage agricole dans les processus de production agricole. Notons que l'eau de drainage avec une faible salinité est utilisée directement ou après un mélange avec l'eau du Nil, tandis que les eaux avec une haute salinité ou contaminées par les déchets municipaux ou industriels ne peuvent pas être utilisées pour l'irrigation.

Le système de drainage dans le Delta du Nil est composé de vingt-deux bassins versants. En fonction de leur qualité, les effluents sont déchargés dans les lacs du nord ou pompés dans des canaux d'irrigation de vingt et un sites le long des principaux drains pour augmenter la provision d'eau douce (DRI-MADWQ, 1998). Plusieurs programmes ont été développés dans l'objectif d'assurer le suivi de la qualité de l'eau du Nil et des eaux de drainage agricole en Égypte. En 1977, le Centre National de Recherches sur l'Eau (The National Water Research Center, NWRC) a continuellement étendu ses activités de suivi pour couvrir le nombre croissant de sites d'échantillonnage et de variables de qualité de l'eau. Le programme de suivi du système de drainage du Delta du Nil vise à évaluer la conformité avec les normes nationales, estimer le transport des charges, et identifier les tendances temporelles et spatiales (NAWQAM, 2001). .

Le réseau de suivi du système de drainage du Delta du Nil est assuré par quatre-vingt-quatorze sites d'échantillonnage à travers lesquels trente-trois variables de la qualité d'eau sont mesurées sur une base mensuelle (Figure 1 et Tableau 1). Toutes les analyses de laboratoire sont effectuées par le « Central Laboratory for Environmental Quality Monitoring ».

Les sites d'échantillonnage sont distribués comme suit: vingt et un sont localisés aux sites de réutilisation des eaux de drainage; dix sont sur le système de drainage (localisés sur les principaux cours d'eau, servent de points de contrôle pour évaluer la charge polluante et la quantité de sels); treize sites de contrôle placés sur les tributaires de petits bassins versants dont les eaux coulent vers les systèmes principaux de drainage; cinquante sites de contrôle sont placés sur des affluents qui fournissent de l'eau aux systèmes de drainage principal. L'annexe montre un exemple typique de l'analyse préliminaire effectuée pour chacun des sites de surveillance.

Tableau 1. Variables de la qualité de l'eau mesurées par le réseau national égyptien de suivi de la qualité de l'eau.

Variable de qualité	**Symbole**	**Unités**	**Variable de qualité**	**Symbole**	**Unités**
Demande biochimique d'oxygène	*BOD*	mg/l	Débit	*Q*	m^3/sec
Demande chimique d'oxygène	*COD*	mg/l	Température	*T*	^{o}C
Oxygène dissous	*DO*	mg/l	Acidité	*pH*	-
Conductivité spécifique	*EC*	dS/m	Solides en suspension	*TSS*	mg/l
Solides dissous	*TDS*	mg/l	Solides volatiles totaux	*TVS*	mg/l
Calcium	*Ca*	mg/l	Turbidité	*Turb*	NTU
Magnésium	*Mg*	mg/l	Visibilité par disque Secchi	*Vis*	cm
Sodium	*Na*	mg/l	Coliformes totaux	*TColi*	MPN/100ml
Potassium	*K*	mg/l	Coliformes fécaux	*FColi*	MPN/100ml
Bicarbonate	HCO_3	mg/l	Cadmium	*Cd*	mg/l
Sulphate	SO_4	mg/l	Manganèse	*Mn*	mg/l
Chlore	*Cl*	mg/l	Cuivre	*Cu*	mg/l
Nitrate	NO_3	mg/l	Fer	*Fe*	mg/l
Ammonium	NH_4	mg/l	Zinc	*Zn*	mg/l
Phosphore total	*TP*	mg/l	Nickel	*Ni*	mg/l
Azote total	*TN*	mg/l	Bore	*B*	mg/l
			Plomb	*Pb*	mg/l

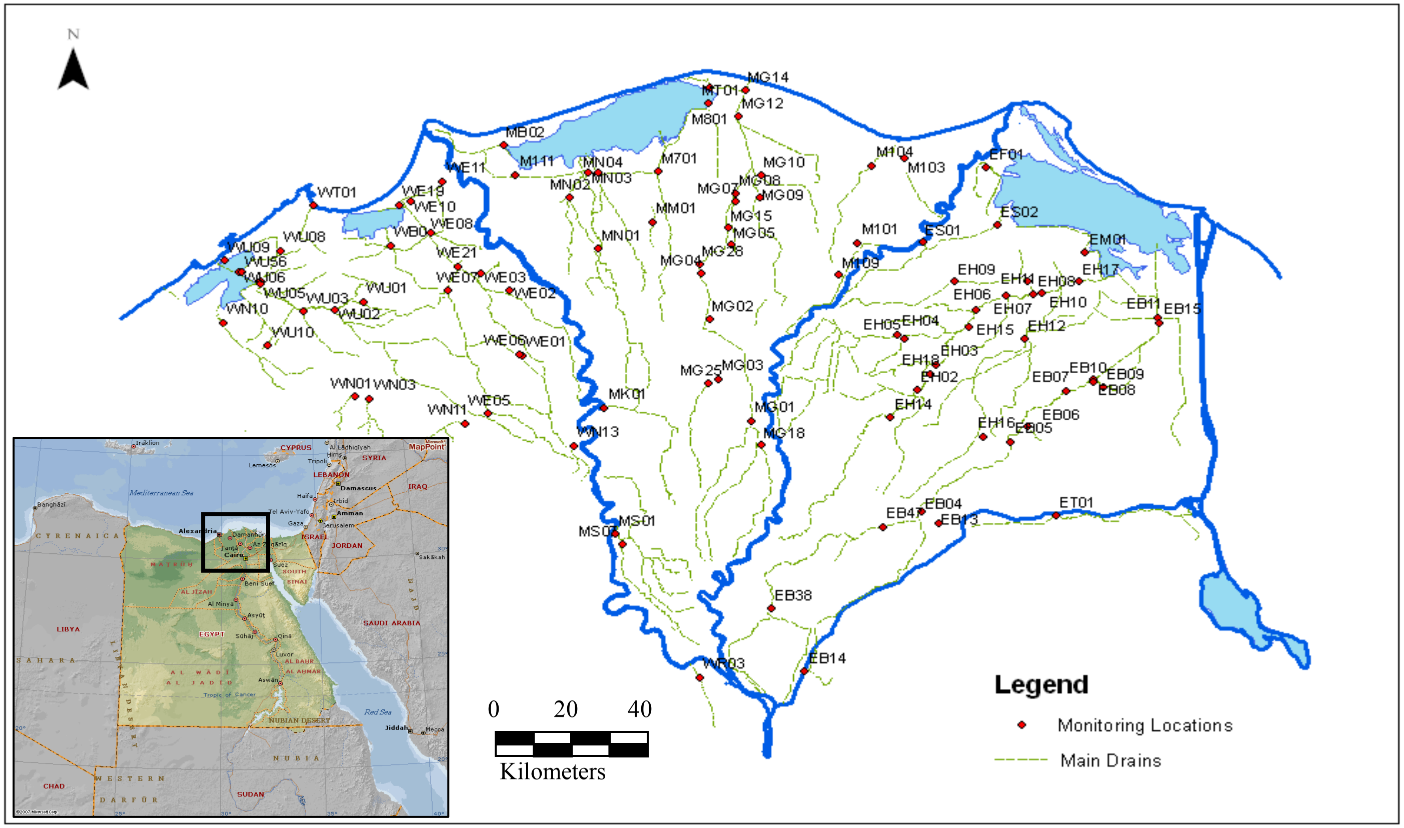


Figure 1. Les sites de suivi de la qualité de l’eau du delta du Nil (source: NWRC, 2001)

4. Méthodologie

Dans cette étude, quatre nouvelles méthodes principales sont proposées. (i) la rationalisation des variables de la qualité de l'eau; (ii) l'extension des séries chronologiques existantes; (iii) l'évaluation et la reconfiguration des sites d'échantillonnage; et (iv) l'estimation des caractéristiques de la qualité de l'eau sur des sites non jaugés.

4.1 La rationalisation des variables de la qualité de l'eau

Peu de travaux ont porté sur la rationalisation des variables de la qualité de l'eau. L'approche de corrélation-régression (CR) a le principal avantage de permettre la reconstitution des informations sur les variables abandonnées à l'aide de l'analyse de régression.

Cependant, trois défauts principaux existent dans l'approche CR tel qu'utilisée en pratique de nos jours pour la réduction de variables de la qualité de l'eau. La première est la méthode utilisée pour identifier les variables hautement associées. Le coefficient de corrélation est couramment utilisé comme un critère pour évaluer le degré de l'association, mais la sélection du bon seuil au-dessus duquel un coefficient de corrélation peut être considéré comme suffisant pour associer deux variables peut être problématique. L'évaluation du coefficient de corrélation est toujours subjective et demeure un choix pour le concepteur. Ainsi, différents chercheurs peuvent arriver à différents résultats en utilisant le même ensemble de variables. La deuxième insuffisance de la méthode CR est l'absence d'un critère pour identifier la combinaison de variables à mesurer en permanence et celles à abandonner. La dernière insuffisance est l'utilisation de l'analyse de régression pour reconstituer les informations sur les variables abandonnées, qui entraîne souvent une sous-estimation de la variance dans les données estimées (Alley et Burns, 1983; Hirsch, 1982). L'objectif principal de cette étude est de modifier l'approche traditionnelle de la méthode CR pour surmonter ces lacunes.

L'approche proposée se compose de trois étapes. La première étape est l'évaluation du degré d'association entre les variables et la définition des ensembles de variables qui sont hautement associés. Des critères sont élaborés pour déterminer un seuil de coefficient de corrélation au-dessus duquel les variables peuvent être considérées comme hautement associées. Ces critères sont basés sur les procédures d'augmentation d'enregistrements. Un seuil de corrélation est déterminé selon les conditions d'estimateurs améliorés (avec une faible variance) de la moyenne (d_m) et de la variance (d_v) pour une variable abandonnée à l'aide des formules fournies par Matalas et Jacobs (1964). Le regroupement hiérarchique est effectué pour les variables, tout en utilisant le coefficient de corrélation comme mesure de proximité. On emploie le seuil du coefficient de corrélation pour définir le degré de différence afin d'identifier les meilleurs groupements.

Après l'identification des groupements de variables hautement associées, la deuxième étape consiste à étudier chaque groupe multi-variable séparément. L'approche suppose que chaque variable au sein du groupement sera abandonnée. Pour chaque variable abandonnée, la meilleure variable auxiliaire pour l'extension des enregistrements est sélectionnée à partir d'autres variables dans le même groupe. La meilleure variable auxiliaire est celle qui réduit la variance de la moyenne estimée de la variable abandonnée.

La troisième étape est d'évaluer les différentes combinaisons de variables qui seront abandonnées et les variables à mesurer en permanence. Par exemple, nous prenons le cas où les réductions budgétaires nécessitent que k des variables soient abandonnées. Quelles k variables entre les w variables dans la liste des variables à mesurer doivent être choisis? Le nombre de combinaisons possibles à abandonner est obtenu par le coefficient binomial, $C(w,k)$. Nous proposons un indice de performance globale pour évaluer les combinaisons diverses, et par conséquent, pour classer les différentes combinaisons. Cette procédure permet l'identification de la combinaison optimale des variables à abandonner et fournit le rang des meilleures combinaisons à abandonner pour le décideur. La Figure 2 illustre le flux des analyses telles que résumées ci-dessus. La méthodologie détaillée et les résultats sont présentés dans l'article II de la partie B.

4.2 La reconstitution des informations sur les variables abandonnées

L'extension mensuelle, hebdomadaire ou quotidienne d'enregistrements dans un site à court enregistrement à partir d'un autre site avec des mesures en continu est appelée extension des séries chronologiques («record extension»). La régression (« ordinary least squares regression » OLS) est une technique couramment utilisée pour l'extension d'enregistrements. Toutefois, son objectif est de générer des estimations optimales pour chaque enregistrement par jour (ou par mois), plutôt que les caractéristiques de la population. En général, l'OLS a tendance à sous-estimer la variance. La ligne de corrélation organique (« line of organic correlation » LOC) a été développée pour corriger ce biais. D'autre part, la méthode de la ligne robuste de Kendall-Theil (« Kendall-Theil robust line » KTRL) a été proposée en tant qu'analogue de l'OLS, mais a l'avantage d'être robuste en présence des valeurs extrêmes.

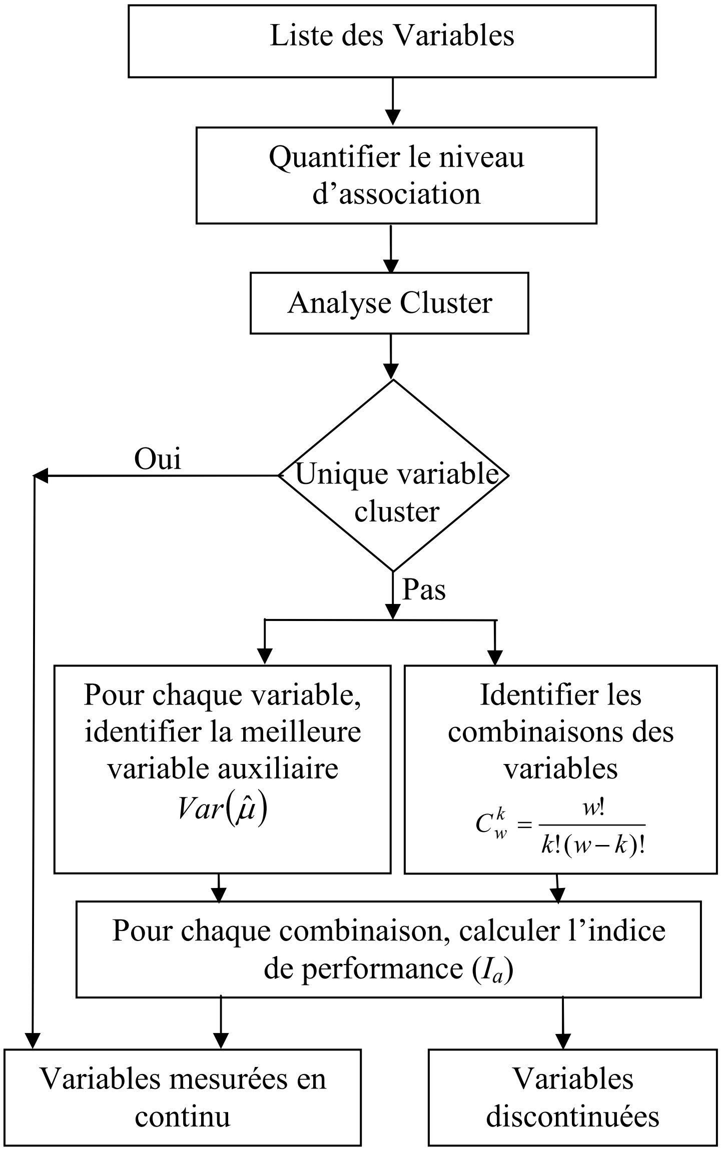

Liste des Variables
Quantifier le niveau d'association
Analyse Cluster
Unique variable cluster
Oui
Pas
Pour chaque variable, identifier la meilleure variable auxiliaire $Var(\hat{\mu})$
Identifier les combinaisons des variables $C_w^k = \frac{w!}{k!(w-k)!}$
Pour chaque combinaison, calculer l'indice de performance (I_a)
Variables mesurées en continu
Variables discontinuées

Figure 2. Schéma de l'approche de rationalisation proposée

Dans cette étude, quatre méthodes d'extension d'enregistrements sont décrites et leurs propriétés sont explorées. Ces méthodes sont notamment : OLS, LOC, KTRL et une nouvelle méthode proposée dans cette thèse (KTRL2) qui combine l'avantage de LOC en réduisant le biais dans l'estimation de la variance et la robustesse de KTRL en présence des valeurs extrêmes.

Afin d'évaluer les quatre techniques d'extension des séries chronologiques, un essai Monte-Carlo et une expérience empirique sont menés. Les essais Monte-Carlo permettent la comparaison et l'évaluation des différentes méthodes d'extension des séries chronologiques à l'aide d'enregistrements avec des propriétés distributives et statistiques prédéfinies. L'expérience empirique permet l'évaluation des quatre méthodes d'extension des séries chronologiques à l'aide de données fiables. Dans les deux expériences Monte-Carlo et empiriques, les quatre techniques d'extension des séries chronologiques sont évaluées pour différents degrés de corrélation entre les variables dépendantes et indépendantes, ainsi que pour différentes tailles d'enregistrements. Les deux expériences, de Monte-Carlo et empiriques, sont menées pour évaluer ces quatre techniques, le biais et l'erreur standard du moment des estimations et la gamme complète des percentiles.

Pour l'essai Monte-Carlo, les séquences de variables dépendantes et indépendantes sont générées pour 120 cas à partir d'une distribution normale bivariée, avec une moyenne de 0 et des écarts-types de 1. Nous considérons trois coefficients de corrélation croisée et différentes combinaisons du nombre d'enregistrements au cours de la période simultanée (n_1) et la période d'estimation (n_2). Les expériences Monte-Carlo sont effectuées pour les valeurs (n_1, n_2), de (96, 24), (72, 48), (48, 72), et (24,96), et pour un coefficient de corrélation (ρ) de 0.5, 0.7, et 0.9. Les essais sont effectués avec chacune des douze combinaisons de ρ et (n_1, n_2) pour évaluer la capacité des quatre méthodes d'extension des séries chronologiques à reproduire les différentes propriétés statistiques de la population dans la série. La série étendue est évaluée selon l'estimation de la moyenne, et de l'écart-type, puis sur la gamme complète des percentiles (du 5^{e} jusqu'au 95^{e} percentile).

Une expérience empirique vise à examiner l'utilité des quatre techniques pour la reproduction d'enregistrements qui préservent les caractéristiques statistiques des solides dissous totaux (TDS) et du chlore (Cl). Trois combinaisons de tailles d'enregistrements au cours de la période simultanée (n_1) et de la période d'extension (n_2) sont considérées pour chacun des deux modèles d'estimation (TDS et CL). Les expériences empiriques sont effectuées pour les valeurs (n_1, n_2) de (108, 12), (84, 36) et (60, 60). Afin d'évaluer la performance des quatre méthodes d'extension des séries chronologiques, une méthode de validation croisée (méthode de rééchantillonnage « jackknife ») est appliquée. Pour la validation croisée, lorsque (n_1, n_2) est égal à (108, 12), un an d'enregistrements mensuels est supprimé des dix années de données disponibles. Les valeurs mensuelles pour l'année supprimée sont estimées en utilisant les quatre techniques d'extension des séries chronologiques calibrées avec les neuf années restantes. De même, quand (n_1, n_2) est égal à (84, 36) ou (60, 60), la validation croisée est appliquée pour estimer trois et cinq ans de données mensuelles respectivement.

Pour chacun des 94 emplacements de mesure, les quatre méthodes d'extension des séries chronologiques sont appliquées pour estimer les TDS et le Cl en utilisant la conductivité électrique (EC) comme variable explicative. Ainsi, pour chacun des deux modèles d'estimation considérés, lorsque (n_1, n_2) est égal à (108, 12), 940 (94 emplacements × 10 différentes combinaisons d'échantillons) réalisations différentes sont considérées pour l'enregistrement étendu de la qualité de l'eau. Mais, quand (n_1, n_2) est égal à (84, 36) et (60, 60), des combinaisons successives ou non successives de trois et cinq années sont considérées. Par conséquent, avec les dix années d'enregistrements mensuels disponibles, $C(10,3) = 120$ et

C(10,5) = 252 combinaisons possibles sont considérées. Ainsi, 11280 (94 sites × 120 combinaisons d'échantillons différentes) réalisations différentes pour l'enregistrement étendu de la qualité de l'eau sont considérées lorsque (n_1, n_2) est égal à (84, 36) et 23688 (94 sites × 252 combinaisons d'échantillons différentes) sont considérées lorsque (n_1, n_2) est égal à (60, 60). Les détails des simulations Monte-Carlo et des études empiriques, ainsi que les résultats obtenus sont présentés dans l'article III de la partie B.

4.3 L'évaluation et le remaniement des sites d'échantillonnages

Malgré les différentes approches statistiques proposées dans la littérature pour l'évaluation et le remaniement des sites de suivi de la qualité des eaux de surface, plusieurs déficiences persistent toujours. Ces lacunes varient d'une approche à l'autre, et comprennent généralement: (i) l'ignorance des attributs du bassin surveillé; (ii) le traitement séquentiel des données de la qualité de l'eau multivariée plutôt que de manière simultanée; (iii) l'accent porté principalement sur les sites qui seront suspendus; et (iv) l'ignorance des informations de reconstitution sur les sites abandonnés.

Dans cette étude, une méthodologie qui surmonte ces lacunes est proposée. Pour ce faire, le bassin surveillé est divisé en sous-bassins et une analyse des groupements hybride est utilisée pour identifier les groupes de sous-bassins avec des attributs similaires. Une stratégie d'échantillonnage optimal stratifié est ensuite employée pour identifier le nombre optimal de sites à mesurer pour chacun des groupes de sous-bassins. Un indice des informations agrégées est utilisé pour déterminer la combinaison optimale des sites qui seront abandonnés.

Afin d’intégrer les attributs de la région surveillée dans l’évaluation et le remaniement des sites de suivi, le Delta du Nil est divisé en unités spatiales (USs). Chaque unité spatiale est un domaine drainé par un seul point sur le système de drainage. Mesurer la qualité de l’eau à ce point décrit l’effet des attributs de l’US sur l’état de la qualité de l’eau. Ainsi, pour chaque unité spatiale neuf des attributs qui expliquent les différents effets naturels et anthropiques sont identifiés. On suppose que les attributs d’une US sont les principales sources de pollution dans les systèmes de drainage. Comme pour les variables de qualité de l’eau mesurées, les variables qui expliquent la variabilité de la qualité de l’eau dans le système de drainage du Delta du Nil sont sélectionnées selon l’analyse en composantes principales (ACP). Basées sur l’ACP, les quatre variables de qualité d’eau sont la demande biochimique en oxygène (BOD), les solides volatiles totaux (TVS), l’azote total (TN) et les solides dissous totaux (TDS). Pour reconstruire l’information concernant les variables aux sites abandonnés, trois méthodes sont employées pour étendre les séries chronologiques : la régression, la maintenance de variance type 3 (MOVE3), et le réseau de neurones artificiels (RNA). Une expérience empirique est conçue pour comparer ces trois techniques selon leur capacité à estimer les mesures qui préservent les principales caractéristiques des séries de la qualité de l’eau.

L’évaluation et le remaniement des sites de suivi se composent de deux étapes principales. La première étape consiste à grouper les USs similaires selon leurs attributs. Dans la seconde étape, l’échantillonnage stratifié optimal est appliqué pour déterminer le nombre optimal des USs à jauger dans chacun des groupes identifiés.

Un algorithme de groupement hybride est utilisé pour grouper les USs selon leurs attributs. Chacun des attributs est normalisé avant l'analyse afin d'éliminer la dimensionnalité et les effets d'échelle. La distance euclidienne est utilisée comme mesure de proximité pour définir la distance entre les USs et l'algorithme de Ward est utilisé pour définir les différents groupes dans un algorithme hiérarchique aggloméré. Avec l'algorithme des groupes hiérarchiques, nous considérons différents nombres de groupes et les centres de chaque groupe sont calculés et ensuite utilisés comme données pour l'algorithme de la méthode des K-moyennes « K-means ».

Dans la première étape, nous déterminons le nombre des groupes, le nombre des USs dans chaque groupe, le centre du groupe et la distance entre chaque US et son centre de groupe. Si l'objectif est d'établir un nouveau réseau de suivi, nous utilisons la stratégie optimale d'échantillonnage stratifié pour distribuer les *M* sites de suivi parmi les groupes identifiés. L'allocation optimale pour chaque groupe des USs est proportionnelle à l'écart-type de la distribution des USs au sein du groupe et au nombre de membres (USs).

Si l'objectif est d'évaluer et de redistribuer les sites de suivi de la qualité de l'eau, il est possible d'identifier le nombre optimal de sites pour chaque groupe. En comparant le nombre optimal des USs jaugées pour chaque groupe au nombre des USs déjà jaugées, trois scénarios sont possibles si le nombre des USs déjà jaugées de chaque groupe est: égal, inférieur ou supérieur au nombre optimal des USs jaugées. Si la situation suit le premier scénario, aucune mesure additionnelle ne doit être prise. Dans la situation du deuxième scénario, il est possible d'ajouter des sites, tandis que dans le troisième scénario, certains des USs déjà jaugées pourraient être abandonnées. Il est donc possible d'ajouter des sites de suivi à des USs non jaugées dans des groupes à faible suivi ou de supprimer des sites de suivi qui font partie des groupes à fort suivi. S'il est nécessaire d'ajouter des sites à des groupes qui n'ont pas suffisamment de sites de suivi, les USs non jaugées les plus éloignées du centre du groupe ont la priorité. S'il est nécessaire de supprimer des sites de groupes à trop fort suivi, on utilise un indice d'information agrégé pour déterminer les combinaisons optimales de sites à supprimer. La Figure 3 montre le flux des analyses résumées ci-dessus. La méthodologie détaillée et les résultats sont présentés dans l'article IV de la partie B.

4.4 L'estimation des caractéristiques de la qualité de l'eau aux sites non jaugés

À notre connaissance, aucun travail n'a été fait auparavant pour l'estimation des caractéristiques de la qualité de l'eau des sites non jaugés. Nous présentons deux modèles pour l'estimation des valeurs moyennes de la qualité de l'eau aux sites non jaugés. Le premier modèle est fondé sur les réseaux de neurones artificiels (RNAs) et le deuxième modèle est basé sur l'analyse de corrélation canonique (ACC) et les RNAs. Un modèle d'ensemble de RNAs est développé pour établir la relation fonctionnelle entre la valeur moyenne de la qualité d'eau et les attributs du bassin. Dans le modèle basé sur l'ACC et les RNA, l'ACC forme un espace d'attributs canonique avec des données de sites jaugés. Ensuite, une analyse de RNA est appliquée pour identifier les relations fonctionnelles entre les valeurs moyennes de qualité de l'eau et les attributs dans l'espace de l'ACC. Les deux modèles sont appliqués à 50 sous-bassins versants dans le delta du Nil en Égypte. Une procédure de validation par ré-échantillonnage « jackknife » est utilisée pour évaluer la performance des deux modèles.

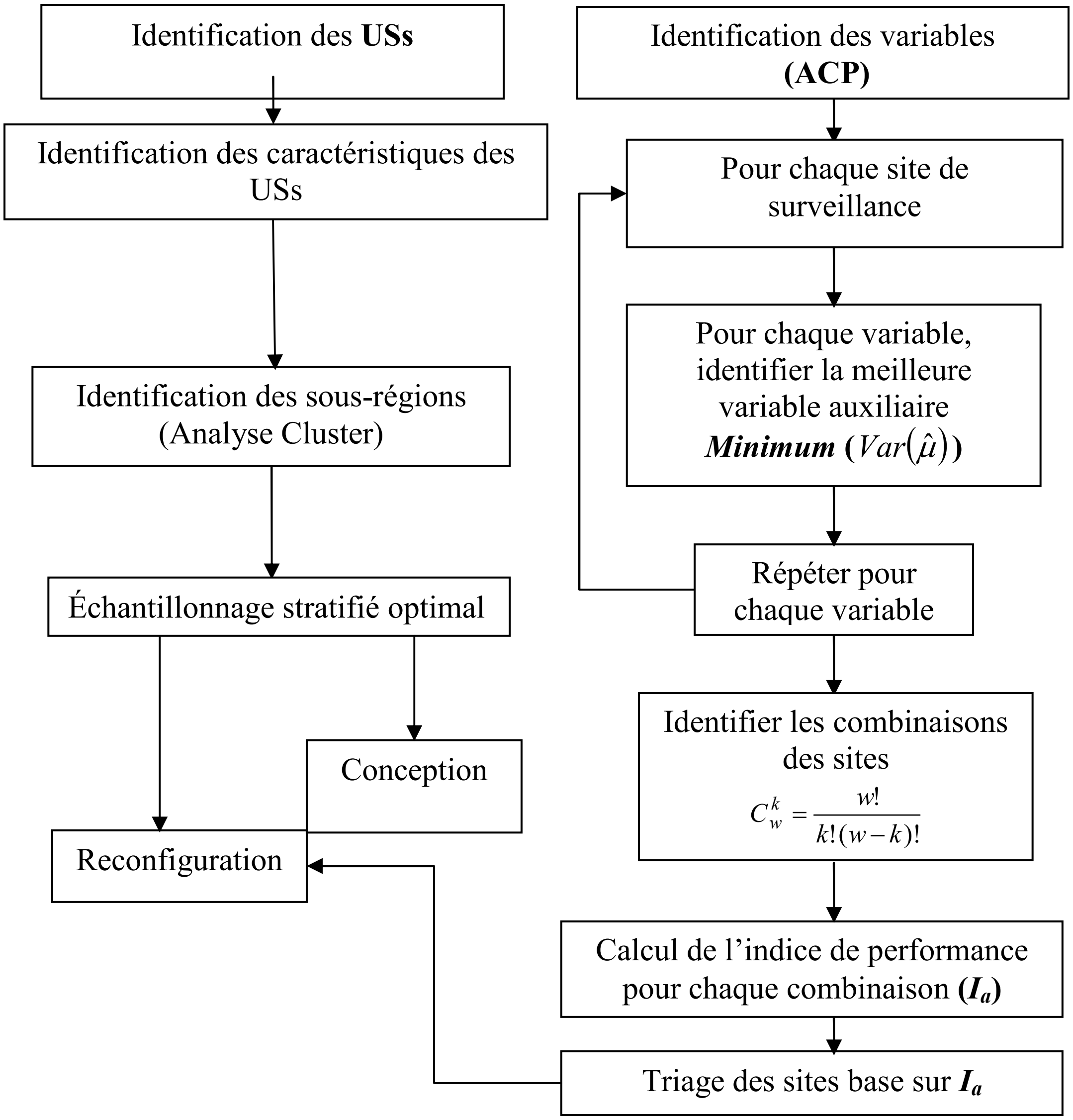

Identification des **USs**
Identification des caractéristiques des USs
Identification des sous-régions (Analyse Cluster)
Échantillonnage stratifié optimal
Conception
Reconfiguration
Identification des variables **(ACP)**
Pour chaque site de surveillance
Pour chaque variable, identifier la meilleure variable auxiliaire ***Minimum*** $(Var(\hat{\mu}))$
Répéter pour chaque variable
Identifier les combinaisons des sites $C_w^k = \frac{w!}{k!(w-k)!}$
Calcul de l'indice de performance pour chaque combinaison (I_a)
Triage des sites base sur I_a

Figure 3. Schéma de l'approche proposée pour la reconfiguration des sites d'échantillonnage

D'abord, on divise le delta du Nil en sous-bassins versants qui sont des zones représentant des USs; ces zones sont chacune drainées par un seul point sur le système de drainage. La qualité de l'eau à chacun de ces points décrit l'effet des attributs naturels et anthropiques de l'US sur l'état de la qualité de l'eau. La préparation des données se compose de deux étapes principales. Dans la première étape, neuf des attributs qui expliquent les différents effets naturels et anthropiques sont identifiés pour chaque US. Dans la seconde étape, l'ACP est utilisée pour choisir quatre indicateurs de la qualité de l'eau à partir des 33 variables mesurées. Les quatre variables de qualité de l'eau sélectionnées sont la demande biochimique en oxygène (BOD), les solides volatiles totaux (TVS), l'azote total (TN) et les solides dissous totaux (TDS).

Les modèles proposés dans cette étude sont basés sur la relation fonctionnelle entre les attributs de l'US et les variables choisies pour la qualité de l'eau à chaque US jaugée. Le premier modèle est basé principalement sur les réseaux de neurones artificiels, tandis que le second modèle utilise l'ACC et les RNA. Dans le premier modèle, on emploie un ensemble de RNAs (« ERNA ») avec une couche d'entrée, une couche cachée et une couche de sortie pour chacune des composantes de l'ERNA. Les entrées sont les attributs de l'US et les sorties sont les valeurs moyennes des indicateurs sélectionnés pour la qualité de l'eau. Une fonction de transfert de type tangente-sigmoïde est utilisée pour les nœuds de la couche cachée, tandis que pour les nœuds de sortie, la fonction de transfert est linéaire. Dans cette étude, une analyse de sensibilité est effectuée afin de déterminer le nombre optimal de nœuds cachés. En faisant varier le nombre de neurones cachés de trois à quinze, on remarque que les RNA avec sept neurones cachés fournissent l'estimation la plus précise quand ils sont appliqués pour estimer les valeurs moyennes pour les indicateurs sélectionnés pour la qualité de l'eau. Différentes tailles d'ensemble, variant de 5 à 20, ont été appliquées dans cette étude. Les résultats indiquent que l'erreur d'estimation diminue progressivement lorsque la taille d'ensemble augmente jusqu'à 15. Au-delà d'une taille de 15, il n'y a aucune amélioration de l'erreur d'estimation. Ainsi, un ensemble de taille 15 est utilisé dans le présent document. La procédure « bagging » est choisie pour générer des réseaux individuels qui comprennent les RNA et la moyenne simple est utilisée pour combiner les résultats de chaque RNA.

Dans le second modèle, le modèle ERNA dans l'espace ACC est utilisé pour établir la relation fonctionnelle entre les valeurs moyennes de la qualité de l'eau et les attributs des USs. L'ACC est utilisée pour former un espace canonique des attributs en utilisant les attributs aux les USs jaugées. Les modèles d'ensemble de RNA sont ensuite utilisés pour identifier les relations fonctionnelles entre les valeurs moyennes de la qualité de l'eau et les attributs de l'US dans l'espace d'ACC. Les modèles ERNA dans l'espace ACC ont la même structure, la même fonction de transfert, le même nombre de neurones dans la couche cachée et le même nombre de composants RNA que ceux définis pour le modèle ERNA. Les réseaux des composants dans les modèles ERNA-ACC sont produits avec l'approche « bagging », et les réseaux qui en résultent sont combinés par moyenne simple.

Une procédure de ré-échantillonnage « jackknife » est utilisée pour comparer les performances relatives des modèles ERNA et ERNA-ACC. Dans cette procédure, les valeurs moyennes de la qualité de l'eau à chaque US jaugée sont temporairement supprimées; ainsi l'US est considérée comme étant non jaugée. Ensuite, chaque modèle est calibré avec les données mesurées des autres USs. Une estimation de la moyenne de chaque variable de la qualité de l'eau est obtenue pour l'US temporairement supprimée en utilisant des modèles calibrés, puis les estimations sont comparées par rapport aux valeurs moyennes calculées à partir des enregistrements observés. Les évaluations sont réalisées en utilisant les cinq indices suivants: le critère de Nash (*NASH*), la racine carrée de l'erreur (*RMSE*), la racine carrée de l'erreur relative (*RMSEr*), le biais moyen (*BIAS*) et le biais relatif moyen (*BIASr*). Les détails concernant les deux modèles et les résultats obtenus sont présentés dans l'article V de la partie B.

5. Résultats

Les résultats de ce travail sont résumés dans les quatre sous-sections. La première sous-section résume les principaux résultats obtenus en appliquant l'approche proposée pour la rationalisation des variables de la qualité de l'eau. La deuxième sous-section résume les résultats obtenus à partir des analyses Monte-Carlo et des expériences empiriques menées pour comparer les techniques d'extension des séries chronologiques. La troisième sous-section résume les résultats obtenus en appliquant l'approche statistique proposée pour l'évaluation et la redistribution des emplacements des sites d'échantillonnage. Finalement, la quatrième sous-section présente les résultats obtenus pour l'estimation des caractéristiques de qualité de l'eau sur des sites non jaugés.

5.1 Rationalisation des variables de la qualité de l'eau

Nous identifions les groupements de variables de la qualité de l'eau hautement associées en utilisant le regroupement hiérarchique et les critères développés pour déterminer le seuil du coefficient de corrélation. En utilisant des groupes hiérarchiques et de critères mis au point pour déterminer le seuil du coefficient de corrélation, des groupes de variables fortement corrélées sont identifiés.

Les variables de la qualité de l'eau sont divisées en 24 groupes (Figure 4). Dix-huit des regroupements sont composés d'une variable unique. Ces 18 variables doivent être suivies en continu, car elles fournissent l'information qu'on ne peut pas estimer à partir d'autres variables. Les six autres groupes sont sélectionnés pour des analyses supplémentaires. Au sein de chaque groupe multivariable, chaque variable est sensée être abandonnée et sa meilleure variable auxiliaire est identifiée. Quatre groupes se composent de seulement deux variables. Dans ces cas, chaque variable fonctionne comme une variable auxiliaire pour l'autre. Une seule variable peut être abandonnée dans chacun de ces groupes. Deux groupes se composent de plus de deux variables chacun. Ici, chaque variable est donc traitée comme étant abandonnée et sa meilleure variable auxiliaire est identifiée à partir des autres variables du même groupe.

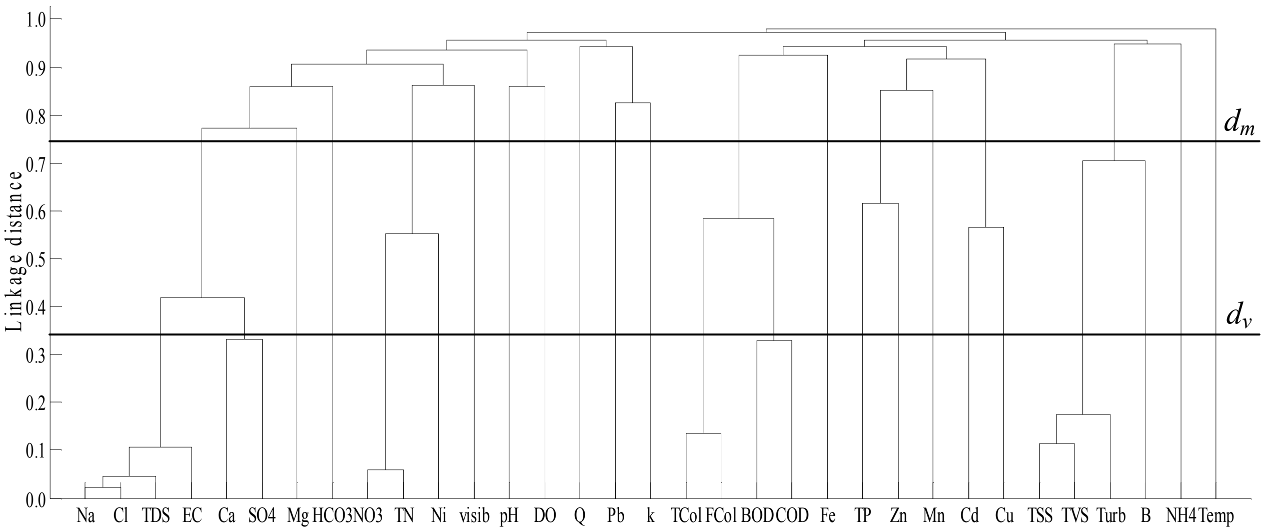


Figure 4. Arbre de groupement pour les variables de la qualité de l'eau sur le site Arin (EH16)

Si une variable doit être retirée de la liste des variables mesurées, l'indice de performance agrégé est appliqué. Les variables à abandonner sont classées en fonction des valeurs d'indice des informations agrégées. L'indice des informations agrégées fournit le rang des variables qui peuvent être abandonnées d'un point de vue statistique. Ainsi, il peut servir à accompagner d'autres critères comme information supplémentaire pour les décideurs et les concepteurs de réseau afin de choisir quelle variable à abandonner. Les autres critères peuvent inclure de préférences des intervenants ou l'importance de la variable pour des études spécifiques. Si plusieurs variables peuvent être abandonnées, l'indice des informations agrégées est appliqué pour classer les différentes combinaisons (Tableau 2).

Tableau 2. Les combinaisons de variables à éliminer

Variable éliminée	Meilleur auxiliaire	$Var(\hat{\mu})$	I_a%	Deux variables		I_a%	Trois variables			I_a%
Cl	*Na*	0.0838	2.9818	*Cl*	*TDS*	2.4406	*Cl*	*TDS*	*TVS*	2.2793
Na	*Cl*	0.0838	2.9818	*Na*	*EC*	2.4420	*Cl*	*TDS*	*Turb*	2.2801
TDS	*Na*	0.0842	2.9825	*Cl*	*EC*	2.4420	*Cl*	*TDS*	*TSS*	2.2806
EC	*TDS*	0.0877	2.9872	*Na*	*TVS*	2.4424	*Na*	*EC*	*TVS*	2.2808
Turb.	*TSS*	0.0872	2.9874	*Cl*	*TVS*	2.4424	*Cl*	*EC*	*TVS*	2.2808
TSS	*Turb.*	0.0872	2.9877	*TDS*	*TVS*	2.4428	*Na*	*Turb*	*EC*	2.2815
TVS	*TSS*	0.0880	2.9878	*Na*	*Turb*	2.4431	*Cl*	*Turb*	*EC*	2.2815
Ca	SO_4	0.0897	2.9918	*Cl*	*Turb*	2.4431	*Na*	*Turb*	*TVS*	2.2819
SO_4	*Ca*	0.0897	2.9918	*TDS*	*Turb*	2.4435	*Cl*	*Turb*	*TVS*	2.2819
COD	*BOD*	0.0924	2.9952	*Na*	*TSS*	2.4437	*Na*	*TSS*	*EC*	2.2820
BOD	*COD*	0.0924	2.9952	*Cl*	*TSS*	2.4437	*Cl*	*TSS*	*EC*	2.2820
TN	NO_3	0.1145	3.0157	*TDS*	*TSS*	2.4440	*TDS*	*Turb*	*TVS*	2.2823
Fcoli	*Tcoli*	0.1165	3.0191	*EC*	*TVS*	2.4442	*Cl*	*TDS*	*Ca*	2.2831
NO_3	*TN*	0.1145	3.0305	*Turb*	*EC*	2.4450	*Cl*	*TDS*	*SO4*	2.2831
Tcoli	*Fcoli*	0.1165	3.0324	*Turb*	*TVS*	2.4453	*Turb*	*EC*	*TVS*	2.2837
Combinaisons avec la plus haute I_a				NO_3	*Fcoli*	2.4801	*BOD*	NO_3	*Tcoli*	2.3435
				NO_3	*Tcoli*	2.5010	*COD*	NO_3	*Tcoli*	2.3442

5.2 Comparaison de quatre techniques d'extension des séries chronologiques

Dans l'expérience Monte-Carlo, 5000 essais sont générés. Ce nombre a été choisi en fonction de la pré-analyse de la convergence de l'erreur dans l'estimation des différentes statistiques. Les résultats de l'expérience Monte-Carlo indiquent que les quatre techniques d'extension des enregistrements préservent la valeur moyenne. Cependant, pour l'estimation de l'écart-type, les méthodes OLS et KTRL entraînent une sous-estimation systématique de l'écart-type, tandis que LOC et KTRL2 préservent la variance des enregistrements étendus.

Les valeurs des biais (*BIAS*) (Tableau 3) et la racine de l'erreur quadratique moyenne («root mean square error» *RMSE*) pour l'estimation de la moyenne et de l'écart-type décroissent avec des coefficients de corrélation élevés et les grandes tailles d'échantillons. Pour l'estimation de l'écart-type, les valeurs de *BIAS* indiquent que l'OLS et le KTRL sous-estiment l'écart-type, tandis que LOC et KTRL2 préservent la variance dans les registres étendus. En utilisant la technique LOC ou KTRL2, les valeurs de *RMSE* sont équivalentes et inférieures à celles correspondant aux techniques OLS et KTRL. Ces résultats peuvent indiquer que les méthodes LOC et KTRL2 évaluent précisément l'écart-type, compte tenu de la disponibilité d'un nombre suffisant d'enregistrements avec un degré élevé d'association.

Tableau 3. Valeurs de *BIAS* pour estimer la moyenne et l'écart-type pour l'expérience de Monte-Carlo

n_1	n_2	ρ	**Moyenne**				**Écart-type**			
			OLS	**LOC**	**KTRL**	**KTRL2**	**OLS**	**LOC**	**KTRL**	**KTRL2**
96	**24**	**0.5**	0.001	0.002	0.001	0.001	-0.080	0.000	-0.079	0.001
78	**48**	**0.5**	0.001	0.002	0.000	0.001	-0.166	0.001	-0.164	0.002
48	**78**	**0.5**	0.002	0.002	-0.001	-0.001	-0.260	0.004	-0.257	0.007
24	**96**	**0.5**	0.000	0.002	-0.001	0.001	-0.366	0.012	-0.361	0.015
96	**24**	**0.7**	0.000	0.000	0.000	-0.001	-0.054	-0.001	-0.053	0.000
78	**48**	**0.7**	-0.001	-0.001	-0.001	-0.001	-0.110	0.000	-0.108	0.001
48	**78**	**0.7**	0.000	0.000	-0.001	-0.001	-0.170	0.002	-0.168	0.004
24	**96**	**0.7**	-0.002	0.001	0.000	0.003	-0.232	0.009	-0.229	0.012
96	**24**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.020	-0.001	-0.020	0.000
78	**48**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.040	0.000	-0.039	0.001
48	**72**	**0.9**	0.000	0.000	0.000	0.000	-0.061	0.000	-0.059	0.001
24	**96**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.081	0.003	-0.078	0.005

Les méthodes OLS et KTRL surestiment les percentiles faibles et sous-estiment les percentiles élevés. Les méthodes LOC et KTRL2 ont tendance à réduire la partialité dans l'estimation des percentiles extrêmes. En général, pour l'estimation des percentiles, les valeurs *BIAS* (Figure 5) et *RMSE* diminuent avec des coefficients de corrélation élevés et les grandes tailles d'enregistrements.

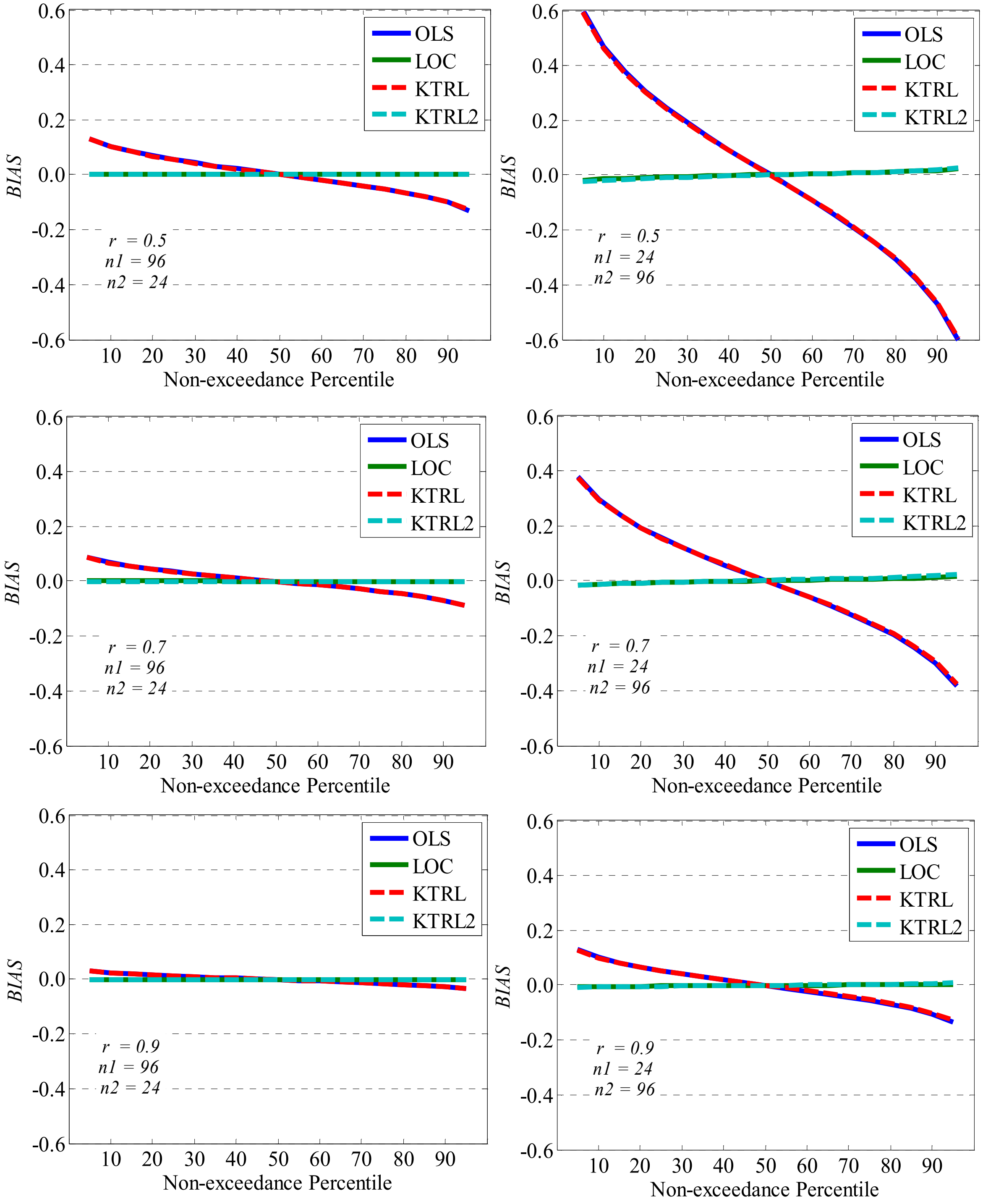


Figure 5. Valeurs de *BIAS* pour l'estimation des percentiles pour l'expérience de Monte-Carlo

Les résultats de l'expérience empirique confirment que les techniques OLS et KTRL réduisent considérablement la variance observée dans la variable dépendante, tandis que LOC et KTRL2 ont tendance à préserver la variance. Les techniques OLS et KTRL sous-estiment les valeurs de concentrations élevées et surestiment les valeurs faibles. D'autre part, les techniques LOC et KTRL2 ont tendance à réduire le biais dans l'estimation de valeurs des hautes et des basses concentrations. Les deux techniques produisent des séries étendues qui préservent relativement bien les percentiles extrêmes.

En utilisant des données réelles de qualité de l'eau, KTRL2 a une meilleure performance que LOC pour l'estimation de percentiles extrêmes. La meilleure performance de KTRL2 découle du fait que les données réelles de qualité de l'eau ne suivent pas une distribution normale en raison de la présence de valeurs extrêmes. Même après la transformation, certaines déviations de normalité peuvent exister. Cette déviation légère de la normalité ou la présence de valeurs extrêmes fait que KTRL2 est la meilleure méthode. Cette performance relativement meilleure est illustrée par KTRL et KTRL2 en comparaison avec OLS et LOC, respectivement. Elle est principalement attribuable à la robustesse de l'estimateur de pente en présence de valeurs extrêmes ou à l'écart par rapport à la normalité.

5.3 L'évaluation et la reconfiguration des sites d'échantillonnages

Les résultats de l'expérience empirique indiquent que les techniques de régression et de RNA réduisent fortement la variance observée dans les registres étendus, tandis que la technique MOVE3 préserve la variance. Les techniques de régression et RNA sous-estiment les concentrations élevées et surestiment les faibles. D'autre part, la méthode MOVE3 réduit le biais dans l'estimation des hautes et des basses concentrations. La méthode MOVE3 produit des séries étendues qui préservent relativement bien les percentiles extrêmes (Figure 6).

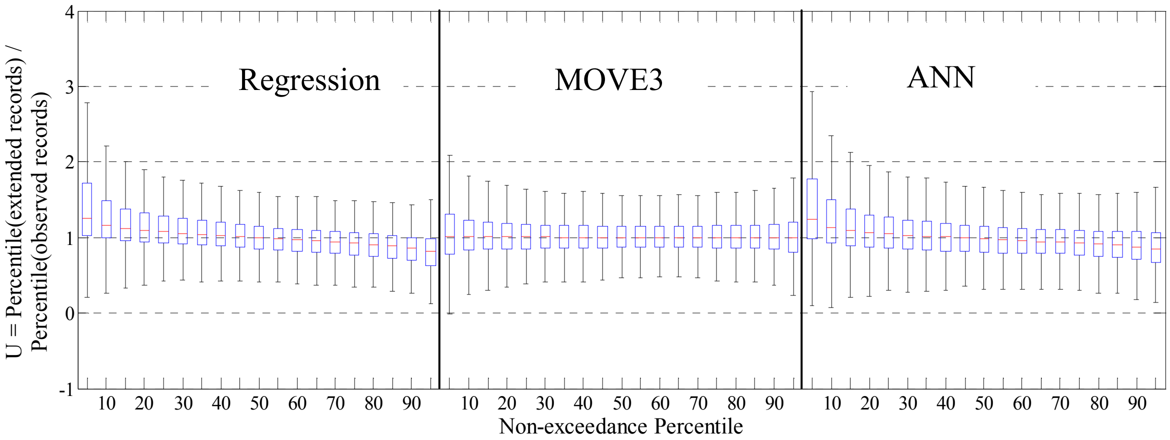


Figure 6. Box-plots de l'*U* ratio pour l'estimation de percentiles

L'algorithme de groupe hybride appliqué dans cette étude divise le Delta du Nil en 11 groupes d'USs similaires. Deux des 11 groupes ont une seule US, et un groupe avec deux USs qui sont à égale distance du centre du groupe. On a décidé de continuer à mesurer les variables de qualité de l'eau aux sites associées aux USs des groupes à US unique et une US de grappe à deux USs. L'échantillonnage optimal stratifié est ensuite appliqué pour distribuer 47 des sites de suivi parmi les huit autres groupes.

En appliquant l'échantillonnage optimal stratifié, le nombre optimal d'USs jaugées est égal au nombre d'USs déjà jaugées dans trois groupes (Tableau 4), supérieur au nombre d'USs déjà jaugées dans quatre groupes et inférieur au nombre d'USs déjà jaugées dans cinq groupes. Pour les groupes dont le nombre d'USs déjà jaugées est inférieur au nombre optimal, la distance entre les USs et le centre du groupe est utilisée pour déterminer l'US prioritaire. Pour les groupes avec plus d'USs que ce qui est requis par l'échantillonnage optimal stratifié, un index des informations agrégées est utilisé pour déterminer la combinaison optimale d'USs à abandonner. La Figure 7 montre les USs du delta du Nil, les sites à mesurer en continu, ceux à éliminer et à ajouter.

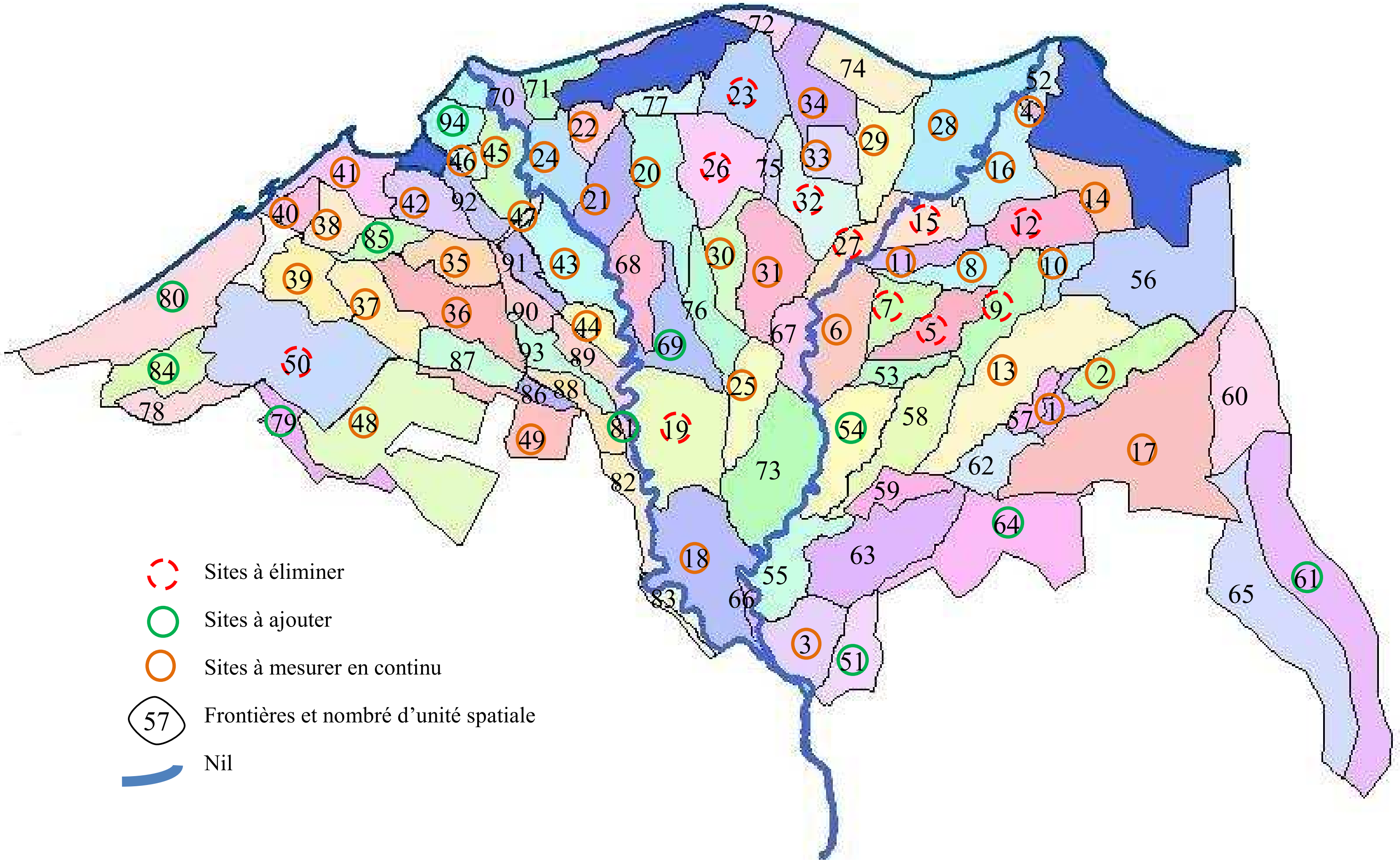


Figure 7. Les USs du delta du Nil, les sites à mesurer en continu, ceux à éliminer et à ajouter.

Tableau 4. Résultats pour l'application de l'échantillonnage optimal stratifié

Cluster	**No. USs**	**No.USs jaugées**	**Ecart-type**	**Ecart-type × no. US**	**No. optimal USs jaugées**
A	13	5	0.075	0.975	5
B	15	10	0.058	0.87	5
C	9	1	0.089	0.801	4
D	5	3	0.097	0.485	2
E	9	4	0.157	1.962	7
F	16	14	0.127	2.032	10
G	7	1	0.087	0.609	3
H	2	2	0	0	1
L	16	8	0.137	2.192	11
O	1	1	0	0	1
Z	1	1	0	0	1

5.4 Estimation des caractéristiques de la qualité de l'eau aux sites non jaugés

Plusieurs indices sont utilisés pour évaluer les modèles ERNA et ERNA-ACC. Les valeurs du critère de *NASH* par le modèle ERNA-ACC pour l'estimation des valeurs moyennes des quatre variables considérées pour la qualité de l'eau sont plus élevées que celles provenant du modèle ERNA. Cela indique que le modèle ERNA dans l'espace ACC (ERNA-ACC) est meilleur que celui de l'ERNA seul.

Les indices *RMSE* et *RMSEr* fournissent respectivement l'évaluation de l'exactitude des prédictions sur une échelle absolue et relative. Le modèle ERNA-ACCA offre de meilleurs résultats que le modèle ERNA en fonction de ces deux indices (Tableau 5). Les indices *BIAS* et *BIASr* fournissent une indication sur la tendance d'un modèle à surestimer les valeurs ou à les sous-estimer. L'analyse fondée sur l'indice *BIAS* suggère que le modèle ERNA produit des résultats légèrement meilleurs que le modèle ERNA-ACC pour l'estimation des valeurs moyennes de la BOD et des TVS. Toutefois, le *BIASr* indique que le modèle ERNA-ACC fournit de meilleurs résultats que le modèle ERNA pour l'estimation des valeurs moyennes des quatre variables considérées. Cela indique que les erreurs obtenues en utilisant le modèle ERNA sont plus symétriques autour de zéro, mais qu'elles ont plus de dispersion que celles obtenues avec le modèle ERNA-ACC.

Tableau 5. Résultats de la validation Jackknife pour la performance des deux modèles

Métrique	Variables	ERNA	ERNA-ACC
NASH	BOD	0.62	0.84
	TVS	0.68	0.78
	TN	0.61	0.72
	TDS	0.80	0.82
RMSE (mg/l)	BOD	7.95	5.15
	TVS	2.85	2.37
	TN	2.26	1.92
	TDS	451.71	424.41
RMSEr (%)	BOD	15.09	8.60
	TVS	31.42	23.42
	TN	21.24	12.87
	TDS	28.35	18.02
BIAS (mg/l)	BOD	-0.02	0.08
	TVS	0.35	0.41
	TN	0.47	0.15
	TDS	127.74	84.36
BIASr (%)	BOD	2.46	1.80
	TVS	8.78	7.44
	TN	6.95	2.72
	TDS	12.31	10.19

Selon les critères *NASH*, *RMSE*, *RMSEr* et *BIASr*, le modèle ERNA-ACC a donné un rendement sensiblement meilleur que le modèle ERNA. Par conséquent, l'application de modèles ERNA dans l’espace des attributs canoniques peut améliorer les performances des modèles ERNA dans l’espace originel des attributs.

6. Conclusions

Dans cette thèse, une revue de littérature à jour des méthodes statistiques utilisées pour l’évaluation et la reconfiguration des réseaux de suivi de la qualité des eaux de surface est effectuée. Deux nouvelles approches statistiques ont été proposées pour l’évaluation et la reconfiguration des réseaux de suivi de la qualité des eaux de surface. La première approche est proposée pour la rationalisation des variables de la qualité de l’eau. La deuxième approche est proposée pour l’évaluation et la redistribution des sites d’échantillonnage. Une nouvelle technique d’extension des séries chronologiques est proposée. Finalement, deux nouvelles approches statistiques sont proposées pour estimer les caractéristiques de la qualité de l’eau sur des sites non jaugés.

La nouvelle approche proposée dans cette thèse pour l’évaluation et la sélection des variables optimales surmonte les insuffisances de la méthode conventionnelle de corrélation-régression. On conclut que l’approche proposée permet d’identifier, d’une manière systématique et objective, la combinaison optimale des variables de la qualité de l’eau à mesurer en continu et les variables qui peuvent être abandonnées. Enfin, on conclut que l’approche proposée pourrait être un outil de décision utile pour la sélection optimisée des variables de la qualité de l’eau. Cette approche peut être appliquée pour la rationalisation des variables de la qualité de l’eau de tout type de réseaux de suivi, tels que: les réseaux de suivi de la qualité de l’eau de surface, les eaux souterraines, des lacs et des milieux marins.

La nouvelle technique d’extension des séries chronologiques (KTRL2) proposée dans cette thèse est comparée à trois autres méthodes: OLS, LOC et KTRL. La technique KTRL2 combine l’avantage de LOC en réduisant le biais dans l’estimation de la variance et la robustesse de KTRL en présence de valeurs extrêmes.

Les expériences Monte-Carlo et empiriques indiquent que les techniques OLS et KTRL ne peuvent pas fournir des enregistrements avec les variances appropriées ou avec les formes de distribution appropriées. L’évaluation des biais de moments et des percentiles indiquent que les techniques LOC et KTRL2 sont plus efficaces que les techniques OLS et KTRL, respectivement. L’estimation de la pente KTRL2 à l’aide de percentiles des variables dépendantes et indépendantes pendant la période d’enregistrements simultanés, on a prouvé que les estimations de la variable dépendante possédant des propriétés semblables à celles attendues de répartition de la variable dépendante ont été obtenues. L’expérience empirique montre que KTRL2 possède des propriétés plus avantageuses que LOC en présence de valeurs extrêmes.

Lorsque plusieurs estimations doivent être générées et des calculs effectués sur les probabilités de dépassement, telles que les probabilités de dépasser une règle ou une norme pour la qualité de l’eau, des inférences qui deviennent dépendantes de la distribution de probabilité des données estimées sont arrêtées. Dans ces cas, il faut utiliser les méthodes LOC et KTRL2, plutôt que OLS et KTRL, pour générer des données. La nouvelle technique KTRL2 est supérieure si aucune transformation ne peut produire une quasi-normalité en raison des distributions à queues lourdes et de présence de valeurs extrêmes.

La nouvelle approche proposée pour l'évaluation et la reconfiguration des sites de suivi de la qualité des eaux de surface surmonte plusieurs lacunes existant dans les approches utilisées. En conclusion, l'approche proposée permet de distribuer spatialement les points de contrôle pour la mise en place d'un nouveau réseau de suivi de la qualité de l'eau. Elle peut être appliquée pour développer ou réduire un réseau de surveillance fonctionnel. Cette approche peut évaluer et identifier, de manière systématique et objective, les sites de suivi à échantillonner en permanence, les sites à abandonner et les sites à ajouter. Cette approche est applicable uniquement pour les réseaux de suivi de la qualité des eaux de surface. Elle peut être appliquée à différentes configurations de réseaux de suivi des eaux de surface. Cependant, les attributs de sous-bassin doivent être choisis sur la base des conditions physiographiques, climatiques, et hydrologiques des bassins suivis.

Deux modèles pour l'estimation des valeurs moyennes de la qualité de l'eau aux sites non jaugés ont été présentés. Le premier modèle est fondé sur les réseaux de neurones artificiels (RNAs) et le deuxième modèle est basé sur l'analyse de corrélation canonique (ACC) et les RNAs. La validation par la méthode de rééchantillonnage «jackknife» évalue la performance des deux modèles. Les résultats indiquent que les deux modèles fonctionnent de manière acceptable. L'application de modèles ERNA dans l'espace des attributs ACC peut considérablement améliorer la performance des modèles ERNA dans l'espace des attributs d'origine. L'utilisation de ces modèles en parallèle aux activités de suivi permet de fournir aux décideurs une information de qualité de l'eau à chaque unité spatiale dans le delta du Nil.

7. Recommandations pour les travaux futurs

L'approche proposée pour de rationalisation des variables de la qualité de l'eau peut être modifiée en utilisant la régression multiple, où plus d'une variable auxiliaire peut être utilisée pour la reconstitution des informations sur une variable à éliminer.

Parallèlement aux approches proposées, une analyse des coûts ou d'un indice d'information peut être introduite pour traiter du compromis entre le nombre des variables de la qualité de l'eau à mesurer, le nombre d'emplacements du suivi de la qualité de l'eau et la fréquence d'échantillonnage.

Les modèles développés ici sont conçus pour estimer les valeurs moyennes pour la qualité de l'eau pour quatre variables sélectionnées sur des sites non jaugés dans le Delta du Nil en Égypte. La modification des méthodes présentées dans le présent document serait nécessaire pour estimer d'autres paramètres statistiques et tenir compte des variables non stationnaires.

Le développement des méthodes qui permettent l'estimation d'une série chronologique détaillée de la qualité de l'eau des sites non jaugés représente une orientation de recherche future très importante.

8. References

[1]. Abdel-Gawad, S.T., Kandil, H.M. and Sadek, T.M. (2004). Water scarcity prospects in Egypt 2000-2050, in: Marquina (ed.) Environmental Challenges in the Mediterranean 2000-2050, Dordrecht: Kluwer Academic Publishers, 187 - 203.

[2]. Abu-salama, M. S. M. (2007). Spatial and temporal consolidation of drainage water quality monitoring networks, Ph.D. dissertation, Universität Lüneburg, Fakultät III, Umwelt und Technik, Germany.

[3]. Alley, W.M. and Burns, A.W. (1983). Mixed-station extension of monthly streamflow records. Journal of Hydraulic Engineering, 109 (10), 1272 - 1284.

[4]. Bartram, J. and R. Balance (1996). Water Quality Monitoring: A practical guide to the design and implementation of freshwater quality studies and monitoring programmes. Published on behalf of UNEP and WHO, Taylor & Francis, 383 p.

[5]. Chapman, D. (1996). Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring. Chapman & Hall, London.

[6]. Dijkman, J.P.M. (1993). Environmental Action Plan of Egypt, A Working Paper on Water Resources. Directorate of General International Cooperation, Ministry of Foreign Affairs, the Netherlands, 116 - 127.

[7]. Dixon, W. and B. Chiswell (1996). Review of aquatic monitoring program design. Water Res. 30(9), 1935 - 1948.

[8]. DRI (Drainage Research Institute) - MADWQ, (1998). Monitoring and analysis of drainage water quality in Egypt, Interim Report, Cairo.

[9]. Ellis, J.C. and Lacey, R.F. (1980). Sampling: defining the task and planning the scheme. Water Pollution Control, 79, 452 - 467.

[10]. Engelman, R. and Le Roy, P. (1993). Sustaining water, population and the future of renewable water supplies. Population Action International, Population and Environment Program, Washington, D.C., 302 - 318.

[11]. Frenken, K., 2005. Irrigation in Africa in Figures, Aquastat Survey (2005). Food & Agriculture Org, Rome, Italy, 88 p.

[12]. GAO, General Accounting Office (2000). Water quality: Key EPA and State decisions limited by inconsistent and incomplete data. GAO Report Number GAO/RCED-00-54, GAO, Washington, DC, p.78.

[13]. GAO, General Accounting Office (2004). Watershed management: better coordination of data collection efforts needed to support key decisions. Report Number GAO-04-382, GAO Washington, DC, p. 155.

[14]. Harmancioglu, N.B. (1984). Entropy concept as used in determination of optimum sampling intervals. Proceedings of Hydrosoft (1984), International conference on hydraulic engineering software, Portoroz, Yugoslavia, 99 - 110.

[15]. Harmancioglu, N.B. and Alpaslan, M.N. (1992). Water quality monitoring network design: a problem of multi-objective decision making. Water Resources Bulletin, 28 (1), 179 - 192.

[16]. Harmancioglu, N.B., Alpaslan, N. and Singh, V.P. (1992). Design of water quality monitoring networks. Geomechanics and Water Engineering in Environmental Management, A.A. Balkema Publishers, Rotterdam, ch. 8, 267 - 296.

[17]. Harmancioglu, N.B., O. Fistikoglu, S.D. Ozkul, V.P. Singh and M.N. Alpaslan (1999). Water Quality Monitoring Network Design. Kluwer Academic Publishers, Dordrecht, the Netherlands, 290 p.

[18]. Hirsch, R.M. (1982). A comparison of four streamflow record extension techniques. Water Resources Research, 18(4), 1081 - 1088.

[19]. Horton, R.E. (1945). Erosional Development of Streams. Geological Society Am. Bull., 56, 281 - 283.

[20]. Khalil, B.M., Abdel-Gawad, S.T., Abdel-Rashid, A. and Morsy, A.M. (2004). Sampling frequency assessment for the drainage water quality monitoring in Egypt, Proceedings, The International IWA Conference, AutMoNet, 19-20 April, Vienna, Austria, 85 - 92.

[21]. Krige, D.G. (1951). A statistical approach to some basic mine valuation problems in the Witwatersrand. Journal of chemical, Mettalurgical and Mining Society of South Africa 52, 119.

[22]. Lettenmaier, D.P. (1976). Detection of trends in water quality data from records with dependent observations. Water Resources Research, 12, 1037 - 1046.

[23]. Lettenmaier, D.P., D.E. Anderson and R.N. Brenner (1984). Consolidation of a Stream Quality Monitoring Network. Water Resources Bulletin, 20(4): 473 - 481.

[24]. Mace, A.E. (1964). Sample-size determination. Reinhold, New York, USA, 226 p.

[25]. MacNeil, V.H., A.G. McNeil, and W.A. Poplawski (1989). Development of water quality monitoring system in Queensland, in: Ward R.C., J.C. Loftis and G.B. McBride (eds.), Proceeding, International Symposium on the Design of Water Quality Information Systems, Fort Collins, CSU Information Series no. 61, 73 - 86.

[26]. Matalas, N.C. and Jacobs, B. (1964). A correlation procedure for augmenting hydrologic data, *U.S. Geol. Surv. Prof. Pap.*, 434-E, E1-E7.

[27]. Matheron, G. (1963). Principles of geostatistics. Economic Geology 58, 1246 - 1266.

[28]. MWRI, Ministry of water resources and Irrigation (1997). Review of Egypt's Water Policies, Strengthening the Planning Sector Project, Ministry OF Water Resources and Irrigation, Cairo, Egypt.

[29]. NAWQAM, National Water Quality and Availability Management Project. (2001). Evaluation and Design of Egypt National Water Quality Monitoring Network, Technical Report No.: WQ-TE-0110-005-DR, NAWQAM, National Water Research Center, Cairo, Egypt.

[30]. Ongley, E.D. and E.B. Ordonez (1997). Redesign and modernization of the Mexican water quality monitoring network. Water International, 22(3), 187 - 194.

[31]. Sanders, T.G. and Adrian, D.D. (1978). Sampling Frequency for River Quality Monitoring. Water Resources Research, 14, 569 - 576.

[32]. Sanders, T.G., R.C. Ward, J.C. Loftis, T.D. Steele, D.D. Adrian and V. Yevjevich (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[33]. Strobl R.O., P.D. Robiliard, R.D. Shannon, R.L. Day and A.J. McDonnell (2006). A Water quality monitoring network design methodology for the selection of critical sampling points: Part I, Environmental Monitoring and Assessment, 112, 137 - 158.

[34]. Strobl, R.O. and P.D. Robillard (2008). Network design for water quality monitoring of surface freshwaters: A review, Journal of Environmental Management, 87, 639 - 648.

[35]. Tirsch, F.S. and J.W. Male (1984). River basin water quality monitoring network design: options for reaching water quality goals, in: T.M. Schad (ed.). Proceeding of Twentieth Annual Conference of American Water Resources Associations, AWRA Publications, 149 - 156.

[36]. Ward, R.C., J.C. Loftis and G.B. McBride (1990). Design of Water Quality Monitoring systems, Van Nostrand Reinhold, New York, USA, p 231.

[37]. Ward, R.C., J.C. Loftis, K.S., Nielsen and R.D. Anderson (1979). Statistical evaluation of sampling frequencies in monitoring networks. J. of WPCF, 51(9), 2292 - 2300.

[38]. Wetering, B.G.M. and S. Groot (1986). Water Quality monitoring in the state-managed waters of the Netherlands. Water Research, 20(8), 1045 - 1050.

[39]. Whitfield, P.H. (1988). Goals and data collection design for water quality monitoring. Water Resources Bulletin, AWRA, 24(4), 775 - 780.

[40]. Wolf, P. (2000). Irrigated agriculture in Egypt - Notes of an external observer, Proceedings Symposium "Sustainable Agriculture and Rural Development in Egypt" Witzenhausen, University of Kassel, Germany.

[41]. Zhou Y. (1996). Sampling frequency for monitoring the actual state of groundwater systems. Journal of Hydrology, 180, 301 - 318.

PART B: ARTICLES

Article I. Statistical Approaches Used To Assess and Redesign Surface Water Quality Monitoring Networks

Statistical Approaches Used To Assess and Redesign Surface Water Quality Monitoring Networks

B. Khalil [1,2] and T.B.M.J. Ouarda [2]

[1] Irrigation and Hydraulics department, Faculty of Engineering, Helwan University, Cairo, Egypt

[2]Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE,
University of Québec, Québec City, Canada

Published in the Journal of Environmental Monitoring, 2009, vol 11 (11), 1915 - 1929

Abstract: An up-to-date review of the statistical approaches utilized for the assessment and redesign of surface water quality monitoring (WQM) networks is presented. The main technical aspects of network design are covered in four sections, addressing monitoring objectives, water quality variables, sampling frequency and spatial distribution of sampling locations. This paper discusses various monitoring objectives and related procedures used for the assessment and redesign of long-term surface WQM networks. The appropriateness of each approach for the design, contraction or expansion of monitoring networks is also discussed. For each statistical approach, advantages and disadvantages are examined from a network design perspective. Possible methods to overcome disadvantages and deficiencies in the statistical approaches that are currently in use are recommended.

Key words – sampling, aquatic, water quality, monitoring network, design, expansion, contraction

1 Introduction

Freshwater is a finite resource that is essential for agriculture, industry and human existence (Bartram and Balance, 1996). Water demand for industrial water supply, irrigation, and hydropower generation is ever increasing with world development. Water of adequate quantity and quality is essential for sustainable development. Water quality is a term used to describe the chemical, physical, and biological characteristics of water with respect to its suitability for a particular use.

Water quality is affected by a wide range of natural and anthropogenic influences. Natural processes (hydrological, physical, chemical and biological) may affect the characteristics and concentration of chemical elements and compounds in freshwater. The following are examples of natural processes:

- Hydrological processes: Dilution, evaporation, percolation, leaching, suspension and settling;
- Physical processes: Volatilization, adsorption, desorption and diffusion;
- Chemical processes: Photodegradation, acid/base reactions, oxidation/reduction (redox) reactions, dissolution of particles and precipitation of minerals;
- Biological processes: Decomposition of organic matter, bioaccumulation and biomagnification.

Aside from natural processes, there are also anthropogenic impacts that affect water quality, such as human-induced point and nonpoint pollution sources, introduction of xenobiotics and alteration of water quality due to water use and river engineering projects (Chapman, 1996).

Assessment of water resources requires knowledge and full understanding of the processes affecting both water quantity and water quality (Harmancioglu et al., 1999). In order to understand the process dynamics of a watershed, a well-designed WQM network is required. Monitoring programs help elucidate the various processes affecting water quality, as well as provide water managers with the necessary information for water resources management in general and water quality management in particular.

WQM programs encompass a variety of activities that include the following: definition of the monitoring purpose and desired information, monitoring network design, sampling protocol design, laboratory analysis, data verification and storage, and data analysis. Several aspects of a monitoring program, such as sampling procedures, sample handling and storage and laboratory analysis, must be performed by properly trained staff to ensure the utility of the generated data. Thus, in the last few decades, researchers have been increasingly turning their attention to the design of monitoring networks. The first step towards establishment of a monitoring network is the definition of the monitoring objectives. The design of the monitoring network is the translation of objectives into a protocol that describes the variables to be measured, as well as the location, timing and duration of monitoring.

In the 1960s and 1970s, WQM programs were developed to describe the general status of water quality. These early efforts typically involved arbitrary approaches that lacked a consistent or logical design strategy (Sanders, et al., 1983). Sampling locations and frequencies were often determined by convenience or by other subjective criteria. Once the network was established, there was often no assessment of the effectiveness of the monitoring design (Tirsch and Male, 1984; Ward et al., 1990). Inadequacy of a proper network design methodology often results in water quality data collected with little analysis or ultimate purpose (GAO, 2000, 2004). For long-term operational WQM networks, what was previously a design problem has become an assessment and redesign problem (Harmancioglu et al., 1999).

Given that water quality is a complex topic, statistical approaches can make a significant contribution. There are several statistical methods used to assess the performance of WQM networks. A significant amount of research has been directed toward evaluating current design procedures and investigating effective means to improve the efficiency of existing networks (Harmancioglu et al., 1999). However, there is presently no established strategy or methodology for designing monitoring networks, particularly regarding the location of sampling stations (Strobl et al., 2006). Thus, a logical and consistent design methodology that allows for more efficient and effective data collection and, consequently, more useful outputs, is required. Such an approach would permit not only better water pollution control recommendations and better allocation of financial resources, but also a better understanding of the ecosystem under study (Strobl and Robillard, 2008).

This paper presents a comprehensive review of the various statistical approaches that have been proposed for the assessment and redesign of surface WQM networks. Each approach is explained, and its advantages and disadvantages from a network design perspective are discussed. The appropriateness of each approach for the design, contraction or expansion of monitoring networks is reviewed. The following definitions are used in this document: design of a network refers to the choice of optimum monitoring sites, sampling frequency and variables measured; expansion of an existing network is an increase in the number of monitoring locations, the sampling frequency, or the variables measured; and finally, contraction of an existing network refers to a decrease in the number of monitoring locations, samples taken or variables measured.

The remaining sections of this paper are as follows. Section 2 presents the monitoring objectives. Section 3 describes the statistical approaches used for the selection of water quality variables. Section 4 discusses the statistical approaches used for the assessment of sampling frequencies. The statistical approaches used for the assessment and redesign of monitoring locations are presented in section 5. Finally, concluding remarks are provided in section 6.

2 Monitoring objectives

Monitoring objectives should define the information output expected from the network. Failure to adequately specify the desired information leads to failure of the network itself (Harmancioglu et al., 1992). The monitoring objectives are the foundation upon which the monitoring program is built. Three main challenges may arise when identifying monitoring objectives: selecting from multiple potential objectives, stating the objective and transforming objectives into statistical questions. One major obstacle is that objectives are often described in global terms, rather than specific, precise, and clear-cut statements (Harmancioglu et al., 1999).

Definition of the monitoring goals is essential for both the design and operation of a network. The literature provides a wide range of possible objectives for water quality monitoring (Ward et al., 1979; Sanders et al., 1983; Whitfield, 1988; Zhou, 1996; and Harmancioglu et al., 1999). The following is a summary of the most common monitoring objectives as presented in the literature, some of which are only applicable for either long-term or short-term monitoring networks:

- Identify spatial and temporal trends;
- Assess compliance with standards;

- Facilitate impact assessment studies;
- Determine suitability for various water uses;
- Execute general surveillance;
- Evaluate various control strategies;
- Estimate mass transport in rivers; and
- Facilitate water quality modeling or other specific research activities.

Ward (1989) introduced a statistical perspective regarding monitoring objectives. He viewed monitoring as a form of statistical sampling, and thus he indicated that the objectives must be specified in statistical terms. The collected data are expected to allow for reliable statistical analyses and the eventual transfer of data into useful information. Ward (1989) described a systematic approach to defining objectives, which includes the following steps:

- Specify objectives of water quality management (general objectives), for example, preservation of water quality;
- Specify objectives of the monitoring network within the framework of water quality management (specific objectives), for example, identification of trends;
- Give a statistical description of the specific objectives, for example, specification of the precision required or accepted error.

Such an approach would foster efficient selection of variables, sampling locations and sampling frequencies. Definition of the statistical methods is also helpful for the assessment of the performance of the network and the quality of the resultant information. Whitfield (1988) recommended identifying the monitoring objectives on a site-by-site basis, designing a sampling strategy for each objective and periodically evaluating the adequacy of the design. The periodic assessment of WQM programs should take into account not only changes in environmental conditions, but also shifts in management objectives. A monitoring program's main goal(s) may be translated into monitoring objectives that vary on a site-by-site basis: What are we trying to measure at this site? What are the water quality variables to be measured? What is the appropriate statistical tool to use in order to obtain the desired information? Answering these three questions will help identify the water quality variable(s) to be measured at each site as well as the sampling strategy needed to meet the monitoring objective(s) based on the chosen statistical approach.

The problem of climate change has become increasingly recognized as a major environmental concern but has not yet been addressed in WQM network design. Climate change refers to any long-term significant change in the variability or mean state of the atmosphere. Since aquatic ecosystems are highly influenced by numerous exchange processes between the atmosphere and hydrosphere, changes in any of the principal atmospheric variables may have a significant effect on water quantity and quality. For example, higher air temperatures are expected to affect the hydrologic cycle and melt snowpacks more rapidly. At the same time, surface water temperatures will rise due to temperature exchange with the warmer atmosphere. In general, changes in water temperatures as well as changes in the timing, intensity, and duration of precipitation can affect water quality. Higher water temperatures reduce dissolved oxygen levels, which in turn may have an effect on aquatic life. Increases in the frequency and intensity of rainfall may result in increased pollution and sedimentation in surface water streams.

Furthermore, a sea level rise may also affect freshwater quality by increasing the salinity of coastal rivers and causing saltwater intrusion. Thus, assessment of the changes in climate conditions and their impact on water quality should be an important monitoring objective, and the design of monitoring networks should be adjusted to incorporate this new objective.

Water quality monitoring is a costly undertaking, as it requires sufficient equipment, instrumentation, specialized personnel and adequate funding. Consequently, availability of resources, as well as other relevant political, legal, and social constraints, should be considered when specifying monitoring objectives. Inadequate monitoring is a poor investment.

3 Water quality variables

The quality of a water body is usually described by a set of interrelated physical, biological and chemical variables. Water quality can be defined in terms of one variable to hundreds of compounds. Thus, choosing appropriate variables to characterize surface water quality is a highly complicated issue (Sanders et al., 1983; Ward et al., 1990; Harmancioglu et al., 1999). Many recognize that it is not possible to measure everything in the environment, and that some logical means of selecting variables should be part of water quality information systems (Ward et al., 1990).

Selection of the set of water quality variables to be measured (design) or addition of new variables to those already being measured (expansion) involves subjective issues, such as monitoring type, monitoring objectives, basin characteristics and available budget. Proposed methods for the assessment and re-selection of variables fall under the category of contraction, which aims to reduce the number of variables being measured. Two main approaches for selecting monitoring variables have been proposed in the literature: the correlation and regression method and principal component analysis (PCA).

3.1 Correlation and Regression analysis

The correlation and regression approach is based on three steps. The first step is to assess the level of association among the variables by correlation analysis. If high correlation exists between variables, it is an indication that some of the information produced may be redundant and perhaps one can stop measuring some. The second step is the selection of water quality variables to be discontinued and variables to be continuously measured. This step is based on some subjective criteria, such as the significance of the variable, the presence of the variable in local or international standards, cost of analysis and so forth. In the third step, regression analysis is used to reconstitute information about the discontinued variables using auxiliary variables from those continuously measured. Thus, the original list of variables being measured becomes partially measured, and partially estimated using regression analysis.

During the 1970s, several studies showed that concentrations of major ionic constituents can be related to specific conductance (McKenzie, 1976; Hawkinson et al., 1977; Briggs and Ficke, 1978). Specific conductance can serve as an indicator variable from which concentrations of major ionic solutes can be determined, in cases where suitable regression functions can be found (Sanders et al., 1983). Yevjevich and Harmancioglu (1985) investigated the transfer of information by bivariate correlations among daily water quality variables observed along the Upper Potomac River Estuary, USA. The objective was to determine pairs of variables that exhibited strong relationships to each other in order to identify which variables that should continue to be sampled and those that can be estimated. Harmancioglu and Yevjevich (1986, 1987) studied the effects of removing deterministic components (trends, periodicity and stochastic dependence) in order to see the effects of these characteristics on the amount of information transfer. They concluded that some basic similarities in deterministic components are the main contributors to transferred information.

The search for correlation among water quality variables or between water quantity and water quality variables represents the first step in defining the dependent and independent variables in a regression model. In a multiple linear regression, the problem of multicolinearity may exist. Nevertheless, when this problem occurs, it means that at least one predictor is redundant with other predictors. In this case, a ridge regression or a stepwise regression may be used rather than a simple multiple linear regression.

3.2 Principal Component Analysis

PCA is one of a number of factor extraction methods. Since Hotelling (1933) introduced PCA, it has been used for data interpretation, pattern recognition, dimensional analysis, and multicolinearity detection. PCA is a data technique rather than a statistical technique, meaning that many of the commonly applied statistical inference tests are not directly applicable (Hintze, 2001). PCA transforms a set of correlated variables into a smaller set of uncorrelated variables, called principal components (Jobson, 1992).

PCA can be applied to a set of water quality variables to discover the variables that form coherent subsets that are relatively independent of one another. Variables that are correlated with one another, but largely independent of other subsets of variables, are combined into one component. The components are thought to reflect the underlying processes that created the correlations between variables. Mathematically, PCA produces several linear combinations of observed variables, and each linear combination forms a component. The components summarize the patterns of correlations in an observed correlation matrix. The factor loading matrix obtained from PCA reflects the characteristics of the extraction procedure, which maximizes the variance extracted from the data in each successive component.

Karpuzcu et al. (1987) used PCA to reduce the number of variables to be directly observed. It is claimed that such methods give better estimates of the most representative water quality variables than those obtained by conventional correlation analysis. For the Richibucto drainage basin, New Brunswick, Canada, St-Hilaire et al. (2004) recommended that water quality variables that were found to explain most of the variance by PCA be monitored more closely, as they are key elements for the understanding of variability in water quality. Similarly, Ouyang (2005) employed PCA to identify important water quality variables in the assessment of the surface WQM network of the lower St. Johns River in Florida, USA.

3.3 Discussion

The main advantage of the correlation and regression approach over the PCA approach is that the former allows for the reconstruction of information regarding discontinued variables. However, these approaches are mainly based on the assumption of a linear relationship among water quality variables, but the relationship between physical, biological and chemical variables may be nonlinear. Thus, the mutual information measure and Artificial Neural Networks (ANNs) may be used instead of linear correlation and regression analyses. Mutual information is a measure of a nonlinear dependence or the amount of redundant information between two variables. The ANN is more flexible than regression models in its ability to capture the relationships among water quality variables and requires less prior knowledge of the system under study.

If the structure of the data is inherently linear (for example, if the underlying distribution is normal), PCA is an optimal feature extraction algorithm; however, data that have a nonlinear lower-dimensional structure will not be detectable by PCA (Monahan, 2000). In the early 1990s, a neural-network-based generalization of PCA to the nonlinear feature extraction problem was introduced in the chemical engineering literature by Kramer (1991), who referred to the resulting technique as nonlinear principal component analysis (NLPCA). The fundamental difference between PCA and NLPCA is that PCA only allows for linear mapping, while NLPCA allows for nonlinear mapping (Hsieh, 2004). Kramer's NLPCA has been applied in various fields, including chemical engineering (Kramer, 1991), sea surface temperature (SST) fields (Monahan, 2001; Hsieh, 2001), Northern Hemisphere atmospheric variability (Monahan et al., 2000; 2001), and Canadian surface air temperature (Wu et al., 2002). However, the NPLCA has not yet been applied to water quality multivariate datasets.

Another deficiency in both the correlation/regression and PCA approaches is the absence of a criterion to identify the combination of variables to be continuously measured and those to be discontinued. Consider the case in which budget cuts require w variables to be discontinued.

Which w variables among the m variables currently being measured should be selected? The number of possible combinations of variables to discontinue is given by the binomial coefficient $C(m, w)$. For each combination, a factor can be computed according to which the combinations may be ranked (Ouarda et al., 1996). It is important to identify the type of desired information. For example, the inverse of the variance of the mean value could be used as a surrogate for the amount of information pertaining to a particular water quality variable. The total cost of laboratory analysis for each combination of variables could also be used as a criterion to rank possible combinations. Ouarda et al. (1996) proposed an information performance index based on the variance of the mean value for the contraction of the hydrometric network in Ontario, Canada. Such a procedure would allow for identification of the best combination of variables to be continuously measured and those to be discontinued.

These assessments can also serve to resolve the trade-off between the number of water quality variables measured and the sampling frequency and locations. Thus, the decision will be either to discontinue more variables in favor of keeping more monitoring locations and/or increasing the sampling frequency, or to keep more water quality variables, while decreasing the number of monitoring locations and/or the sampling frequency.

4 Sampling Frequency

Sampling frequency is a very important aspect of WQM network design, as it affects not only data utility but also operation costs. By sampling too frequently, the obtained information is redundant and expensive, while infrequent sampling may limit the precision. Statistical methods proposed for the assessment and calculation of sampling frequencies are directly related to the monitoring objectives and the data analysis methods. The following subsections describe different approaches used to assess and select sampling frequencies. The suitability of each method for network contraction or expansion is examined. The last subsection discusses deficiencies in these strategies from a design perspective and possible areas for further research.

4.1 Trend analysis

Lettenmaier (1976) proposed a method to determine optimum sampling intervals based on the parametric trend test, where the required sampling frequency corresponds to a specified power of the trend test. This is known as the effective sample method (ES). The word "effective" indicates the maximum number of independent samples that can be collected on an annual basis. The basis of ES was developed by Bayley and Hammersley (1946) and was later expanded upon by Lettenmaier (1976) and Sanders and Adrian (1978).

This approach is divided into two steps. The first step is to define the maximum number of samples that can be collected per year in order to avoid autocorrelation, or at least to decrease its effect. The second step is to estimate the length of record (number of years) needed to reliably detect trends at specified confidence levels and test powers. The sampling frequency required to obtain roughly independent samples is based on the autocorrelation between measurements with different time lags. The equation developed by Bayley and Hammersley (1946) that is used to determine the number of effective samples from a dependent time series is given by:

$$\frac{1}{n^*} = \frac{1}{n} + \frac{2}{n^2} \sum_{k=1}^{n-1} (n-k)\rho_k \qquad (1)$$

where n^* is the effective number of samples per year, n is the number of samples taken per year, k is the number of lags between samples and ρ_k is the autocorrelation coefficient for lag k. An effective sample size per year can then be determined for each of the water quality variables. If the data obtained from the current sampling frequency are totally independent ($\rho_k = 0$), then equation (1) will be reduced to $n^* = n$, which is the existing sampling frequency. Since n^* is either smaller than or equal to n, this method is considered to be a contraction approach.

Lettenmaier (1976) and LaChance et al. (1989) furthered the application of the ES method using the standard error and trend magnitude of a time series to determine the length of record needed to reliably determine trend magnitudes at a specified statistical significance level. Lettenmaier (1976) presented equation (2), which was developed from the power function of a classical t-test:

$$N^* = \frac{12(t_{\alpha/2,(n-2)} + t_{\beta,(n-2)})^2}{\left(\dfrac{Tr}{\sigma_\varepsilon}\right)^2} \qquad (2)$$

where N^* is the total number of independent samples needed, t is the Student's "t" statistic, α is the type I error risk of inferring a difference when none exists, β is the type II error risk of failing to detect a trend when in fact it is present, Tr is the absolute value of the change at the beginning and ending predicted values along a regression line for a period of study and σ_ε is the standard deviation of the residuals. N^* (total number of samples) from equation (2) can be divided by n^* (number of samples per year) from equation (1) to determine the number of sampling years needed (sampling duration) to reliably detect an existing trend with the desired confidence and power.

The ES method, which was used by Lettenmaier (1976) and LaChance et al. (1989), makes the general assumption that the data are normally distributed. However, in both of those studies some non-normality in the data was accepted. The autocorrelation coefficients of equation (1) must be generated from a stationary time series, which is accomplished by removing trends and seasonality (LaChance et al., 1989). Recently, Yue and Wang (2004) examined the ability of ES to eliminate the influence of autocorrelation when using the non parametric Mann-Kendall (MK) rank correlation test by Monte-Carlo simulation. Their results showed that when no trend exists within the time series, ES can effectively limit the effect of serial correlation. In contrast, if a trend exists within the time series, the existence of the trend will contaminate the estimate of the magnitude of sample autocorrelation, and the ES computed from such a series cannot properly eliminate the effect of autocorrelation. However, if the ES is computed from a sample autocorrelation estimated from a de-trended series, then ES can still effectively reduce the influence of autocorrelation.

“Step trends” in quality may be detected from discontinuous monitoring as a before-and-after approach in a single watershed, by testing for differences in mean constituent concentrations or mean indicator values between two or more sequential time periods (Lettenmaier, 1976; Sanders et al., 1983; NRCS, 2003). The step trend approach is best used when the watershed in question is paired with a control (Green, 1979; NRCS, 2003) because of the potential confounding effects of climatic conditions and annual variability. The formula applied by Lettenmaier (1976) was modified by Zar (1996) to allow the user to specify the power to detect a difference between time periods, as follows:

$$n = \frac{2 \times S_P^2 (t_{\alpha,2(n-1)} + t_{\beta(1),2(n-1)})^2}{\delta^2} \tag{3}$$

where n is the size of each sample, S_P^2 is the estimate of pooled variance between the two time periods and δ is the minimum detectable difference between the time periods. Values from the t-distribution are for a one-tailed β and a one-tailed or two-tailed α, depending on the type of hypothesis to be tested.

4.2 Confidence Interval

Sanders and Adrian (1978) and Sanders et al. (1983) recommended that the confidence interval about the mean be used as the main criterion for the selection of sampling frequency. The purpose is to select a sampling frequency that yields an estimate of the mean ($\bar{x}$) within a prescribed degree of accuracy (confidence limits) using the following equation:

$$n \geq \left[\frac{(t_{\alpha/2})\, s}{E} \right]^2 \tag{4}$$

where n is the number of samples, t is the Student's "t" statistic, s is the sample standard deviation and E is half the confidence interval (expected error).

At this point, the sampling frequency can be defined for a specific water quality variable at a specific monitoring location. The second step is to allocate the number of samples to different monitoring locations using proportional sampling techniques, which is often based on variability (Ward and Nielsen, 1978; Sanders et al., 1983; Chen and Cheng, 1989). In practice, autocorrelation may be present, so that part of the information contained in one measurement is also contained in the next measurements. In this case the variance of $\bar{x}$ is:

$$Var(\bar{x}) = \frac{\sigma^2}{n}\left[1 + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\rho_k\right] \qquad (5)$$

where σ^2 is the variance of the population of measurements and n is the number of samples (Loftis and Ward, 1979). When $Var(\bar{x})$ given by equation (5) is substituted into equation (4), we obtain a quadratic equation that can be solved for *n*. Gilbert (1987) obtained an approximate expression by ignoring the term in the quadratic equation that has n^2 in the denominator:

$$n = D\left[1 + 2\sum_{k=1}^{n-1}\rho_k\right] \qquad (6)$$

where $D = \left[\frac{(t_{\alpha/2})\,s}{E}\right]^2$.

The relationship between the confidence interval and the number of samples as shown in equation (4) becomes theoretically valid if the variance of the stationary component computed from different sampling intervals stabilizes after a certain sampling interval. After this sampling interval, the variance becomes independent of the sampling interval and any change in the number of samples will only affect the expected error. Using the confidence interval about the mean as the main criterion to assess and redesign the sampling frequency may lead to either an increase or decrease in the sampling frequency.

Whitfield (1983) applied this approach to identify the sampling frequency, which allows the detection of a 10 percent difference between annual mean concentrations for one of the monitoring locations on the Yukon River in Canada. High variability in water quality records resulted in large estimations of the required frequencies, which undermines the applicability of such undertaking. Tokogz (1992) and Harmancioglu and Tokgoz (1995) assessed sampling frequencies of the WQM network in the Prosuk River basin, Turkey. Their results showed that this method is not applicable to water quality time series with a short duration of observations and a large number of missing values, due to the difficulty in meeting the underlying assumptions of the method. It is difficult to meet the assumption that states that the variance of the stationary component calculated from different sampling intervals stabilizes after a certain sampling interval in case of limited water quality records with large gaps.

4.3 Harmonic analysis

Zhou (1996) proposed an approach aimed at defining the sampling frequency in terms of periodicity based on harmonic analysis, h_t, as follows:

$$h_t = A_0 + \sum_{j=1}^{k} \left[A_j \cos(2\pi f_j t) + B_j \sin(2\pi f_j t) \right] + \varepsilon_t \qquad (7)$$

where A_0 is a constant, A_j and B_j are the harmonic series coefficients, j is the index of the j^{th} harmonic at time *t*, and k is the total number of harmonics to be fitted to the data. k is equal to *N/2* for an even sample size and *(N-1)/2* for an odd sample size, where *N* is the sample size. $(A_j{}^2 + B_j{}^2)^{1/2}$ and $f_j = j/N$ are the amplitude and frequency for the j^{th} harmonic. ε_t is an independent random variable with a mean of zero and a variance of σ_ε^2. The details concerning the estimation of the constant and the harmonic coefficients are described in Zhou (1996).

The frequency of the harmonics chosen to fit the data series must be restricted to $0 \le f_j \ge (1/2\Delta t)$, where the frequency $(1/2\Delta t)$ is called the Nyquist frequency (fn). The Nyquist frequency is the minimum frequency that will faithfully sample a signal, which is twice the maximum frequency of the signal. Thus, fn gives the minimum sampling frequency required. Harmonic analysis may reveal the highest frequency of significant periodic fluctuations in the real time series; let this frequency be fs. Then the sampling frequency fp should be more than twice fs, in order to capture significant frequent fluctuations.

Harmonic analysis is carried out on a de-trended time series. Thus, the first step is to determine if there is a trend in the time series, and if a trend is detected, then it should be removed before performing harmonic analysis.

The cyclic variability present in water quality time series should be considered before a final selection of sampling frequencies is made. Three main cyclic variations are commonly found in water quality data: the annual cycle, weekly cycle and diurnal variation. Annual cycles may reflect the hydrology (high and low flow seasons), the weather (winter and summer) or the activities (cultivation seasons) present in the watershed. Reaches immediately downstream of wastewater treatment plants often exhibit weekly cycles. Highly regulated streams with dams for power production also exhibit a weekly cycle. Since power production and flow are reduced during weekends, water quality may be degraded as well. Some water quality variables such as dissolved oxygen (DO) and biochemical oxygen demand (BOD) commonly exhibit diurnal variability. Such variables may require several samples per day to define their periodicity. Detection of high-frequency variations in national and regional long-term monitoring networks would be very costly. Thus, if detection of weekly or diurnal variability is desired, implementing a parallel monitoring program (at selected sites) is recommended. In general, collecting monthly samples imply that temporal variations on a time scale less than a month are not the major concern. Similarly, if a special survey is carried out involving intensive monitoring for short period (e.g. two years), indicates that time variations over several years are not the major concern.

4.4 Binomial test

Extreme values or excursions are most commonly defined as observations that exceed any preset limit, such as a stream standard, drinking water standard, compliance limit, action limit, or control limit (Ward et al., 1990). Mace (1964) and Ward et al. (1990) described an approach for estimating the sample size required to control the risk of type I and type II errors when assessing the proportion of time that a criterion is exceeded. This approach is applicable when individual values are determined to be either "above" or "below" a criterion (nominal scale). The number of exceedances of a limit always follows the binomial distribution regardless of the distribution the data follow (Ellis and Lacey, 1980). An iterative approach is used to estimate the number of exceedances (e) in a specified number of samples (n) that meets the following conditions:

$$1 - B(e; n, 1 - p^*) \leq \alpha \tag{8}$$

$$\beta = B(e; n, 1 - p) \tag{9}$$

where B is the cumulative binomial probability, p^* is the proportion of time that concentrations must be within the criterion and p is a selected proportion associated with an unacceptable frequency of exceedances. The results of the calculation using equations (8) and (9) can be displayed graphically, enabling the appropriate values of e and n to be read for given values of α, β, p^* and p (Miller and Ellis, 1986). Mace (1964) and Klotz and Meyer (1985) used an arcsin normal approximation to estimate the required sample size:

$$n = \left[\frac{z_\alpha + z_\beta}{2 \arcsin \sqrt{p} - 2 \arcsin \sqrt{p^*}} \right]^2 \tag{10}$$

where z_α and z_β are quantiles from the standard normal distribution. By fixing α, p^* and p, n increases as β decreases and vice versa. Thus, this approach could be used to increase or decrease the number of samples. Hapuarachchi and Macpherson (1992) described an approach for estimating sample size when data are autocorrelated. When the measurements follow an autoregressive process of order one, or AR(1), as is commonly the case with water quality data, the authors suggested using an approximate sample size given by:

$$n \approx \left(\frac{z_{\alpha} + z_{\beta}}{z_{p^*} - z_p} \right)^2 \left(\frac{1 + \hat{\rho}_1}{1 - \hat{\rho}_1} \right) \qquad (11)$$

where $\hat{\rho}_1$ is the estimated lag-1 autocorrelation, and z_{p^*} and z_p represent the p^* and p quantiles of the standard normal distribution, respectively. Smith et al. (2003) compared error rates for the binomial test and the variable acceptance sampling approach. Acceptance sampling by variable is a major field in statistical quality control, where the view is that a group of manufactured items is to be accepted or rejected (Smith et al., 2003). In the acceptance sampling by variable approach, the value of the mean that results in the distribution having a standard with an associated probability of exceedance is calculated. The mean of the observed data is then compared with the derived mean. Smith et al. (2003) showed that, as autocorrelation increases, the required sample size increases. They also concluded that the variable acceptance approach is superior to the binomial method in the sense that it requires smaller sample sizes to achieve the same error rates.

4.5 Bayesian analysis

Bayesian analysis has not yet been proposed for the assessment of sampling frequency. However, it has been proposed for the assessment of compliance with standards (Smith et al., 2001; McBride and Ellis, 2001). The following is a brief description of how Bayesian analysis is applied to the assessment of compliance as proposed by McBride and Ellis (2001).

In Bayesian analysis, the exceedance probability is regarded as a continuous variable about which a statement of a confidence is desired. The obtained data are used to update a prior belief in order to obtain a posterior belief, which is known as confidence of compliance (McBride and Ellis, 2001). The prior belief, as well as the posterior information, are stated as a probability density function (*pdf*). Thus, the required probabilities can be obtained by integrating the posterior *pdf* up to *x*, where *x* is the probability of exceedance allowed in water quality standards. The equation for the posterior *pdf* is as follows:

$$h\langle x \mid e, n\rangle = \left[\frac{L\langle e \mid n, x\rangle}{\int_0^1 L\langle e \mid n, x\rangle\, g(x)\, dx}\right] g(x) \qquad (12)$$

where *e* is the number of exceedances in *n* samples, $h\langle x|e,n\rangle$ is the posterior *pdf* of *x* for given values of *e* and *n*, and *g(x)* is the prior *pdf* of *x*. $L\langle e|n,x\rangle$ is the likelihood function for any *n* and *x* given by:

$$L\langle e|n,x\rangle = {}^{n}C_{e}\; x^{e}\; (1-x)^{n-e} \qquad (13)$$

where ${}^{n}C_{e}$ is the binomial coefficient. Thus, the posterior probability density is given by:

$$h\langle x \mid e, n\rangle = \left[\frac{x^{e}\; (1-x)^{n-e}}{\int_0^1 x^{e}\; (1-x)^{n-e}\; g(x)\, dx}\right] g(x) \qquad (14)$$

The probability of compliance or Confidence of Compliance (CC) is calculated from the distribution function:

$$CC = F\langle x \leq X \,|\, e, n \rangle = \int_0^x h\langle x \,|\, e, n \rangle \, dx \qquad (15)$$

where *X* is the value of *x* where the tested hypothesis is true. The confidence of failure (*CF*) is given as *CF = 1- CC*. Noting that the likelihood function follows a binomial distribution, McBride and Ellis (2001) chose four different priors, or *g(x),* to demonstrate the influence of the prior assumptions on the confidence results. Two reference priors (uniform and Jeffreys') and two other user-defined priors were used. The results showed that using the Bayesian technique (with Jeffreys' reference prior) makes compliance rules less onerous, particularly for smaller numbers of samples, while still affording the desired degree of protection (McBride and Ellis, 2001).

Smith et al. (2001) showed that if the sample size is smaller than 20, the binomial method cannot adequately control the error rates. Given sufficient prior information, Bayesian methods may be used with smaller sample sizes to help select the error rates of concern. The selection of the prior *pdf g(x)* can be difficult when there is little information, and the analysis becomes subjective. Support for these probabilities can come from previous reports and surrounding sites. This would lead to a more objective formulation of priors and would make the Bayesian approach a sound alternative (Smith et al., 2001).

4.6 Regression analysis

Tirsch and Male (1984) addressed the temporal design of networks by multivariate linear regression models. Monitoring precision, as described by the corrected regression coefficient of determination, is expressed as a function of sampling frequency. This method can be applied to examine sampling frequencies lower than the one currently in use. Such an approach aims to either continue with the current sampling frequency or to reduce the number of required samples if no significant change appears in the coefficient of determination. Thus, using regression analysis in this framework is considered a contraction approach. Tirsch and Male (1984) applied this method using daily specific conductance records of the Shoshone River basin monitoring network and daily stream flow records of the Millers River basin in north central Massachusetts, USA.

4.7 Entropy concept

Entropy theory provides a measure of information contained in a set of data or the distribution of a random variable. It was first developed by Shannon (1948). In environmental and water sciences, entropy theory has been applied to a wide spectrum of problems (Singh, 1997). Harmancioglu (1984) introduced the entropy concept as a way to determine the optimal sampling intervals in water quality monitoring. The entropy principles in this case were applied to determine the information content of stochastic dependent variables in order to identify the optimum sampling intervals with respect to time. The application of entropy principles to the design of WQM networks was subsequently extended to the assessment of network efficiency and cost effectiveness (Harmancioglu and Alpaslan 1992). The following is a brief description of how the entropy technique is used to assess sampling frequency. The marginal entropy $H(X)$ is first determined for the water quality variable under assessment (Equation 16), where the total range of X is divided into intervals of equal size (Δx).

$$H(X) = (M / 2)\ln 2\pi + (1/2)\ln\left|C\right| + M / 2 - M\ln(\Delta x) \qquad (16)$$

where M is the number of variables, $|C|$ is the determinant of the covariance matrix and Δx is the class interval size for the M-variables. In this case, M is replaced by one and the covariance matrix C converts directly into the variance. In the second step, the water quality variable X is described as the subseries X_k for time lags k = 0, 1, …, K. A maximum lag K is assumed and the subseries X_0, X_1, …, X_k are considered separate variables. The total entropy is computed for each time lag using equation (16), where M in this case is the number of lagged subseries and the conditional entropy $H\langle X_0 | X_1,...,X_k \rangle$ is computed using equation (17). The conditional entropy quantifies the remaining entropy of a random variable given that the value of the second random variable(s) is known.

$$H\langle X_0 | X_1,........., X_k \rangle = H(X_0, X_1,....X_k) - H(X_1,........X_k) \tag{17}$$

In the third step, the transinformation is computed for lag k = 1, 2, 3, …, K, using equation (18). Transinformation or mutual information is an entropy quantity that measures the redundant information between variables. It is the difference between the total entropy and the conditional entropy of the dependent variable (Harmancioglu, 1981).

$$T(X_0, X_k) = H(X_0) - H\langle X_0 | X_k \rangle \tag{18}$$

where $T(X_0, X_k)$ is the transinformation between lag zero (the original data set) and lag k. High transinformation means that there is high information redundancy between successive measurements. When no further reductions are observed in the transinformation at lag $k = K$, it is considered that the lags beyond this point do not contribute significantly to the reduction of uncertainty. Thus, the entropy concept is considered a contraction approach.

Although not specifically a parametric approach, the data must be normal or lognormal for the multivariate case. The entropy-based approach is not well-suited for other skewed distributions with multivariate data (Harmancioglu and Alpaslan, 1992; Yang and Burn, 1994). The entropy method requires approximating a continuous probability distribution function with a discrete function. This discretization was shown by Harmancioglu et al. (1985) to be critical to the value of the entropy provided by the analysis, and has the potential to change the decision arising from entropy-based analysis. The choice of the time interval for which the water quality values are assigned (period used for time-averaging) also appears to have a significant impact on the accuracy of the results.

4.8 Semivariogram

The semivariogram method is based on the fundamental work of Krige (1951) and Matheron (1963), and is the first stage in a geo-statistical analysis. The semivariogram is a graphical representation of how the similarity between values varies as a function of the distance (and direction) or time separating them. The theoretical semivariogram is a plot of one distance, or "lag" separating pairs of points (x-axis). The general equation for the semivariogram is:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)-1} \left[Z(x_{i+h}) - Z(x_i) \right]^{2} \qquad (19)$$

where $\gamma(h)$ is the semivariance, $n(h)$ is the number of points separated by the time or distance h (the "lag") and $\left[Z(x_{i+h}) - Z(x_i)\right]$ is the difference between the values of variables separated by the lag h.

The procedure begins by plotting an experimental semivariogram using the available data, and a theoretical model is fitted to the resulting plot. The y-axis represents the variance and the x-axis represents the sampling frequency. Several models are frequently used to describe the theoretical semivariogram, including the Gaussian, linear and exponential models. The curve fitted to the data usually consists of two distinct parts: an initial rising segment and a horizontal flat segment (Figure 1). The point at which the curve levels off is used to calculate both the sill (y-axis) and the range of correlation (x-axis). The sill is the maximum semivariance exhibited by the data set and the range of correlation is the lag (sampling frequency) at which the sill value is reached. Pairs of points separated by a distance or time greater than the range of correlation are considered to be spatially or temporally uncorrelated. A sample can be considered representative of the time defined by the effective sampling frequency, which is the range of serial correlation.

(Figure 1)

The semivariogram was introduced to define the effective range of correlation and consequently to assess the sampling frequency. In order to have a serially independent data series, the sampling frequency is selected based on the effective range of correlation. Dowdall et al. (2005) utilized geostatistical techniques in the optimization and design of sampling regimes to monitor temporal fluctuations in the levels of technetium in the Norwegian Arctic marine environment.

Khalil et al. (2004) used the semivariogram to assess sampling frequency in the Nile Delta of Egypt; one year of daily records of specific conductance (EC) and dissolved oxygen (DO) were used, and the effective range of correlation as an indicator of autocorrelation was chosen as the design criterion.

4.9 Discussion

Calculation of sampling frequencies by statistical methods often results in optimization of sampling around one purpose or goal (Ward et al., 1990). Water quality monitoring cannot address every information need through one data collection procedure (Whitfield, 1988), yet the system must attempt to meet several information goals simultaneously. Whitfield (1988) suggested that different sampling frequencies should be used for different monitoring objectives in order to maximize the information gain. Zhou (1996) assessed different sampling frequencies for different monitoring objectives, and selected the highest sampling frequency in order to fulfill multiple objectives. Ward et al. (1990) recommended evaluating logical alternative sampling frequencies in terms of their ability to satisfy multiple objectives, rather than optimize for any one objective.

The approaches described in this paper are used to assess the sampling frequency of a specific water quality variable at a specific monitoring location. However, in practice, water quality monitoring programs usually measure several water quality variables at several monitoring locations. Ward (1978) and Sanders et al. (1983) used proportional sampling to address this issue. Proportional sampling consists of distributing a pre-identified total number of samples among monitoring locations and/or variables based on a given weight for each variable and location. The weights are typically assigned as a function of population density, historical variance, mean variable values, and flow (Ward, 1978). However, the number of samples selected may not satisfy all the monitoring objectives at all the monitoring locations for all the measured variables. Thus, the consequences of applying proportional sampling should be evaluated. The assessment of the sampling frequency for different objectives simultaneously may be a possible area for further research.

Another challenge in determining optimal sampling frequencies is related to the time scale of monitoring. For example, to estimate the mean value it is important to set the monitoring objective whether to estimate the long-term annual mean or specific interval mean (year, month or day). An explicit consideration of scale in monitoring system design and data analysis is important to produce meaningful statistical information (Loftis et al., 1991). Loftis et al. (1991) discussed how the time scale in water quality monitoring can have a marked effect on the information extracted from monitoring data. As demonstrated by simulation results, a process that is stationary over a long period of time may contain short-term runs that are important from a management standpoint. In a given time series, runs above or below the long-term mean might be described either as autocorrelation or short-term trends (Loftis et al., 1991). These authors also showed that long-term mean would not be determined with high precision by intensive sampling when observations are autocorrelated. However, autocorrelation works to reduce rather than increase the variance of error in estimating specific-interval (annual, for example) means for a given number of equally spaced observations. They concluded that the distinction between autocorrelation and trends in a time series is scale dependent. They also determined that autocorrelation has a much different effect on the estimation of long-term means than it does on the estimation of specific-period means. Thus, the importance of scale should be considered while assessing the sampling frequency.

Several studies have demonstrated the usefulness of continuous monitoring using data loggers equipped with water quality sensors. Continuous monitoring improves the resolution of the signal being measured and is able to capture transient events (Harriman et al., 1990; Wade and Whitfield, 1994; Whitfield and Wade, 1992; Weatherley and Ormerod, 1991; Whitfield and Dalley, 1987). These technologies will likely transform our view of catchment processes, by allowing observation of hydro-chemical changes at temporal resolutions that are orders of magnitude finer than before (Kirchner et al., 2004). In order to determine the input-output mass balance of a particular catchment, weekly or monthly chemical monitoring may not be adequate. Pollutant loads are estimated by multiplying pollutant concentrations in water by discharge rates. Fixed interval sampling (weekly or monthly) may miss important events, especially in streams in which pollutant concentrations and/or discharge levels are highly variable, or where stream flow responds quickly to precipitation events. In such a flashy stream, it is almost impossible to synchronize grab sampling with peak flows, making continuous monitoring essential for accurate estimation of flux (Whyte and Kirchner, 2000). Continuous monitoring avoids under- and over-sampling storm events, and thus also avoids under- and over-estimation of cumulative flux (Whyte and Kirchner, 2000). Weekly or monthly monitoring programs cannot capture short-term chemical dynamics, which closely reflect hydrological processes. Thus, high-frequency chemical observations will be essential in developing, calibrating, and validating the next generation of catchment models (Kirchner et al., 2004).

5 Monitoring locations

The choice of monitoring locations is an important aspect in the design of a monitoring network. If the location is not representative of the water body, the data interpretation and presentation becomes inconsequential (Sanders et al., 1983; Ward et al., 1990). Early practices in water quality sampling focused on sites with easy access, without any systematic approach to selection of sampling locations (Harmancioglu et al., 1999). Over time, the number of sites has increased to include stations at points of interest such as those located upstream and downstream of highly industrialized or highly populated areas, areas with point pollution sources, or areas of intensive land use (Tirsch and Male, 1984). Later, various methodologies were proposed for the selection of both the number and location of sampling stations. The following subsections describe the different approaches used to assess and redesign sampling locations, organized by the statistical analysis method employed. Each approach is described, and its suitability for the design, contraction or expansion of the number of monitoring locations is discussed. The last subsection discusses the deficiencies in the proposed approaches from a design perspective as well as possible ways to overcome those drawbacks.

5.1 Stream order hierarchical approach

The stream order hierarchical approach proposed by Sanders et al. (1983) is based on using stream ordering to describe a stream monitoring network. The stream ordering procedure (Horton, 1945) assigns each unbranched small tributary the order of one, a stream made up of only first order tributaries the order of two, and so on. Later, Sharp (1970; 1971) used stream ordering procedures to measure the uncertainty involved in locating the source of pollutants observed at the outlet of a network. Sanders et al. (1983) followed Sharp's procedure (1970;

1971) by selecting sampling locations on the basis of the number of contributing tributaries, pollutant discharge or BOD loads. This approach systematically locates sampling sites so as to divide the river network into sections that are equal with respect to the number of contributing tributaries, discharge or pollutant loading. Three levels of design criteria are considered by Sanders et al. (1983) for location sampling: the macrolocation, the microlocation and representative sampling.

The macrolocations are the river reaches that will be sampled within the river basin, and they are defined using the stream ordering approach. The microlocation is a point within the reach that is completely well-mixed relative to outfalls or point sources of pollution. The microlocation provides representative sampling of the reach, and requires an analysis of mixing. The representative locations are the points in the river cross-section that provide a lateral profile of the stream.

In addition, the mixing process is an important process in case of pollutant released into a water body. Regardless of the pollutant nature, the mixing process consists of a one-dimensional movement induced by turbulent mean flow (advection) and three-dimensional spreading action produced by the turbulent flow components (diffusion) (Abbott and Basco, 1989). Numerical advection-diffusion models are intended to make predictions by solving the advection-diffusion equation (Abbott and Basco, 1989; Sloan and Pender, 1998). The advection-diffusion equation uses, time, velocity and the diffusion coefficient with spatial variability. Typically, these models use medium time steps (days to months) and are generally limited to small spatial scales. Details of different methods used to calculate vertical, transverse, longitudinal mixing distances and pollutant transport are available in Fischer et al. (1979), Abbott and Basco (1989), and Czernuszenko and Rowinsk (2005).

When no data are available, the stream order hierarchical approach can be applied to select monitoring locations. However, the role of tributary order may be overly emphasized. Assigning each unbranched small tributary the order of one assumes that each of these tributaries has the same contribution to the system. This assumption may be valid if all tributaries drain the same area and have the same activities within these areas. Moreover, in the presence of a point source of pollution (such as industrial effluent), assigning an order of one to the source assumes that it has the same weight as an unbranched tributary, yet it is completely different in the type and loading of pollution. Using the discharge of the tributaries or pollution sources instead of order gives a true weight to each tributary. In operational monitoring networks, pollutant loading can be used to assess the existing monitoring locations. Although assessing pollutant loads may produce a rather different system of locations, the hierarchical approach works well in initiating a network when no data or very limited amounts of data are available.

5.2 Regression approach

Spatial design of water quality networks is also performed using regression techniques. Tirsch and Male (1984) proposed a multivariate linear regression model where the adjusted regression coefficient of determination between sampling locations is considered a measure of monitoring precision. In this approach, a regression model is performed in which each monitoring location is considered the dependent variable, and different combinations of the remaining locations are considered the independent variables. The adjusted coefficient of determination is obtained for each model, and the monitoring precision changes with the addition or deletion of some number of stations within the network. A high coefficient of determination indicates that a high degree of redundancy exists and that the station selected as a dependent variable might not be needed. Tirsch and Male (1984) applied this approach using daily specific conductance records from the network monitoring the Shoshone River basin, USA.

The advantage of using regression is that it allows data regarding discontinued locations to be reconstructed. However, assessment using regression utilizes only one water quality variable to assess the spatial distribution of locations. Usually, in monitoring networks, several water quality variables are measured. Thus, if this approach is applied to all the measured variables sequentially, as many designs as the number of variables could be obtained.

5.3 Entropy concept

Another statistical approach to spatial sampling is the entropy-based method. Using this method, the amount of transinformation between sampling locations is determined based on the degree of uncertainty (Harmancioglu and Alpaslan, 1992; Harmancioglu et al., 1999). Dependence between sampling locations results in reduced entropy, or uncertainty, between the locations. If the dependence is consistent over time, one or more of the sampling sites may be discontinued with a minimal loss of information.

For a multivariate case and assuming multivariate normal data, the joint entropy of X is defined by equation (16). This equation results in a single value that expresses the joint entropy over the whole network for M locations. Marginal entropy for each sampling location is computed using the same equation by letting $M = 1$, adding the subscript m to X to represent each sampling location, and substituting the variance (σ^2) for $|C|$.

The entropy method has been applied to multivariate data, but only one variable is analyzed at a time (Harmancioglu et al., 1994; Ozkul et al., 1995, 2000). However, handling multivariate data should be conducted simultaneously rather than sequentially. Another deficiency in this method is the discretization problem, as explained in subsection (4.7). This approach reflects the amount of redundant information among monitoring locations, and may give an indication to increase the number of locations. However, the method cannot define where exactly new stations should be located.

5.4 Multivariate data analysis

Different multivariate data analysis techniques have been employed for the redesign of water quality monitoring locations. These include principal component analysis (PCA), cluster analysis (CA), and discriminant analysis (DA). Cluster analysis (CA) is an exploratory data technique used to group similar observations into clusters, where the within-cluster variance is minimized and the between-cluster variance is maximized (Peck et al., 1989; Jobson, 1992). In earlier years, the validity of clustering techniques was questioned because of the lack of inferential tests offered by many other statistical techniques (Baker and Hubert, 1975; Wong, 1982). However, using additional tests such as estimating the bootstrap confidence intervals (Peck et al., 1989), performing a discriminant analysis on the final clusters (Jobson, 1992), or combining hierarchical and non-hierarchical clustering techniques (Jobson, 1992) can provide validation for the chosen clusters.

Discriminant analysis (DA) is a multivariate statistical method that is somewhat similar to CA except for one major difference: in DA, the groups are already known and the user is testing the ability of a set of variables to discriminate between the various groups. DA has been developed for both parametric (Rao, 1973) and nonparametric (Rosenblatt, 1956; Parzen, 1962) cases. The nonparametric form is also known as the k-nearest neighbor method. The nonparametric form is appropriate when the data do not have a multivariate normal distribution. DA can be used as a validation measure to test the discriminating ability of the clusters from a previous CA. Details concerning PCA, CA and DA are available in reference text books (e.g. Tabachnick and Fidell, 1996; Jobson, 1992).

Odom (2003) used PCA, CA, and DA to assess and redesign the WQM network in the Great Smoky Mountains National Park, Tennessee, USA. He used the average pollutant values at each location in the PCA. Then, he used the first three principal components in a cluster analysis to define similar locations, and finally applied DA to verify the groups found in CA. Ouyang (2005) applied PCA to evaluate the effectiveness of the surface WQM network of the lower St. Johns River in Florida, USA, where the variables evaluated in PCA are the monitoring locations. He also used the water quality mean values at each location as variables for different cases. In these two studies, the authors summarized the information content of the data series using mean values only, and did not consider all of the available data.

Often, analysts tend to standardize the variables before performing PCA. In water quality data, only one water quality variable may be the main source of the variability. This variable will figure disproportionately in the covariance matrix and consequently in the principal components. The first principal component will largely represent that variable, and the other principal components will have negligibly small variances. Thus, principal components based on the covariance matrix will not involve the other variables. In this case, the variables can be standardized before extracting the eigenvalues and eigenvectors. This is equivalent to finding principal components from the correlation matrix. Zitko (2006) asserted that standardization of variables may not eliminate the influence of high variability in few variables on the PCA outcome. Scaling the variables to a fraction of their range $\left((x_{\max} - x_i)/(x_{\max} - x_{\min})\right)$ is recommended by Zitko (2006), and is believed to better remove the effect of such variables on the outcome of PCA.

When the variances differ greatly, or if the measurement units are not commensurate as is common in water quality data, the components of the covariance matrix will be dominated by the variables with large variances. The other variables will contribute very little. For a more balanced representation in such cases, the components of the correlation matrix may be used. Extracting principal components using the covariance matrix with standardized variables is equivalent to extracting principal components using the correlation matrix. Standardizing variables or rescaling the variables to a fraction of their range aims to eliminate the overwhelming effect of largely different variances. Standardization, i.e. scaling the variables to a fraction of their range or centering the variables by removing the means preserves the distribution of the variable. However, scaling the variables may lead to less variability than standardization or centering the variables.

5.5 Optimization approaches

Some researchers stress the use of optimization techniques for the selection of both sampling locations and sampling frequencies (e.g. Skalski and MacKenzie, 1982; Bernstein and Zalinski, 1983; Reinelt et al., 1988; Palmer and MacKenzie, 1985; Spooner et al., 1985; MacKenzie et al., 1987). In these studies, several experimental designs were suggested, mainly related to impact assessment and ecological monitoring. In these designs, two requirements are expected to be fulfilled by the network: cost-effectiveness and statistical power. The latter is often investigated by analysis of variance (ANOVA) techniques, and optimization methods are used to maximize the statistical power of the network while minimizing the costs (Harmancioglu et al., 1992).

Some design approaches combine both the spatial and temporal design criteria to evaluate the space-time trade-off. The approach in these combined designs is to compensate for a lack of information in one dimension by increasing the intensity of efforts in the other dimension (Harmancioglu and Alpaslan, 1992). Tirsch and Male (1984) combined spatial and temporal design using multivariate linear regression. Regression analysis, as applied for the assessment of sampling frequency (4.6) and monitoring locations (5.2), can be combined to evaluate the space-time trade-off. This may take the form of a plot with a temporal axis and a spatial axis to define a precision space. Different location combinations appear on the y-axis and different sampling frequencies appear on the x-axis. In the x-y space, curves of the adjusted coefficient of determination define the trade-off between the spatial and temporal measures. Then, the best combination is decided on an economic basis, by maximizing the net benefits resulting from the monitoring network. The assessment of monitoring benefits is derived from monitoring reliability, as measured by the coefficient of determination.

Ozkul (1996) investigated space-time dimensions of the WQM network in the Mississippi River basin using the entropy concept. They derived curves of redundant information with respect to both the number of stations and sampling frequencies, where redundant information (transinformation) increases with an increase in the number of sampling locations and decreases with a decrease in temporal sampling frequencies (Harmancioglu et al., 1999).

The basic entropy measure used in this combined approach is transinformation. The objective is to select the space-time combination that produces the least amount of transinformation. For a constant level of transinformation, a number of space-time alternatives exist. From such alternatives, one may evaluate whether to increase the number of monitoring locations and decrease the sampling frequency or to decrease the number of monitoring locations and increase the sampling frequency. The final decision depends on the evaluation of cost reduction with respect to decreases in number of monitoring locations or sampling frequencies (Harmancioglu et al., 1999).

Schilperoort et al. (1982) emphasized the need to optimize monitoring networks to achieve cost-effective designs while fulfilling the monitoring objectives. This approach enables the space-time trade-offs in the design to be evaluated. In general, economic analysis can be considered the main tool used to evaluate the space-time trade-off, where the options are either to increase the number of monitoring locations and decrease the sampling frequency or to increase the sampling frequency and decrease the number of monitoring locations.

5.6 Discussion

The statistical approaches that have been proposed for the assessment of monitoring locations are mainly contraction approaches that aim to decrease the number of locations in the monitoring network. The stream ordering approach is the most frequently used approach to design a monitoring network when no water quality data are available. In this method, one of the stream attributes, such as the stream order, stream length, area served by each stream or stream flow, is used to weight each stream. This approach can be used to select monitoring locations during the establishment of a new monitoring program. Although one of the suggested weights is pollutant loading, this approach is valid only if the pollutant loads are measured at all streams. Furthermore, different stream attributes can be used at the same time by giving a combined weight to different attributes for each stream. In this case, the placement of the monitoring locations is done in accordance with several stream attributes simultaneously. Using several attributes sequentially may result in a different spatial design.

The main disadvantage of the entropy and regression methods is that both approaches are generally applied using only one water quality variable. However, assessment and redesign of the water quality monitoring locations is more reliable when based on several water quality indicators. Multivariate data should be handled simultaneously rather than sequentially.

Several inconsistent spatial designs are obtained in case of applying these methods to several water quality variables sequentially. For example, Ozkul et al. (2000) applied the entropy based approach to the case of basin segment 07 of the Mississippi River basin in Louisiana in the USA. They used a water quality monitoring network consisting of 12 locations, where 26 water quality variables are measured on a monthly basis. The entropy approach was applied for each water quality variable separately to select the best combination of monitoring locations. Using the electric conductivity (EC), results indicated that three locations can be excluded from the monitoring network; these are locations number 049, 051 and 055. Using the DO, results indicated that only two locations may be excluded (locations 054 and 055), while using the chemical oxygen demand (COD) indicates that all the aforementioned locations (049, 051, 054 and 055) should be included.

The multivariate data analysis approaches overcome this disadvantage by employing several water quality variables simultaneously. However, there are two main disadvantages in using multivariate data analysis. The first disadvantage is that the reconstruction of information at discontinued locations is not considered. The second disadvantage is that the multivariate data analyses are based on a linear structure in the data set. Given that the relationship between physical, chemical and biological variables may be nonlinear, application of conventional multivariate analyses may be not appropriate.

The common disadvantage of the proposed approaches is that they only focus on identifying monitoring locations to be discontinued. However, the optimum spatial design may consist in the discontinuation of a number of existing monitoring locations while adding other locations at ungauged sites. This downfall arises from excessive focus on assessment using the water quality data already obtained while ignoring the characteristics of the basins being monitored.

Assessment and redesign of monitoring locations should involve several water quality indicators simultaneously as well as the basin characteristics. Different basin characteristics may affect the spatial allocation, such as climatic region, land use, geology, existence of point and nonpoint sources of pollution and human activities within the region. In addition, the assessment and redesign of monitoring locations should also assess whether monitoring locations can be discontinued and if new monitoring locations should be added. The redesign approach to be used should always allow for reconstituting information at discontinued locations.

The classical tools for multivariate statistical analysis include multiple linear regression, PCA and Canonical Correlation Analysis (CCA). CCA is a way of explaining the linear relationship between two sets of variables (Shu and Ouarda, 2007). Torranin (1972) and Rice (1972) presented the first applications of CCA in hydrology. Torranin (1972) demonstrated the potential of CCA applications for hydrological problems through a study of coastal monthly precipitation forecasts and seasonal snowmelt runoff. Rice (1972) applied CCA to hydrological prediction. More recently, CCA has been used with regression or ANN to estimate flood quantiles at ungauged sites (see Ouarda et al., 2000, 2001; Shu and Ouarda, 2007). In this approach, CCA is used to measure similarity between ungauged sites and gauged drainage basins using hydrological, meteorological and basin characteristics. Then, regression or ANN is used to estimate flood quantiles at ungauged sites. In the assessment of spatial distribution of water quality monitoring locations, CCA can play an important role by identifying the correlation between basin characteristics and measured water quality variables. Consequently, this strategy can be employed as the first step to estimate water quality at ungauged locations or discontinued locations. Thus, using CCA with regression or ANN may overcome the disadvantages of using multivariate data analysis.

In general, popular multivariate data analysis methods suffer from the limitation of being linear. Since the late 1980s, ANN methods have been increasingly used to perform nonlinear regression. More recently, ANN methods have been extended to perform nonlinear PCA (NLPCA) and nonlinear CCA (NLCCA) (Hsieh, 2004). In PCA, a given dataset is approximated by a straight line, which minimizes the mean square error (MSE), whereas in NLPCA, the straight line is replaced by a curve that minimizes the MSE. Similarly, NLCCA removes the restriction of detecting only linear oscillations in two spaces. Given that water quality data time series often have a nonlinear structure (Lettenmaier, 1988; Berryman et al., 1988), it may be more appropriate to use the recently-developed nonlinear methods instead of classical multivariate data analysis.

6 Conclusions

Research into important aspects of WQM network design has been ongoing since the 1970s. This paper focused mainly on statistical approaches as a quantitative tool for the assessment and redesign of WQM networks. The design of water quality monitoring programs is not a straightforward process, and is based mainly on the monitoring objectives. It is therefore very important that the objectives of the monitoring program be carefully defined.

Identification of the monitoring objectives should be specific and clearly stated, preferably on a site-by-site basis. The objectives statement should include why we want to monitor at this location, what we would like to monitor, what is the type of information required, and what is the analysis tool we intend to use in order to obtain the desired information. Assessing climate change impacts on water quality should be one of the objectives of surface water-quality-monitoring programs. Input from different stakeholders such as policy makers, water managers, researchers and the public should be considered in the various steps of assessment and redesign, especially when identifying the monitoring objectives.

This review reveals that, although much research has been undertaken to evaluate the performance of monitoring networks, several deficiencies in the proposed approaches still exist. The proposed approaches for the assessment and rationalization of the water quality variables are based mainly on the assumption of linear structure among the variables. Applying nonlinear techniques such as mutual information and ANN will overcome this deficiency. Another deficiency is the absence of a criterion to identify a combination of variables to be continuously measured and those to be discontinued. A performance index based on information criterion may help overcome this downfall. Such an index would also help to integrate the assessment of water quality variables with the assessment of sampling frequency and sampling locations.

The approaches proposed for the assessment of the sampling frequency address a specific water quality variable at a specific monitoring location, and often result in the optimization of sampling around only one of the monitoring objectives. Proportional sampling is recommended to distribute a pre-identified total number of samples among monitoring locations for multiple variables. Such an allocation of the number of samples may result in a fewer number of samples at some locations than required. Thus, an assessment of the consequences of applying proportional sampling should be conducted, especially in monitoring networks with multiple objectives.

Most of the approaches proposed for the assessment of spatial distribution of monitoring locations are based on only one water quality variable. In addition, these methods are based mainly on the assumption of a linear structure in the data, reconstitution of information about discontinued locations is not considered. These approaches focus mainly on identifying monitoring locations to be discontinued, when it may be that the optimal spatial distribution of monitoring locations involves discontinuing of a number of existing locations and while also adding other locations at ungauged sites. The incorporation of basin characteristics in the assessment is believed to overcome this deficiency. New technologies and analysis techniques need to be explored with respect to monitoring network design. Artificial intelligence technologies as well as the new generation of nonlinear multivariate data analysis (NLPCA and NLCCA) induce innovation and new problem-solving criteria.

This review has also revealed that, in the past, the majority of water-quality-monitoring programs have been designed on a rather arbitrary basis. Although many studies have sought to improve the performance of monitoring networks, most have focused mainly on only one of the network design aspects. Few researchers have examined the optimization of different aspects simultaneously. Network assessment and redesign require combining the assessment of the monitoring objectives, the variables to measure, the sampling frequency and the sampling locations into one framework. The three main aspects in monitoring design should be assessed simultaneously, and may be linked by a criterion of either cost or information.

Finally, the design of a monitoring network needs to be periodically re-assessed and modified accordingly to changing environmental conditions and/or shifts in management priorities. In order to facilitate the periodic assessment of the monitoring network performance, the design or the assessment and redesign should be well documented. The documentation should indicate the approaches followed, the standardized laboratory procedures and techniques employed and the data analysis procedures to follow in order to obtain the desired information.

7 References

[1]. Abbott, M.B. and D.R. Basco (1989). Computational fluid dynamics: an introduction for engineers. Longman Scientific & Technical, Harlow.

[2]. Baker, F.B. and L.J. Hubert (1975). Measuring power of hierarchical cluster analysis. Journal of the American Statistical Association, 70(349), 31 - 38.

[3]. Bartram, J. and R. Balance (1996). Water Quality Monitoring: A practical guide to the design and implementation of freshwater quality studies and monitoring programmes. Published on behalf of UNEP and WHO, Taylor & Francis, 383 p.

[4]. Bayley, G. V. and J.M. Hammersley (1946). The “Effective” Number of Independent Observations in an Autocorrelated Time Series. Journal of the Royal Statistical Society, 8(2), 184 - 197.

[5]. Bernstein, B.B. and J. Zalinski (1983). An optimum sampling design and power tests for environmental biologists. Journal of Environmental Management, 16, 35 - 43.

[6]. Berryman, D., B. Bobée, D. Cluis and J. Haemmerli (1988). Nonparametric Tests for Trend Detection in Water Quality Time Series. Water Resources Bulletin, 24(3), 545 - 556.

[7]. Briggs, J.C. and Ficke, J.F. (1978). Quality of rivers of the United States, (1975) water year- based on the National Stream Quality Accounting Network, U.S. Geological Survey Open-File Report 78-200, 436 p.

[8]. Chapman, D. (1996). Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring. Chapman & Hall, London.

[9]. Chen, M. and Cheng, S. (1989). Method for the optimum sampling frequency selection in a regular water quality monitoring system. Huanjing Kexue, 10(3), 58 - 63.

[10]. Czernuszenko, W. and Rowinsk, P. (2005). Hazards and dispersion of pollutants, Springer, 250 pp.

[11]. Dowdall, M., Gerland, S., Karcher, M., Gwynn, J.P., Rudjord, A.L. and Kolstad, A.K. (2005). Optimisation of Sampling for the temporal monitoring of technetium-99 in the Arctic marine environment. Journal of Environmental Radioactivity, 84, 111 - 130.

[12]. Ellis, J.C. and Lacey, R.F. (1980). Sampling: defining the task and planning the scheme. Water Pollution Control, 79, 452 - 467.

[13]. Fischer, H.B., List, E.J., Koh, R.Y.C., Imberger, J. and Brooks, N.H. (1979). Mixing in inland and coastal waters. Academic Press, New York.

[14]. GAO, General Accounting Office (2000). Water quality: Key EPA and State decisions limited by inconsistent and incomplete data. GAO Report Number GAO/RCED-00-54, GAO, Washington, DC., p.78.

[15]. GAO, General Accounting Office (2004). Watershed management: better coordination of data collection efforts needed to support key decisions. Report Number GAO-04-382, GAO Washington, DC., p. 155.

[16]. Gilbert, R.O. (1987). Statistical methods for environmental pollution monitoring. New York, John Wiley & Sons, 320 p.

[17]. Green, R.H. (1979). Sampling design and statistical methods for environmental biologists. New York, John Wiley & Sons, 257 p.

[18]. Hapuarachchi, K.P. and Macpherson, B.D. (1992). Autoregressive processes applied to acceptance sampling by variables. Communications in Statistics: Simulation, 21(3): 833 - 848.

[19]. Harmancioglu, N.B. (1981). Measuring the information content of hydrological processes by the entropy concept. Journal of civil engineering, Ege University, Faculty of Engineering, 13 - 38.

[20]. Harmancioglu, N.B. (1984). Entropy concept as used in determination of optimum sampling intervals. Proceedings of Hydrosoft (1984), International conference on hydraulic engineering software, Portoroz, Yugoslavia, 99 - 110.

[21]. Harmancioglu, N.B. and Alpaslan, M.N. (1992). Water quality monitoring network design: a problem of multi-objective decision making. Water Resources Bulletin, 28 (1), 179 - 192.

[22]. Harmancioglu, N.B., Alpaslan, N., Alkan, A., Ozkul, S., Mazlum, S. and Fistikoglu, O. (1994). Design and Evaluation of Water Quality Monitoring Networks for Environmental Management (in Turkish). Report prepared for the research project granted by TUBITAK, Scientific and Technical Council of Turkey, Project code: DEBAG-23, 514 p.

[23]. Harmancioglu, N.B., Alpaslan, N. and Singh, V.P. (1992). Design of water quality monitoring networks. Geomechanics and Water Engineering in Environmental Management, A.A. Balkema Publishers, Rotterdam, ch. 8, 267 - 296.

[24]. Harmancioglu, N.B., Fistikoglu, O., Ozkul, S.D., Singh, V.P. and Alpaslan, M.N. (1999). Water Quality Monitoring Network Design. Kluwer Academic Publishers, Dordrecht, the Netherlands, 290 p.

[25]. Harmancioglu, N.B. and Tokgoz, S. (1995). Selection of sampling frequencies in water quality monitoring network design. Journal of Water Pollution Control, 5(1), 9 - 20.

[26]. Harmancioglu, N.B. and Yevjevich, V. (1986). Transfer of Information among Water Quality Variables of the Potomac River, Phase III: Transferable and Transferred Information. Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 1986, 81 p.

[27]. Harmancioglu, N.B. and Yevjevich, V. (1987). Transfer of hydrologic information among river points. Journal of Hydrology, 91, 103 - 118.

[28]. Harmancioglu, N.B., Yevjevich, V. and Obeysekera, J.T.B. (1985). Measures of information transfer between variables, Proceedings of the Fourth International Hydrology Symposium, 481 - 499.

[29]. Harriman, R., Gillespie, E., King, D., Watt, A.W., Christie, A.E.G., Cowan, A.A., and Edwards, T. (1990) Short-term ionic response as indicators of hydrochemical processes in the Allt a' Mharcaidh catchment, Western Cairngorms, Scotland, J.Hydrology, 116, 267 - 285.

[30]. Hawkinson, R.O., Ficke, J. F. and Saindon, L. G. (1977). Quality of rivers of the United States, 1974 water year-based on the National Stream Quality Accounting Network (NASQAN), U.S. Geological Survey Open- File Report 77-151, 158 p.

[31]. Hintze, J. (2001). NCSS Help System. Kaysville, Utah, NCSS (Number Cruncher Statistical Systems).

[32]. Horton, R.E. (1945). Erosional Development of Streams. Geological Society Am. Bull., 56, 281 - 283.

[33]. Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6), 417 - 441.

[34]. Hsieh, W.W. (2001). Nonlinear principal component analysis by neural networks, Tellus, Ser. A, 53, 599–615.

[35]. Hsieh, W.W. (2004). Nonlinear multivariate and time series analysis by neural network methods, Rev. Geophys., 42, RG1003, doi:10.1029/2002RG000112.

[36]. Jobson, J.D. (1992). Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods. New York, Springer-Verlag, 768 p.

[37]. Karpuzcu, M., Senes, S. and Akkoyunlu, A. (1987). Design of monitoring systems for water quality by principal component analysis a case study. Proceeding, INT. Symp. On Environmental Management (Environment 87), 673 - 690.

[38]. Khalil, B.M., Abdel-Gawad, S.T., Abdel-Rashid, A. and Morsy, A.M. (2004). Sampling frequency assessment for the drainage water quality monitoring in Egypt, Proceedings, The International IWA Conference, AutMoNet, 19-20 April, Vienna, Austria, 85 - 92.

[39]. Kirchner, J.W., Feng, X., Neal, C. and Robson, A.J. (2004). The fine structure of water quality dynamics, the (high-frequency) wave of the future. Hydrological Processes, 18, 1353 - 1359.

[40]. Klotz, J.H. and Meyer, R.D. (1985). Biostatistical microcomputing in Pascal. Rowman and Allenheld. Totowa, NJ, USA, 168 p.

[41]. Kramer, M.A. (1991). Nonlinear principal component analysis using autoassociative neural networks, AIChE J., 37, 233–243.

[42]. Krige, D.G. (1951). A statistical approach to some basic mine valuation problems in the Witwatersrand. Journal of chemical, Mettalurgical and Mining Society of South Africa 52, 119.

[43]. LaChance, M., Bobee, B. and Haemmerli, J. (1989). Methodology for the planning and operation of a water quality network with temporal and spatial objectives: application to acid lakes in Québec, in: R.C. Ward, J.C. Loftis and G.B. McBride (eds.). Proceedings, International Symposium on the Design of Water Quality Information Systems, Fort Collins, CSU Information Series no. 61, 145 - 162.

[44]. Lettenmaier, D.P. (1976). Detection of trends in water quality data from records with dependent observations. Water Resources Research, 12, 1037 - 1046.

[45]. Lettenmaier, D.P. (1988). Multivariate nonparametric tests for trend in water quality, AWRA, Water Resources Bulletin (24)3, 505 - 512.

[46]. Loftis, J.C., McBride,G.B. and Ellis, J.C. (1991) Considerations of scale in water quality monitoring and data analysis. Water Resources Bulletin, 27(2): 255 - 264.

[47]. Loftis, J.C. and Ward, R.C. (1979). Regulatory water quality monitoring networks-statistical and economic considerations. U.S. Environmental Protection Agency Report No. EPA-600/4-79-055.

[48]. Mace, A.E. (1964). Sample-size determination. Reinhold, New York, USA, 226 p.

[49]. Mackenzie, M.C., Palmer, R.N. and Millard, S.P. (1987). Analysis of Statistical Monitoring Network Design. Journal of Water Resources Planning and Management, ASCE, 113(5), 599 - 615.

[50]. Matheron, G. (1963). Principles of geostatistics. Economic Geology 58, 1246 - 1266.

[51]. McBride, G.B. and Ellis, J.C. (2001). Confidence of compliance: A Bayesian approach for percentile standards. Water Research, 35(5): 1117 - 1124.

[52]. McKenzie, S.W. (1976). Long-term water quality trends in Delaware streams: U.S. Geological Survey, Report 76-71, p.85.

[53]. Miller, D.G. and Ellis, J.C. (1986). Derivation and monitoring of consents. Water pollution control, 85: 249 - 258.

[54]. Monahan, A.H. (2000). Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System. Journal of Climate, (13)4, 821 - 835.

[55]. Monahan, A.H. (2001). Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure, J. Clim., 14, 219–233.

[56]. Monahan, A.H., Fyfe, J.C. and Flato, G.M. (2000). A regime view of Northern Hemisphere atmospheric variability and change under global warming, Geophys. Res. Lett., 27, 1139–1142.

[57]. Monahan, A.H., Pandolfo, L. and Fyfe, J.C. (2001). The preferred structure of variability of the Northern Hemisphere atmospheric circulation, Geophys. Res. Lett., 28, 1019–1022.

[58]. NRCS, (2003). National handbook of water quality monitoring. U.S. Department of Agriculture Natural Resources Conservation Service, 450-VI-NHWQM.

[59]. Odom, K.R. (2003). Assessment and Redesign of the Synoptic water quality monitoring network in the Great Smoky Mountains National Park. Ph.D. Dissertation, University of Tennessee, Knoxville, USA, 268 p.

[60]. Ouarda, T.B.M.J., Girard, C., Cavadias, G.S. and Bobée, B. (2001). Regional flood frequency estimation with canonical correlation analysis, Journal of Hydrology, 254, 157 - 173.

[61]. Ouarda, T.B.M.J., Haché, M., Bruneau, P. and Bobée, B. (2000). Regional flood peak and volume estimation in a northern Candian basin. Journal of cold region Engineering, 14, 176 – 191.

[62]. Ouarda, T.B.M.J., Rasmussen, P.F., and Bobée, B., Sorin, J. (1996) Ontario Hydrometric Network Rationalization, Statistical Consideration, Research Report No. R-470, National Institute for Scientific Research, INRS-ETE, University of Québec, Québec, Canada, 75 p.

[63]. Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. Water Research, 39, 2621 - 2635.

[64]. Ozkul, S.D. (1996). Space / Time Design of Water Quality Monitoring Networks by the Entropy Method, Ph.D. Thesis on Civil Engineering, Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Izmir, 196 p.

[65]. Ozkul, S.D., Alkan, A., Harmancioglu, N.B. and Alpaslan, N. (1995). Evaluation of sampling frequencies in the design of water quality monitoring networks. Proceedings, Advances in Civil Engineering, Second Technical Congress, Bagazici University, Istanbul, 302 - 312.

[66]. Ozkul, S.D., Harmancioglu, N.B. and Singh, V.P. (2000). Entropy-based assessment of water quality monitoring networks in space/time dimensions. Journal of Hydrologic Engineering, ASCE, 5 (1), 90 - 100.

[67]. Palmer, R.N. and Mackenzie, M.C. (1985). Optimization of Water Quality Monitoring Networks. Journal of Water Resources Planning and Management, ASCE, 3(4), 478 - 493.

[68]. Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics, 33, 1065 - 1076.

[69]. Peck, R., Fisher, L. and Van Ness, J. (1989). Approximate Confidence Intervals for the Number of Clusters. Journal of the American Statistical Association, 84(405), 184 - 191.

[70]. Rao, C.R. (1973). Linear Statistical Inference and Its Applications, Second Edition. New York, John Wiley & Sons, 625 p.

[71]. Reinelt, L.E., Horner, R.R. and Mar, B.W. (1988). Nonpoint Source Pollution Monitoring Program Design. Journal of Water Resources Planning and Management, ASCE, 114(3), 335 - 352.

[72]. Rice, R.M. (1972). Using canonical correlation for hydrological predictions. Bull.Int.Assoc.Hydrol.Sci. XVII 3(10), 315 - 321.

[73]. Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. Annals of Mathematical Statistics, 27, 832 - 837.

[74]. Sanders, T.G. and Adrian, D.D. (1978). Sampling Frequency for River Quality Monitoring. Water Resources Research, 14, 569 - 576.

[75]. Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V. (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[76]. Schilperoort, T., Groot, S., Wetering, B.G.M. and Dijkman, F. (1982). Optimisation of the sampling frequency of water quality monitoring networks, “Waterloopkundig” Laboratium Delft, Hydraulics Lab, Delft, the Netherlands.

[77]. Shannon, C.E. (1948). A mathematical theory of communication, Bell system Technical Journal, 27, 397 - 423.

[78]. Sharp, W.E. (1970). Stream Order as a Measure of Sample Source Uncertainty. Water Resources Research, 6(3), 919 - 926.

[79]. Sharp, W.E. (1971). A topologically optimum water sampling plan for rivers and streams. Water Resour. Res. (7)6, 1641 - 1646.

[80]. Shu, C. and Ouarda, T.B.M.J. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, vol. 43, W07438, doi:10.1029/2006WR005142.

[81]. Singh, V.P. (1997). The use of Entropy in Hydrology and Water Resources. Hydrological Processes, 11, 587 - 626.

[82]. Skalski, J.R. and MacKenzie, D.H. (1982). A design for aquatic monitoring programs. Journal of Environmental management, 14, 237 - 251.

[83]. Sloan, S. and Pender, G. (1998). Evaluation of Numerical Modelling Techniques for Solving the Advection-Diffusion Equation. International Conference on Hydro-Science and -Engineering, pp. 1-13.

[84]. Smith, E.P., Ye, K., Hughes, C. and Shabman, L. (2001). Statistical assessment of violations of water quality standards under section 303(d) of the Clean Water Act. Environmental Science & Technology, 35(3): 606 - 612.

[85]. Smith, E.P., Zahran, A., Mahmoud, M. and Ye, K. (2003). Evaluation of water quality using acceptance sampling by variables. Environmetrics, 14: 373 - 386.

[86]. Spooner, J., Mass, R.P., Dressing, S.A., Smolen, M.D. and Humenik, F.J. (1985). Appropriate designs for documenting water quality improvements from agricultural NPS control programs. In prespective on Nonpoint source Pollution, US EPA440 5-85-001, 30 - 34.

[87]. St-Hilaire A., Brun, G., Courtenay, S. C., Ouarda, T.B.M.J., Boghen, A.D. and Bobée, B. (2004). Multivariate Analysis of Water Quality in the Richibucto Drainage Basin (New Brunswick, Canada). Journal of the American Water Resources Association (JAWRA), June, 691 - 703.

[88]. Strobl R.O., Robiliard P.D., Shannon R.D., Day R.L. and McDonnell A.J. (2006). A Water quality monitoring network design methodology for the selection of critical sampling points: Part I, Environmental Monitoring and Assessment, 112, 137 - 158.

[89]. Strobl, R.O. and Robillard, P.D. (2008). Network design for water quality monitoring of surface freshwaters: A review, Journal of Environmental Management, 87, 639 - 648.

[90]. Tabachnick, B.G. and Fidell, L.S. (1996). Using Multivariate Statistics. Allyn and Bacon, Boston, London, 879 p.

[91]. Tirsch, F.S. and Male, J.W. (1984). River basin water quality monitoring network design: options for reaching water quality goals, in: T.M. Schad (ed.). Proceeding of Twentieth Annual Conference of American Water Resources Associations, AWRA Publications, 149 - 156.

[92]. Torranin, P. (1972). Applicability of canonical correlation in hydrology. Hydrology paper 58. Colorado State University, Fort-Collins, CO 30 pp.

[93]. Tokgoz, S. (1992) Temporal design of water quality monitoring networks, Master of Science thesis in Civil Engineering, Dokuz Eylul University, Graduate School of Natural and Applied Science, Izmir.

[94]. Wade, N.L. and Whitfield, P.H. (1994). Observing transient water quality events using electronic sensors, in National Symposium on Water Quality, American Water Resources Association, 105 - 112.

[95]. Ward, R.C. (1978). Evaluating the sampling frequencies of water quality monitoring networks. EP 1.23/8:600/7-78-169, US EPA, Las Vegas, NV.

[96]. Ward, R.C. (1989). Water quality monitoring – a systems approach to design, in: R.C. Ward, J.C. Loftis, and G.B. McBride (eds.), Proceedings, International Symposium on the Design of Water Quality Information Systems, Fort Collins, CSU Information Series no. 61, 37 - 46.

[97]. Ward, R.C., Loftis, J.O. and McBride, G.B. (1990). Design of Water Quality Monitoring systems, Van Nostrand Reinhold, New York, USA, p 231.

[98]. Ward, R.C., Loftis, J.C., Nielsen, K.S. and Anderson, R.D. (1979). Statistical evaluation of sampling frequencies in monitoring networks. J. of WPCF, 51(9), 2292 - 2300.

[99]. Ward, R.C. and Nielsen, K.S. (1978). Evaluating the sampling frequencies of water quality monitoring networks: U.S. Environmental Protection Agency. Report No. EPA-600/7-78-169.

[100]. Weatherley, N.S. and Ormerod, S.J. (1991). The importance of acid episodes in determining faunal distributions in Welsh streams, Freshwater Biology, 25, 71 - 84.

[101]. Whitfield, P.H. (1983). Evaluation of water quality sampling locations on the Yukon River. Water Resources Bulletin, AWRA, 19(1), 115 - 121.

[102]. Whitfield, P.H. (1988). Goals and data collection design for water quality monitoring. Water Resources Bulletin, AWRA, 24(4), 775 - 780.

[103]. Whitfield, P.H. and Dalley, N.E. (1987). Rainfall driven Ph depressions in a British Columbia Coastal Stream, in Symposium on Monitoring Modeling and Mediating Water Quality, American Water Resources Association, 285 - 294.

[104]. Whitfield P.H. and Wade, N.L. (1992). Monitoring transient water quality events electronically, Water Resources Bulletin, 28, 703 - 711.

[105]. Whyte D.C., Kirchner J.W. (2000). Assessing water quality impacts and cleanup effectiveness in streams dominated by episodic mercury discharges. Science of the Total Environment 260, 1 - 9.

[106]. Wong, M. A. (1982). A Hybrid Clustering Method for Identifying High-Density Clusters. Journal of the American Statistical Association, 77(380), 841 - 847.

[107]. Wu, A., Hsieh, W.W. and Shabbar, A. (2002). Nonlinear characteristics of the surface air temperature over Canada, J. Geophys. Res., 107(D21), 4571, doi:10.1029/2001JD001090.

[108]. Yang, Y. J. and Burn, D. H. (1994). An Entropy Approach to Data Collection Network Design. Journal of Hydrology, 157(4), 307 - 324.

[109]. Yevjevich, V. and Harmancioglu, N.B. (1985). Modeling Water Quality Variables of Potomac River at the Entrance to its Estuary, Phase II (Correlation of Water Quality Variables within the Framework of Structural Analysis). Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 59p.

[110]. Yue, S. and Wang, C. (2004). The Mann-Kendal test modified by effective sample size to detect trend in serially correlated hydrological series. Water Resources Management, 18, 201 - 218.

[111]. Zar, J.H. (1996). Biostatistical analysis (3rd ed.), Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 929 p.

[112]. Zhou Y. (1996). Sampling frequency for monitoring the actual state of groundwater systems. Journal of Hydrology, 180, 301 - 318.

[113]. Zitko, V. (2006). Comments on Ouyang, Y. Evaluation of river water quality monitoring stations by principal component analysis. Water Research 39(2005), 2621 – 2635. Water Res. 40, 3141 - 3143.

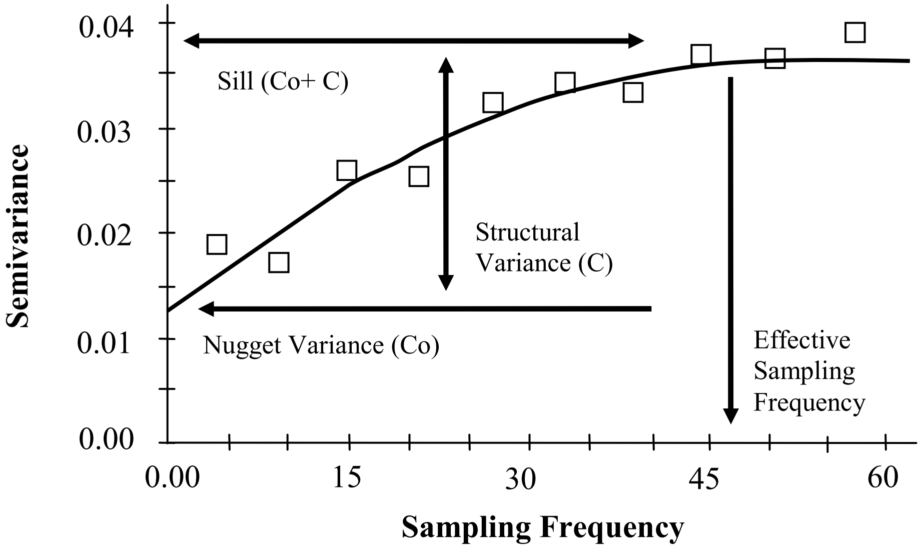


Figure 1. Semivariogram parameters

Article II. A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks

A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks

B. Khalil[1,2], T. B.M.J. Ouarda[2], A. St-Hilaire[2] and F. Chebana[2]

[1] Irrigation and hydraulics department, Faculty of Engineering, Helwan University, Cairo, Egypt.

[2] Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE, Québec City, Canada.

Published in Journal of Hydrology, 2010, vol 386 (1-4), 173-185

Abstract

Despite several decades of operation and the increasing importance of water quality monitoring (WQM) networks, authorities still rely on subjective or semi-subjective decision processes to identify the optimal combination of water quality variables to measure. For this purpose, a statistical approach is developed for the assessment and selection of the optimal combination of water quality variables. The proposed approach overcomes deficiencies in the conventional correlation-regression approach used to assess and eventually reduce the number of water quality variables in WQM networks. For the reduction of water quality variables, criteria developed from record-augmentation procedures are integrated with correlation analysis and cluster analysis to identify highly associated water quality variables. This step is followed by the application of an information performance index to systematically identify the optimal combination of variables to be continuously measured and those to be discontinued. The linear regression and maintenance of variance (MOVE) record-extension techniques are employed to reconstitute information about discontinued variables. The proposed approach is applied for the rationalization of water quality variables in the Nile Delta surface WQM network in Egypt. Results indicate that the proposed approach represents a useful decision support tool for the optimized selection of water quality variables. The MOVE record-extension technique is shown to result in better performance than regression for the estimation of discontinued variables.

Keywords: sampling; water quality indicator; monitoring network; rationalization; record extension; record augmentation.

1 Introduction and review

Assessment of water resources requires knowledge and a full understanding of both the water quantity and the water quality processes (Harmancioglu et al., 1999). Water quality monitoring programs aid in understanding various water quality processes as well as provide water managers with the necessary information for water resources management in general and water quality management in particular. The design of monitoring networks is the translation of the monitoring objectives to specify sampling sites, sampling frequency and the variables to be measured. Both sampling frequency and sampling sites are influenced by water quality variables being monitored, and therefore the selection of the specific variables of interest is intrinsic to the design and subsequent operation of a water quality monitoring network (Strobl and Robillard, 2008).

The quality of a water body is usually described by sets of physical, chemical and biological variables that are mutually interrelated. Water quality can be defined in terms of one variable to hundreds of compounds and for multiple uses. This is a very complex issue since there are numerous variables to choose from in representing surface water quality (Sanders et al., 1983; Harmancioglu et al., 1999). Many researchers have recognized that it is impossible to measure everything in the environment, and that some logical means of selecting the variables to measure has to be part of every water quality information system (Ward et al., 1990). Consideration should be given to reduce the number of variables sampled without a substantial loss of information. Fewer variables would be easier and less costly to analyze. Dependencies or correlations between various water quality variables become easier to establish if they are few, saving time and effort (Strobl and Robillard, 2008).

A review of the literature reveals that the main statistical approach proposed to reduce the number of variables being measured is the Correlation-Regression (CR) approach. The CR approach is based on three steps. The first step is the assessment of the level of association among the variables being measured by correlation analysis. If high correlation exists among variables, this is an indication that some of the information produced may be redundant. The second step is the selection of water quality variables to be continuously measured and those to be discontinued. This step is based on some subjective criteria, such as the significance of the variable, the presence of the variable in local or international standards, etc. It may also be based on the cost of the laboratory analysis. The third step is the reconstitution of information about discontinued variables using auxiliary variables from the continuously measured ones.

During the 1970s, several studies showed that the concentrations of major ionic constituents can be related to specific conductance (McKenzie, 1976; Briggs and Ficke, 1978). Specific conductance can thus serve as an indicator variable from which the concentrations of major ionic solutes can be determined in cases where suitable regression functions can be found between the former and the latter (Sanders et al., 1983). Yevjevich and Harmancioglu (1985) investigated the transfer of information using bivariate correlations between daily water quality variables observed along the Upper Potomac River Estuary, USA. The objective was to determine pairs of variables that exhibited strong relationships in order to define variables to be sampled continuously and variables that can be estimated. Similar analyses were carried out by Harmancioglu et al. (1987) on monthly data from a highly polluted river basin in Turkey. Harmancioglu and Yevjevich (1986, 1987) studied the effects of removing deterministic components (trends, periodicity and stochastic dependence) in order to see the effects of these characteristics on the amount of information transfer. They concluded that some basic similarities in deterministic components are the main contributors to transferred information.

The main advantage of the CR approach is that it allows for the reconstitution of information about discontinued variables using regression analysis. However, three main deficiencies exist in the CR approach as commonly practiced for water quality variables reduction. The first deficiency is the method used to identify highly associated variables. The correlation coefficient is commonly used as a criterion to assess the level of association, but selection of the proper threshold above which a correlation coefficient can be considered sufficient to associate two variables can be problematic. Assessment of the correlation coefficient is always left to a subjective designer preference. Thus, studies of the same set of variables by different investigators may lead to different results. The second deficiency is the absence of a criterion to identify the combination of variables to be continuously measured and those to be discontinued. The third deficiency is that the use of regression analysis to reconstitute information about discontinued variables often results in an underestimation of the variance in the extended records (Alley and Burns, 1983; Hirsch, 1982).

The main goal of this study is to modify the conventional CR approach to overcome these deficiencies, summarized as follows: 1) Subjective assessment of the correlation coefficient; 2) Subjective selection of discontinued variables; and 3) Underestimation of the variance in the extended records of the discontinued variables.

In the following section, a description of the study area is provided. In section 3, the methodology is presented. In section 4, the results obtained are presented and discussed. Finally, the conclusions from this work are presented in section 5.

2 Egyptian national WQM network

In a survey of world freshwater, it has been reported that Egypt is among the ten countries to be plagued with scarcest water resources by the year 2025 due to its rapidly increasing population (Engelman and Le Roy, 1993). The distribution of the Egyptian share of Nile River water to its population is near the water poverty threshold and will fall well below this threshold in the years to come (MWRI, 1997; Wolf, 2000). The per-capita share of fresh water resources in Egypt is now almost 800 m^3 per person per year (Abdel-Gawad et al., 2004; Frenken, 2005). One of the applied solutions to stretch limited Egyptian water resources is the reuse of any kind of drainage water (agricultural/industrial/municipal or most often a mixture thereof) in agricultural production processes. Drainage water with low salinity is used directly or after mixing with fresh Nile water in the case of slightly brackish water. Drainage water with high salinity or that is contaminated by municipal and industrial wastes cannot be used in irrigation.

The drainage system in the Nile Delta is composed of 22 catchment areas. Depending on their quality, effluents are either discharged into the Northern Lakes or pumped into irrigation canals at 21 sites along the main drains to augment freshwater supply (DRI-MADWQ, 1998). Numerous programs have been developed in the past to monitor the water quality of the Nile and of agricultural drainage in Egypt. In 1977, the National Water Research Center (NWRC) had started to monitor a few volumetric and qualitative water parameters (predominantly concerning salinity) in some of the main drains in the Nile Delta. Since 1997, the NWRC continuously had to expand its monitoring activities to include an ever-increasing number of sampling sites and water quality variables. The monitoring program of the Nile Delta drainage system aims at assessing its compliance with the national standards, estimating mass transport and identifying temporal and spatial trends (NAWQAM, 2001).

Twenty-eight water quality variables are measured monthly at 94 sites on the Nile Delta drainage system (Figure 1). Two more variables were added in August 1999: the Fecal Coliform (*FCol*) and the Total Nitrogen (*TN*). Three other variables were added in August 2000: Manganese (*Mn*), Boron (*B*), and Nickel (*Ni*). Turbidity (*Turb*) records are available from August 1997 to July 2005. The available data for the 33 water quality variables from the 94 monitoring sites from August 1997 to July 2007 are used in this study (Table 1).

(Figure 1) and (Table 1)

3 Methodology

The methodology is divided into two main subsections. The first subsection deals with the background on the record augmentation and extension approaches often used to reconstitute information concerning short-record streamflow stations as adapted to water quality monitoring networks. The second subsection presents the proposed approach.

3.1 Background on record augmentation and extension approaches

The estimation of the mean and variance of streamflows or other hydrological variables at a short-record gauge from another longer continuously measured gauge is termed record augmentation (Vogel and Stedinger, 1985). However, the problem of actually extending monthly, weekly or daily records is termed record extension. The following subsections present respectively a short review of record-augmentation and record-extension approaches.

3.1.1 Record augmentation

We assume that the measured variable y has n_1 years of data and the measured variable x has n_1+n_2 years of which n_1 are concomitant with the data observed for y, illustrated as follows:

$$\begin{array}{l} x_1, x_2, x_3, \ldots\ldots, x_{n_1}, x_{n_1+1}, x_{n_1+2}, \ldots\ldots\ldots, x_{n_1+n_2} \\ y_1, y_2, y_3, \ldots\ldots, y_{n_1} \end{array}$$

In the case of water quality variable reduction, one can consider that year n_1 is the year where assessment and selection took place. After n_1 years, the decision is to stop measuring the variable y and continue measuring the variable x. Assume after n_2 years that our interest is to estimate the mean μ_y or the variance σ_y^2 of the variable y as accurately as possible. Matalas and Jacobs (1964) developed a procedure for obtaining unbiased estimators of both μ_y and σ_y^2, showing that the mean value ($\hat{\mu}_y$) of the extended series can be determined by:

$$\hat{\mu}_y = \bar{y}_1 + \frac{n_2}{n_1+n_2}\hat{\beta}(\bar{x}_2 - \bar{x}_1) \qquad (1)$$

where $\bar{y}_1$ and $\bar{x}_1$ are the mean values of y_i and x_i based on short records $i = 1,...,n_1$, $\bar{x}_2$ is the mean value of x_i observed during the period $i = n_1 + 1,....,n_2$ and the parameter $\hat{\beta}$ is the estimated regression coefficient. Based on this formulation, it is possible to show (Cochran, 1953) that the variance of $\hat{\mu}_y$ is given by:

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

where σ_y^2 is the population variance of y and ρ is the population correlation between x and y. For practical use, these values may be replaced by their estimates based on the n_1 years of data (Ouarda, et al., 1996). In order to assess whether the extended series provides additional information on the variable y, the variance above must be compared with the variance obtained from the short record (σ_y^2/n_1), and the condition for an improved estimator (smaller variance) of the mean is given by:

$$\rho^2 > 1/(n_1 - 2) \qquad (3)$$

Thus, estimating the mean from the extended series is profitable only if the correlation coefficient between the two variables exceeds $(|n_1 - 2|)^{-1/2}$ (Ouarda et al., 1996). If the variance of the y-series is of interest, one can proceed as in the case of the mean. Matalas and Jacobs (1964) obtained the following expression for the unbiased variance estimator $\hat{\sigma}_y^2$:

$$\hat{\sigma}_y^2 = \hat{\beta}^2 s_x^2 + \left[1 - \frac{n_1 + n_2 - 3}{(n_1 - 3)(n_1 + n_2 - 1)}\right]\frac{n_1 - 1}{n_1 - 2}(s_{y1}^2 - \hat{\beta}\ s_{x1}^2) \tag{4}$$

where s_x^2 is the variance estimate based on the entire x-series and s_{y1}, s_{x1} are the standard deviations of y and x based on the short records $i = 1,\ldots,n_1$. Moreover, Matalas and Jacobs (1964) showed that the variance of the variance estimator ($Var\{\hat{\sigma}_y^2\}$) is given by:

$$Var\{\hat{\sigma}_y^2\} = \frac{2\sigma_y^4}{n_1 - 1} + \frac{n_2\sigma_y^4}{(n_1 + n_2 - 1)^2 (n_1 - 3)}(A\rho^2 + B\rho + C) \tag{5}$$

where

$$A = \frac{(n_2 + 2)(n_1 - 6)(n_1 - 8)}{(n_1 - 5)} + (n_1 - 4)\left(\frac{n_1 n_2 (n_1 - 4)}{(n_1 - 3)(n_1 - 2)} - \frac{2n_2 (n_1 - 4)}{(n_1 - 3)} - 4\right) \tag{6}$$

$$B = \frac{6(n_2 + 2)(n_1 - 6)}{(n_1 - 5)} + 2(n_1^2 - n_1 - 14) + (n_1 - 4)\left(\frac{2n_2 (n_1 - 5)}{(n_1 - 3)} - 2(n_1 + 3) - \frac{2n_1 n_2 (n_1 - 4)}{(n_1 - 3)(n_1 - 2)}\right) \tag{7}$$

$$C = 2(n_1 + 1) + \frac{3(n_2 + 2)}{(n_1 - 5)} - \frac{(n_1 + 1)(2n_1 + n_2 - 2)(n_1 - 3)}{(n_1 - 1)} + (n_1 - 4)\left(\frac{2n_2}{(n_1 - 3)} + 2(n_1 + 1) + \frac{n_1 n_2 (n_1 - 4)}{(n_1 - 3)(n_1 - 2)}\right) \tag{8}$$

The first term on the right-hand side in equation 5 is equal to the variance of the variance estimator based on the n_1 years of the y-series, and estimation is therefore profitable when:

$$\rho^2 > (-B \pm \sqrt{B^2 - 4AC}) / 2A \tag{9}$$

Referring to our main objectives, one can use either equation 3 or equation 9, or both, as criteria to assess the required correlation coefficient threshold and consequently to identify highly associated pairs. Thus, using such criteria, one can overcome the first disadvantage of the CR approach. In this study, the higher correlation coefficient obtained from equations 3 and 9 was used as a criterion to evaluate the correlation coefficients among water quality variables.

3.1.2 Record extension

The literature review reveals that several approaches are available for record extension. The first approach is linear regression. To estimate records of the discontinued variable y for the period $n_1 + 1$ through n_2 years, one can use simple linear regression of y on x.

$$\hat{y}_i = a + bx_i \tag{10}$$

where $\hat{y}_i$ are the estimated values of y for $i = n_1 + 1, \ldots n_2$ and a and b are the constant and slope of the regression equation, respectively. The parameters a and b are the values that minimize the sum of the squared difference between estimated and measured y values. The solutions of a and b are found by solving the normal equations (Draper and Smith, 1966, p. 59). The optimal solution to equation 10 is:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \tag{11}$$

where r is the product-moment correlation coefficient between the n_1 concurrent measurements of x and y. The use of regression analysis often results in underestimation of the variance in the extended records (Alley and Burns, 1983). Matalas and Jacobs (1964) demonstrated that unbiased estimates of the mean ($\hat{\mu}_y$) and variance ($\hat{\sigma}_y^2$) are achieved if the following equation is used:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) + \alpha\left(1 - r^2\right)^{1/2} s_{y1}\, e_i \qquad (12)$$

where α is a constant that depends on n_1 and n_2 (see Hirsch, 1982), r is the product-moment correlation coefficient between the n_1 concurrent measurements of x and y and e_i is a normal independent random variable with zero mean and unit variance. However, due to the presence of an independent noise component (e_i), the problem in using equation 12 is that studies of the same sequence of x and y by different investigators will almost surely lead to different values of $\hat{y}_i$ (Hirsch, 1982; Alley and Burns, 1983).

Hirsch (1982) suggested two other methods referred to as MOVE1 and MOVE2 (Maintenance of Variance, Types 1 and 2). In MOVE1, Hirsch (1982) chose the estimators of a and b so that if equation 10 is used to generate an entire sequence $\hat{y}_i$, for $i = 1,\ldots, n_1 + n_2$, the short sample moments $\bar{y}_1$ and s_{y1}^2 would be reproduced. Similarly in MOVE2, Hirsch (1982) chose a and b so that if equation 10 is used to generate an entire sequence $\hat{y}_i$ for $i = 1,\ldots, n_1 + n_2$, the unbiased estimates $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ would be estimated.

Hirsch (1982) evaluated the MOVE1, MOVE2, regression and regression-plus-noise methods using a Monte-Carlo study as well as an empirical analysis. Samples for the Monte-Carlo study were generated under the assumption that concurrent observations of x and y are stationary, serially independent and have a bivariate normal probability distribution. The empirical analysis used real data to explore the performance of the four approaches under conditions of non-normal distribution, serial dependence and seasonal cycles. Both the Monte-Carlo study and the empirical analysis showed that regression cannot be expected to provide records with the appropriate variability.

In practice, one uses equation 10 to generate the $\hat{y}_i$ only for $i = n_1 + 1,..., n_1 + n_2$ not for $i = 1,..., n_1 + n_2$. This suggests that Hirsch used estimators of a and b that did not achieve what he intended (Vogel and Stedinger, 1985). MOVE3 was then proposed by Vogel and Stedinger (1985). In MOVE3, the main goal is to select a and b in equation 10 so that the resultant sequence of $n_1 + n_2$ values $\{y_1,..., y_{n1}, \hat{y}_{n1+1},..., \hat{y}_{n1+n2}\}$ has a mean $\hat{\mu}_y$ and variance $\hat{\sigma}_y^2$ (the Matalas and Jacobs estimators of equations 1 and 4). Estimates of a and b for the MOVE3 method are obtained by rewriting equation 10 as:

$$\hat{y}_i = a + b(x_i - \bar{x}_2) \tag{13}$$

where estimates of a and b are obtained from equations 14 and 15:

$$a = \left\lfloor (n_1 + n_2)\hat{\mu}_y - n_1 \bar{y}_1 \right\rfloor / n_2 \tag{14}$$

$$b^2 = \left[(n_1 + n_2 - 1)\hat{\sigma}_y^2 - (n_1 - 1)s_{y1}^2 - n_1 \left(\bar{y}_1 - \hat{\mu}_y\right)^2 - n_2 \left(a - \hat{\mu}_y\right)^2 \right] \left[(n_2 - 1)s_{x2}^2 \right]^{-1} \tag{15}$$

Vogel and Stedinger (1985) carried out a Monte-Carlo experiment, which indicated that MOVE2 and MOVE3 techniques are nearly indistinguishable with respect to the mean square error of the estimators of the mean and variance of the complete extended record.

Record augmentation and extension techniques are widely applied in hydrometric networks to reconstitute information about flow in short-gauged stations. Hirsch (1979, 1982), Vogel and Stedinger (1985) and Moog and Whiting (1999) applied the regression and MOVE techniques to the logarithm of the streamflow records rather than the raw data. This transformation tends to improve the normality of discharge histograms, which generally exhibit a strongly positive skew (Hirsch, 1982). In addition, it is preferable to scale the residuals by computing errors not in flow rate, but in the logarithm of flow (Hirsch, 1979). In this way, errors are appropriately weighted relative to the discharge. Otherwise, the error measure would be dominated by the largest discharges in the largest streams (Moog and Whiting, 1999).

Similarly, water quality variables generally exhibit positive skew (Lettenmaier, 1988; Berryman et al., 1988), which is also confirmed in our case by a preliminary analysis of the Nile Delta data. Consequently, in this study, record-extension techniques are applied on the logarithms of the water quality variables. The consequences of this transformation are that, for any of the techniques, the sample mean of the extended record of the logarithms is an unbiased estimate of the mean of the logarithms, but the sample mean of the extended record of concentrations is not an unbiased estimate of the mean of the concentration (Hirsch, 1982; Sydor, 1998). However, the objective is to produce records with sample cumulative distribution functions (CDFs) that are close approximates of the CDF of the actual records (Hirsch, 1982).

3.2 Proposed approach

The proposed approach consists of four main steps. The first step is to assess the level of association among the variables being measured and to define groups of variables that are highly associated. Then, for each highly associated group of variables, the second step is to assume that each variable within the group would be discontinued and to identify its best auxiliary variable from the same group. The third step is to assess different combinations of variables to be discontinued and variables to be continuously measured. The last step is to build models with which the information about discontinued variables can be reconstituted from continuously measured variables.

3.2.1 Association assessment

Given the large number of variables that can be measured to assess surface water quality, it is desirable to pre-classify them into smaller groups such that variables of the same group are highly correlated. This can be done using cluster analysis (CA). Hierarchical clustering is carried out in two consecutive steps: 1) define dissimilarity between variables, and 2) define the linkage function between clusters. A matrix of association (correlation coefficients) is converted into a dissimilarity matrix by substituting each (r) with $(1-r^2)$. In the second step, the average linkage function is used to define the various clusters.

To define at which level of dissimilarity the clusters are best identified, the conditions for an improved estimator (smaller variance) of the mean or the variance as expressed by equations 3 and 9 are employed. Equations 3 and 9 are used to identify dissimilarity measures (d_m and d_v) for the cluster analysis as follows:

$$d_m < 1 - \frac{1}{n_1 - 2} \tag{16}$$

$$d_v < 1 - \left(-B \pm \sqrt{B^2 - 4AC} \right) / 2A \tag{17}$$

Thus, one can use d_m and d_v as criteria to identify the distance linkage at which the cluster can be created. Using such criteria, one can assess the correlation coefficients between variables. This allows the designer to overcome the first disadvantage in the conventional CR approach. Assessment of correlation among water quality variables is applied at each monitoring site separately. The separate treatment of each site is based on the preliminary analysis, which indicates that the levels of correlation between different variables vary from site to site. It should be noted that, as a particular case, some of the clusters may contain only one variable each (single-variable cluster). The final rationalized list of variables should ideally contain variables from all identified clusters of variables. If all variables of a particular cluster are discontinued, it will no longer be possible to extend data within that cluster. Thus, variables that form single-variable clusters should be continuously measured.

3.2.2 Best auxiliary variables

After identifying clusters of highly associated variables, the following step is to study each multiple-variable cluster separately. The approach assumes that each variable within the cluster is the variable to be discontinued. For each discontinued variable, the best auxiliary variable for record extension is selected from the other variables in the same cluster. This selection depends on the correlation between discontinued and auxiliary variables, and the length of the common period of record. In this study, Equation 2 is used to identify the best auxiliary variable that minimizes the variance of the estimated mean of each discontinued variable.

Using Equation 2, the choice of the best auxiliary variable is based on the number of concurrent years of measurements, the correlation coefficient and also on the number of years after the assessment and reselection took place (n_2). One can assess the precision of the variance of the mean value estimator (Equation 2) after a certain number of years, assuming σ_y^2 and ρ remain unchanged and equal to their estimates based on n_1 years of data. In this study, n_2 is assumed to be two years, thus one would like to reconstitute information about discontinued variables after two years from assessment and reselection took place.

3.2.3 Selection of discontinued variables

This step aims to define the variables to be continuously measured and those to be discontinued. For instance, we consider the case where budget cuts require k variables to be discontinued. Which k variables among the w variables in the list of variables being measured should be selected? The number of possible combinations of variables to discontinue is given by the binomial coefficient, $C(w,k)$. For each combination, one may compute an information index according to which the combinations may be ranked. Such a procedure allows the identification of the best combination of variables to discontinue, or provides the decision maker with the rank of the best combinations to discontinue. For practical comparison of the combinations, the information index must be based on some kind of aggregated information (Ouarda et al., 1996). An aggregated performance index, I_a, can be defined as follows:

$$I_a = \sum_{\text{var}\,iables\ X} \sqrt{Var\{\hat{\mu}\{X\}\}} \qquad (18)$$

where X is the water quality variable and $Var\{\hat{\mu}\{X\}\}$ is the variance of the mean value estimator expected after n_2 years (Equation 2). This summation is carried out over all variables under study. For the discontinued variables, the variance of the mean value estimator after n_2 years is estimated using Equation 2. The population parameters in Equation 2 are replaced by their estimates based on the n_1 years of data. For continuously measured variables, the variance of the mean value after n_2 years is assumed to be equal to the variance of the mean after n_1 years multiplied by $(n_1\text{-}1)/(n_1\text{+}n_2\text{-}1)$. The performance index is applied to the standardized variables in order to remove the dimensionality and scale effects from the variables.

The aggregated performance index (Equation 18) assumes that all water quality variables have equal importance (equal weights). The performance index can be modified by assigning a weight to each variable in the case where the designers or decision makers assume that some of the variables are more important than others. In this study, equal importance is assumed for all water quality variables. Applying the aggregated performance index (Equation 18), one can evaluate each of the possible combinations. After having examined all possible combinations, one can identify the optimal combination that has the minimum I_a, or one can even provide the decision makers with the rank of all possible combinations. Thus, using such an aggregated performance index, one can overcome the second disadvantage in the conventional approach. Figure 2 illustrates the flow of the analyses as described above.

(Figure 2)

Up to this step, the proposed approach to rationalize the number of water quality variables can identify in a systematic and objective way the optimal combination of variables to be continuously measured and those to be discontinued. The following step is to reconstitute information about discontinued variables, which normally takes place a few years after discontinuation.

3.2.4 Information transfer

In the case of reconstituting information about discontinued water quality variables, the objective is to estimate monthly records of the discontinued variable for the n_2 years $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$, while maintaining the main statistical characteristics of the historical records. In water quality, one may be interested not only in the statistical moments but also in extreme values. If the technique used for record extension introduces a bias into the value of the more extreme-order statistics, this will lead to bias in the estimates of the probability of exceedance of selected extreme values or, conversely, bias in the estimation of distribution percentiles (Hirsch, 1982). In this study, the linear regression and MOVE3 record-extension techniques are carried out to reconstitute information about discontinued variables.

3.2.5 Performance evaluation of the two record-extension techniques

An empirical experiment is designed to examine the usefulness of the simple linear regression and MOVE3 techniques for preserving the statistical characteristics of the discontinued water quality variables. In order to evaluate the performance of the two record-extension techniques, a cross-validation (jackknife) is conducted. In the cross-validation, two years of monthly records are in turn removed from the available ten years of data. All possible combinations of successive or non-successive two-year periods are considered. Thus, from the available ten years of monthly records $C(10,2) = 45$ possible combinations are considered. The values for these two years of monthly observations are then estimated using the two record-extension techniques calibrated with the remaining eight years.

The experimental design is as follows: For each pair of water quality variables identified in the previous step as the variable to be discontinued and its best auxiliary, the two record-extension techniques are applied. Thus, different realizations of extended water quality variable records (i.e., the possible pairs × 45 in all 94 locations) are generated for cross-validation. Important characteristics of the observed and generated records during the extension period are computed for the two record-extension techniques. The evaluation of records generated by the extension techniques involves determining the ability of the techniques to reproduce the various statistical properties of the observed records.

The extended records $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$ are compared to the observed records $\{y_{n_1+1},....,y_{n_2}\}$ based on the estimation of the mean, standard deviation and over the full range of percentiles (from the 5th to the 95th percentile). Moreover, the extended series $\{y_1, y_2,...., y_{n_1}, \hat{y}_{n_1+1},...., \hat{y}_{n_2}\}$ is compared to the observed series $\{y_1, y_2,...., y_{n_1}, y_{n_1+1},...., y_{n_2}\}$ in terms of the estimation of the *y*-time rate of change (trend magnitude). Two methods are used to estimate the trend magnitude; (i) the simple linear regression and (ii) Sen's slope. The slope of the simple linear regression of *y* over time is an indicator of the *y*-time rate of change. The Sen's slope estimate is a nonparametric alternative to the trend magnitude (EPA, 2000). This approach involves computing slopes for all pairs of ordinal time records and then using the median of these slopes as an estimate of the overall slope. In cases where there are *n* time records, where y_i denotes the record value for the i^{th} time record, there will be *n* (*n*-1)/2 possible pairs of time points (i, j) in which $i > j$. The slope for each pair is called a pair-wise slope, b_{ij}, and is computed as $b_{ij} = (y_i\text{-}y_j) / (i - j)$. Sen's slope estimator is then the median of the *n* (*n*-1)/2 pair-wise slopes.

Different performance measures are applied. First, the ratio *U*, of each statistic for the extended series over the historic records is computed. The ratio *U* is used to assess the performance of the record-extension techniques in preserving the historic characteristics. If the ratio *U* for a given statistical parameter is larger than 1, it means that the applied technique overestimates this parameter. If it is less than 1, the technique underestimates the parameter.

Concurrently, three performance measures are used to assess the two record-extension techniques. They are the mean multiplicative error (*MME*), the relative bias (*BIASr*) and the relative root mean square error (*RMSEr*), which are defined as follows:

$$MME = \exp\left[\frac{\sum_{i=n_1+1}^{n_2} \ln \hat{y}_i - \ln y_i}{n_2}\right] \qquad (19)$$

$$BIASr = \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \frac{\hat{y}_i - y_i}{y_i} \qquad (20)$$

$$RMSEr = \sqrt{\frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \left[\frac{\hat{y}_i - y_i}{y_i}\right]^2} \qquad (21)$$

where $\hat{y}_i$ and y_i are, respectively, the estimated and measured values of the dependent variable for $i = n_1 + 1, \ldots n_2$. The *MME* of the logarithms is equal to the geometric mean of ($\hat{y} / y$), and thus, *MME* values that are equal to unity indicate an ideal performance of the record-extension technique. The ratio *U* and *MME* are applied to the logarithms of the extended records, while the *BIASr* and the *RMSEr* are applied to the reverse transformed records (original records). The ratio *U* is used to compare each statistic with its historical value. The three performance measures of equations 19 to 21 are applied to compare the estimated records with the observed records.

4 Results

In this section, results obtained from the application of the proposed approach to the Nile Delta drainage system WQM network are presented. The proposed approach is applied to each of the 94 monitoring sites under study. Results are presented in two subsections. The first subsection presents the results of the application of the proposed approach to the Arin drain monitoring site (EH16), of the Bahr-Hadus drainage catchment, in the eastern Delta. The second subsection presents the results obtained from the empirical experiment to identify the best technique to use in reconstituting information about discontinued variables.

4.1 Rationalization results

Using CA and the criteria developed for the assessment of the correlation coefficient (d_m and d_v), groups of highly correlated water quality variables are identified. Figure 3 shows the cluster tree for EH16, where the x-axis indicates the water quality variables and the y-axis indicates the linkage distance between clusters. The lowest criterion from d_m and d_v (equations 16 and 17) is applied. Results show that $d_m = 0.75$ and $d_v = 0.33$. Thus, the d_v criterion is applied as indicated by the horizontal line at a cutoff distance of 0.33 (Figure 3). Using the higher value (d_m instead of d_v), one cannot guarantee that extension will be profitable within each cluster of variables, i.e. that variables within each cluster can serve as auxiliaries for a precise estimation of the variance for any of the discontinued variables in the same cluster.

Using d_v as a criterion, the water quality variables are divided into 24 clusters, 18 of which are single-variable clusters. These 18 variables should be continuously monitored as the information they provide cannot be estimated from other variables. Six other clusters are considered for further analysis. The Specific conductance (*EC*), Total Dissolved Solids (*TDS*), Sodium (*Na*) and Chlorine (*Cl*) form the first cluster. The second cluster consists of the Calcium (*Ca*) and Sulphate (SO_4). Nitrate (NO_3) and Total Nitrogen (*TN*) form the third cluster. The fourth cluster consists of the Total Coliform (*TCol*) and the Fecal Coliform (*FCol*). Biochemical Oxygen Demand (*BOD*) and Chemical Oxygen Demand (*COD*) form the fifth cluster. The last cluster consists of Total Suspended Solids (*TSS*), Total Volatile Solids (*TVS*) and Turbidity (*Turb*). One should assure that at least one variable from each of the six clusters is continuously measured. Given that the six clusters contain a total of 15 water quality variables, the maximum number of variables to discontinue is 9.

(Figure 3)

Within each multiple-variable cluster, each variable is assumed to be discontinued and its best auxiliary variable is identified based on Equation 2. Four clusters consist of only two variables each. It is clear in these cases that each variable works as an auxiliary for the other variable, and only one variable can be discontinued from each one of these clusters. Two clusters consist of more than two variables each. Each variable is hence considered to be discontinued and Equation 2 is applied to identify its best auxiliary from the other variables of the same cluster.

Using Equation 2, *Na* is found to be the best auxiliary for *Cl* and *TDS*. *Cl* is the best auxiliary for *Na* and *TDS* is the best auxiliary for *EC*. For the last cluster, *TSS* is found to be the best auxiliary for *Turb* and for *TVS*. *Turb* is the best auxiliary for *TSS*. Table 2 shows all alternatives, assuming each variable to be discontinued (first column), the best auxiliary variable is identified (second column). The estimated variance based on Equation 2 is presented in the third column.

If one variable is to be discontinued from the list of variables being measured, the aggregated performance index is applied (Equation 18). The variables listed in Table 2 are ranked based on the aggregated performance index (fourth column). If only one variable is to be discontinued, from a statistical point of view, it can be either *Na* or *Cl*. Discontinuation of either of these two variables leads to an equal performance index. Table 2 provides the rank of the variables to be discontinued from a statistical point of view. Thus, it can be used along with other criteria as support for decision makers and network designers to choose which variable to discontinue. Other criteria may include stakeholders' preference or the significance of the variable in specific studies.

(Table 2)

Where more than one variable is to be discontinued, the aggregated information index is applied to rank the different combinations. For example, if two water quality variables must be discontinued, $C(15,2)=105$ different combinations may be considered. Table 2 shows the first fifteen combinations ranked based on the aggregated information index for the case of two variables to be discontinued (col. 5 to 7), and for the case of three variables to be discontinued (col. 8 to 11). Similarly, the proposed approach can provide decision makers, WQM network designers and stakeholders with a variety of scenarios (combinations) for any number of variables to be discontinued.

The proposed approach can also be applied using spatial correlation between variables measured at different sites. In this case, instead of evaluating the correlations between variables measured at the same site, one can evaluate the correlation between the same variable measured at different monitoring sites. Thus, the decision could be to discontinue the measurement of some variables at some sites, and continuously measure the same variables at auxiliary sites.

4.2 Empirical experiment

The results of the empirical experiment are presented herein. Figures 4 and 5 summarize the results obtained for the ratio *U*. Figure 4 shows the ratio *U* distribution for the estimation of the statistical moments and Figure 5 shows the ratio *U* distribution for the estimation of different non-exceedance percentiles. From the previous step, 962 pairs of variables are considered at all 94 monitoring locations. Box plots in Figures 4 and 5 are constructed from 42872 records. This number of records is less than the expected number (possible pairs × 45 at all 94 monitoring sites). This reduction is mainly due to the absence of the *FCal* and the *TN* measurements in the first two years, and the turbidity in the last two years of the monitoring program.

The box plots in Figures 4 and 5 represent the distribution of the ratio *U* for a given statistic and record-extension technique. The accuracy of each approach can be judged by the degree of dispersion in the box plots, by the closeness of the median to a value of 1 and by the symmetry of the box plot around the value of 1 (Hirsch, 1982; Vogel and Stedinger, 1985).

(Figures 4 and 5)

In Figure 4, box plots represent the distribution of the ratio *U* for the estimation of the mean and the standard deviation. For the estimation of the mean, the regression and MOVE3 techniques lead to median values of *U* equal to 1. Boxes are symmetric around 1 and have almost the same dispersion. Figure 4 shows that the regression technique tends to underestimate the standard deviations. The cross-validation shows that approximately 75 percent of the regression standard deviations are lower than the historical values, with a median value of 0.91. As for the MOVE3 technique, the median value is 1.02. The MOVE3 standard deviation box plot is more symmetric around 1 and shows relatively less dispersion than the one corresponding to the regression.

Figure 5 shows that the median values of the ratio *U* for low percentiles are higher than 1 while those corresponding to high percentiles are lower than 1 for both record-extension techniques. In general, the median values of the ratio *U* for the MOVE3 percentiles are very close to 1. The box plot values range between 0.95 and 1.06 for the regression and between 0.98 and 1.02 for MOVE3. In general, MOVE3 box plots are symmetric around 1 and show relatively less dispersion than those corresponding to the regression. These results suggest that the regression tends to overestimate low percentiles and underestimate high percentiles. MOVE3 reduces the bias exhibited by the regression.

In Figure 5, the box plots representing low percentiles show more dispersion than those corresponding to high percentiles, especially for regression. To further study the impact of the percentile value on estimation error, one needs to study the estimation error ($\hat{y}_i - y_i$) as a function of the values of several water quality records. Figure 6 illustrates the estimation error for the *COD* when *BOD* is used as the auxiliary variable at monitoring site EH16. Figure 6 shows that the error generally increases as *COD* values increase. Figure 7 shows the percentile estimation error and relative error for one trial of the generated records. Results indicate that the low percentile estimation error is lower than the error corresponding to high percentiles. On the other hand, relative error evolves in the opposite direction. These results are more intuitive as one usually expects the error to increase as the variable values increase.

(Figures 6 and 7)

In order to confirm the results presented in Figures 4 and 5, the *t*-test is applied to compare the ratio *U* mean values with 1. Table 3 shows the ratio *U* means and standard deviations for both record-extension techniques. Asterisks shown beside the mean values indicate that the hypothesis that the ratio *U* is equal to one is rejected at the 5% significance level in these cases. Table 3 shows that the tested hypothesis is rejected for all statistical parameters for both record-extension techniques. While the ratio *U* mean values for high and low percentiles show a significant difference from 1 for both record-extension techniques, mean values derived using MOVE3 are always closer to 1 than the corresponding values obtained by regression. In addition, standard deviations obtained from MOVE3 are less than those from regression. These results also confirm that MOVE3 performance is better than the regression for preserving the statistical parameters of the discontinued water quality variables.

(Table 3)

Figure 8 illustrates the *MME* exhibited by the two extension techniques in estimating the water quality concentration percentiles. Results indicate that MOVE3 *MME* values are closer to 1 than regression *MME* values. The curves corresponding to MOVE3 and the regression method intersect at the median as the regression approach overestimates low percentiles and underestimates high percentiles. Figure 9 illustrates the *BIASr* exhibited by the two extension techniques in estimating the water quality concentration percentiles. *BIASr* shows an identical behavior to *MME*, as *BIASr* values for the MOVE3 are, in general, closer to zero than regression

BIASr values. Figures 5, 8 and 9 clearly illustrate the regression overestimation of low concentrations and underestimation of high concentrations, as would be expected from its tendency to produce an extended record with a lower variance than that of the observed record.

(Figures 8 and 9)

Table 4 shows the ratio *U*, *BIASr* and *RMSEr* for the estimation of the trend magnitude in the extended series. Using the simple linear regression slope or Sen's slope, the ratio *U* of the trend magnitude estimated from both the regression and MOVE3 extended series to that estimated from the observed series are close to 1. The *BIASr* shows similar results, as the *BIASr* values are close to zero. For the *RMSEr*, results indicate that there is no significant difference between the regression and MOVE3, though the *RMSEr* for the MOVE3 is higher than that for regression.

(Table 4)

The *MME*, *BIASr* and the *RMSEr* are also computed for individual water quality records, where each error represents the difference between the extended and actual water quality record. Table 5 shows that the *MME* are 1.014 and 1.030 for the regression and MOVE3 techniques, respectively. The *t*-test is applied to determine the significance of the difference between the two techniques based on average errors. The last column in Table 5 represents the probability of accepting the null hypothesis that there is no significant difference between the error mean values based on two-tailed *t*-test. The *t*-test indicates that there is significant difference between the two average *MME* values. Similarly the *t*-test indicates that there are significant differences between the two average *BIASr* and *RMSEr* values. In general, the *RMSEr* shows high values and relatively large differences between the two extension techniques. These high values of the *RMSEr* are due to the back-transformation of the estimated records, which consequently shows the dimensionality and the scale effects of the variables. These results indicate that the regression performance is better than MOVE3 on the estimation of individual water quality monthly records.

(Table 5)

The results in this section can be summarized as follows. The two extension techniques produce extended records that are unbiased in the mean. Furthermore, results indicate that regression substantially reduces variability and MOVE3 tends to preserve variability. The regression technique underestimates high concentration values and overestimates low values. On the other hand, the MOVE3 technique tends to reduce the bias in the estimation of both high and low concentration values. The MOVE3 technique produces extended records that preserve both high and low percentiles relatively well. The linear regression technique is generally better than MOVE3 for the estimation of individual water quality records. The *t*-test shows that there is a significant difference in the average *MME*, *BIASr* and *RMSEr*, which indicates that the regression technique is relatively better than MOVE3 for the estimation of individual water quality records.

5 Conclusions

A statistical approach for the assessment and selection of the optimal combination of water quality variables to measure in surface WQM networks is presented here. Criteria developed from record-augmentation procedures are employed with CA to identify highly correlated variables. An information performance index is then applied to identify optimal combinations of variables to be continuously measured and those to be discontinued.

Simple linear regression and the MOVE3 technique are applied to reconstitute information about discontinued variables using the data of the case study. Different statistical performance measures are used to assess each of the extension techniques and their ability to maintain statistical characteristics of the water quality records. The MOVE3 technique showed better performance in preserving the statistical characteristics of the water quality discontinued variables. Conversely, regression shows better performance for the estimation of individual water quality records. Therefore, the MOVE3 technique is recommended for the reconstitution of information about discontinued water quality variables in the Nile Delta WQM network, while the regression technique is recommended for the reconstitution of missing values.

It can be concluded that the proposed approach can identify, in a systematic and objective way, the optimal combination of variables to be continuously measured and variables to discontinue. Using the proposed approach and the MOVE3 record-extension technique allows the deficiencies inherent in the conventional correlation-regression approach to be overcome. Finally, it can be concluded that the proposed approach could be a useful decision support tool for the optimized selection of water quality variables.

It should be emphasized that the proposed approach provides the decision maker with the optimal combinations of variables to discontinue from a statistical point of view. However, various qualitative criteria could be integrated when deciding which variables to discontinue and which variables to be continuously measured. Such criteria may vary based on the specific nature of the monitoring program under assessment.

In addition, the decision may be in the form that some variables could be determined less frequently instead of being terminated. For instance, instead of terminating the monthly measurements of two variables, the decision could be to measure four variables bimonthly. This can be illustrated as follows:

$$\begin{array}{l} x_1, x_2, x_3, \ldots\ldots, x_{n_1}, x_{n_1+1}, x_{n_1+2}, x_{n_1+3}, x_{n_1+4}, x_{n_1+5}, x_{n_1+6}, \ldots\ldots, x_{n_1+n_2} \\ y_1, y_2, y_3, \ldots\ldots, y_4, \qquad y_{n_1+2} \qquad y_{n_1+4} \qquad y_{n_1+6}, \qquad , y_{n_1+n_2} \end{array}$$

Such a decision may help in reassessing the inter-variables correlation and updating record-extension techniques parameters frequently. This may improve reconstitution of information about the less frequently sampled variables on the short as well as on the long terms. It may also assist in integrating the assessment of the variables to be measured with the sampling frequency.

Parallel to the proposed evaluation of the number of variables to discontinue, a cost analysis can also be introduced. This would help to address the trade-off between the number of water quality variables to be measured on one hand and the sampling frequency and the number of sampling sites on the other hand. Thus, the decision will be either to discontinue more variables in favor of keeping more monitoring sites and increasing the sampling frequency, or to keep more water quality variables while reducing the number of monitoring locations and/or the sampling frequency.

The proposed approach may be modified by using multiple regression or stepwise regression, where more than one auxiliary variable can be used in reconstituting information about the discontinued variable. In this case, the proposed approach should be adapted accordingly. For instance, the criteria to identify the correlation coefficient threshold and the criteria to identify the optimal combination of variables to discontinue should be modified.

Acknowledgment: The authors wish to thank Prof. Shaden Abdel-Gawad, president of the National Water Research Center of Egypt, for providing the data used in this study. The financial support provided by Helwan University, Cairo, Egypt, and the Natural Sciences and Engineering Research Council of Canada (NSERC) is acknowledged. The authors are grateful to the Editor, Laurent Charlet, the Associate Editor, and the two anonymous reviewers whose comments and suggestions greatly improved the quality of the paper.

6 References

[1]. Abdel-Gawad, S.T., Kandil, H.M. and Sadek, T.M. (2004). Water scarcity prospects in Egypt 2000-2050, in: Marquina (ed.) Environmental Challenges in the Mediterranean 2000-2050, Dordrecht: Kluwer Academic Publishers, 187 - 203.

[2]. Alley, W.M. and Burns, A.W. (1983). Mixed-station extension of monthly streamflow records. Journal of Hydraulic Engineering, 109 (10), 1272 - 1284.

[3]. Berryman, D., Bobée, B., Cluis D. and Haemmerli, J. (1988). Nonparametric Tests for Trend Detection in Water Quality Time Series. Water Resources Bulletin, 24(3), 545 - 556.

[4]. Briggs, J.C. and Ficke, J.F. (1978). Quality of rivers of the United States, (1975) water year- based on the National Stream Quality Accounting Network, U.S. Geological Survey Open-File Report 78-200, 436 p.

[5]. Cochran, W.G. (1953). Sampling Techniques, John Wiley, New York, p. 428.

[6]. Draper, N.R. and Smith, H. (1966). Applied regression analysis, John Wiley, New York, 736 p..

[7]. DRI (Drainage Research Institute) - MADWQ, (1998). Monitoring and analysis of drainage water quality in Egypt, Interim Report, Cairo.

[8]. Engelman, R. and Le Roy, P. (1993). Sustaining water, population and the future of renewable water supplies. Population Action International, Population and Environment Program, Washington, D.C., 302 - 318.

[9]. EPA, United States Environmental Protection Agency, (2000). Guidance for data quality assessment, EPA QA/G-9, QA00 update, EPA/600/R-96/084. 219 p.

[10]. Frenken, K., 2005. Irrigation in Africa in Figures, Aquastat Survey (2005). Food & Agriculture Org, Rome, Italy, 88 p.

[11]. Harmancioglu, N.B., Fistikoglu, O., Ozkul, S.D., Singh, V.P. and Alpaslan, M.N. (1999). Water Quality Monitoring Network Design. Kluwer Academic Publishers, Dordrecht, the Netherlands, 290 p.

[12]. Harmancioglu, N.B., Ozer, A. and Alpaslan, N. (1987). Procurement of Water quality information (in Turkish). IX. Technical Congress of Civil Engineering, Proceedings, the Turkish Society of Civil Engineers, II, 113 - 129.

[13]. Harmancioglu, N.B. and Yevjevich, V. (1986). Transfer of Information among Water Quality Variables of the Potomac River, Phase III: Transferable and Transferred Information. Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 1986, 81 p.

[14]. Harmancioglu, N.B. and Yevjevich, V. (1987). Transfer of hydrologic information among river points. Journal of Hydrology, 91, 103 - 118.

[15]. Hirsch, R.M., 1979. An evaluation of some record reconstruction techniques. Water Resources Research 15, 1781–1790.

[16]. Hirsch, R.M. (1982). A comparison of four streamflow record extension techniques. Water Resources Research, 18(4), 1081 - 1088.

[17]. Lettenmaier, D.P. (1988). Multivariate nonparametric tests for trend in water quality, AWRA, Water Resources Bulletin (24)3, 505 - 512.

[18]. Matalas, N.C. and Jacobs, B. (1964). A correlation procedure for augmenting hydrologic data, *U.S. Geol. Surv. Prof. Pap.*, 434-E, E1-E7.

[19]. McKenzie, S.W. (1976). Long-term water quality trends in Delaware streams: U.S. Geological Survey, Report 76-71, p.85.

[20]. Moog, D.B. and Whiting P.J. (1999). Streamflow record extension using power transformations and application to sediment transport, Water Resources Research, vol. 35 (1), 243 - 254.

[21]. MWRI, Ministry of water resources and Irrigation (1997). Review of Egypt's Water Policies, Strengthening the Planning Sector Project, Ministry OF Water Resources and Irrigation, Cairo, Egypt.

[22]. NAWQAM, National Water Quality and Availability Management Project. (2001). Evaluation and Design of Egypt National Water Quality Monitoring Network, Technical

Report No.: WQ-TE-0110-005-DR, NAWQAM, National Water Research Center, Cairo, Egypt.

[23]. Ouarda, T.B.M.J., Rasmussen, P.F., Bobée, B., and Morin, J. (1996) Ontario Hydrometric Network Rationalization, Statistical Considerations, Research Report No. R-470, National Institute for Scientific Research, INRS-ETE, University of Québec, Québec, Canada, 75 p.

[24]. Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V. (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[25]. Strobl, R.O. and Robillard, P.D. (2008). Network design for water quality monitoring of surface freshwaters: A review, Journal of Environmental Management, 87, 639 - 648.

[26]. Sydor, K. (1998). Comparison of parametric and nonparametric streamflow record extension techniques, MSc thesis, University of Manitoba, Winnipeg, Manitoba, Canada, 268 p.

[27]. Vogel, R.M. and Stedinger, J.R. (1985). Minimum variance streamflow record augmentation procedures. Water Resources Research, 21(5), 715 - 723.

[28]. Ward, R.C., Loftis, J.O. and McBride, G.B. (1990). Design of Water Quality Monitoring systems, Van Nostrand Reinhold, New York, USA, 231 p.

[29]. Wolf, P. (2000). Irrigated agriculture in Egypt - Notes of an external observer, Proceedings Symposium "Sustainable Agriculture and Rural Development in Egypt" Witzenhausen, University of Kassel, Germany.

[30]. Yevjevich, V. and Harmancioglu, N.B. (1985). Modeling Water Quality Variables of Potomac River at the Entrance to its Estuary, Phase II (Correlation of Water Quality Variables within the Framework of Structural Analysis). Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 59 p.

Table 1. Water Quality variables measured at the Egyptian national WQM network

Water Quality Variable	**Symbol**	**Unit**	**Water Quality Variable**	**Symbol**	**Unit**
Biochemical Oxygen Demand	*BOD*	mg/l	Streamflow	*Q*	m^3/sec
Chemical Oxygen Demand	*COD*	mg/l	Temperature	*T*	^oC
Dissolved Oxygen	*DO*	mg/l	Acidity	*pH*	-
Specific Conductance	*EC*	dS/m	Total Suspended Solids	*TSS*	mg/l
Total Dissolved Solids	*TDS*	mg/l	Total Volatile Solids	*TVS*	mg/l
Calcium	*Ca*	mg/l	Turbidity	*Turb*	NTU
Magnesium	*Mg*	mg/l	Visibility disc	*Vis*	cm
Sodium	*Na*	mg/l	Total Coliform	*TColi*	MPN/100ml
Potassium	*K*	mg/l	Fecal Coliform	*FColi*	MPN/100ml
Bicarbonate	HCO_3	mg/l	Cadmium	*Cd*	mg/l
Sulphate	SO_4	mg/l	Manganese	*Mn*	mg/l
Chloride	*Cl*	mg/l	Copper	*Cu*	mg/l
Nitrate	NO_3	mg/l	Iron	*Fe*	mg/l
Ammonium	NH_4	mg/l	Zinc	*Zn*	mg/l
Total Phosphorus	*TP*	mg/l	Nickel	*Ni*	mg/l
Total Nitrogen	*TN*	mg/l	Boron	*B*	mg/l
			Lead	*Pb*	mg/l

Table 2. Variables to discontinue and their best auxiliary variables

One variable discontinued	Best auxiliary	$Var(\hat{\mu})$	I_a%	Two variables		I_a%	Three variables			I_a%
Cl	*Na*	0.0838	2.9818	*Cl*	*TDS*	2.4406	*Cl*	*TDS*	*TVS*	2.2793
Na	*Cl*	0.0838	2.9818	*Na*	*EC*	2.4420	*Cl*	*TDS*	*Turb*	2.2801
TDS	*Na*	0.0842	2.9825	*Cl*	*EC*	2.4420	*Cl*	*TDS*	*TSS*	2.2806
EC	*TDS*	0.0877	2.9872	*Na*	*TVS*	2.4424	*Na*	*EC*	*TVS*	2.2808
Turb.	*TSS*	0.0872	2.9874	*Cl*	*TVS*	2.4424	*Cl*	*EC*	*TVS*	2.2808
TSS	*Turb.*	0.0872	2.9877	*TDS*	*TVS*	2.4428	*Na*	*Turb*	*EC*	2.2815
TVS	*TSS*	0.0880	2.9878	*Na*	*Turb*	2.4431	*Cl*	*Turb*	*EC*	2.2815
Ca	SO_4	0.0897	2.9918	*Cl*	*Turb*	2.4431	*Na*	*Turb*	*TVS*	2.2819
SO_4	*Ca*	0.0897	2.9918	*TDS*	*Turb*	2.4435	*Cl*	*Turb*	*TVS*	2.2819
COD	*BOD*	0.0924	2.9952	*Na*	*TSS*	2.4437	*Na*	*TSS*	*EC*	2.2820
BOD	*COD*	0.0924	2.9952	*Cl*	*TSS*	2.4437	*Cl*	*TSS*	*EC*	2.2820
TN	*NO3*	0.1145	3.0157	*TDS*	*TSS*	2.4440	*TDS*	*Turb*	*TVS*	2.2823
Fcoli	*Tcoli*	0.1165	3.0191	*EC*	*TVS*	2.4442	*Cl*	*TDS*	*Ca*	2.2831
NO3	*TN*	0.1145	3.0305	*Turb*	*EC*	2.4450	*Cl*	*TDS*	*SO4*	2.2831
Tcoli	*Fcoli*	0.1165	3.0324	*Turb*	*TVS*	2.4453	*Turb*	*EC*	*TVS*	2.2837

Table 3. *t*-test for the ratio *U* of various statistics

Technique	**Regression**		**MOVE3**	
Statistic	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Mean	1.016*	0.149	1.019*	0.143
Standard deviation	0.913*	0.259	1.082*	0.388
5th percentile	1.228*	0.588	1.088*	0.570
10th	1.176*	0.477	1.082*	0.436
15th	1.150*	0.425	1.080*	0.385
20th	1.125*	0.377	1.072*	0.343
25th	1.104*	0.330	1.064*	0.309
30th	1.088*	0.299	1.059*	0.288
35th	1.067*	0.262	1.048*	0.262
40th	1.052*	0.237	1.040*	0.241
45th	1.034*	0.211	1.029*	0.220
50th	1.017*	0.183	1.019*	0.194
55th	1.008*	0.171	1.015*	0.181
60th	0.996*	0.163	1.006*	0.171
65th	0.983*	0.158	0.996*	0.163
70th	0.972*	0.154	0.988*	0.157
75th	0.961*	0.150	0.981*	0.149
80th	0.951*	0.146	0.974*	0.142
85th	0.942*	0.144	0.968*	0.136
90th	0.932*	0.141	0.962*	0.131
95th	0.918*	0.142	0.953*	0.128

* The hypothesis that the ratio U is equal to 1 is rejected at the 5% level of significance.

Table 4. Average error measures for the trend magnitude estimation

Error measures	**Regression slope**		**Sen's slope**	
	Regression	**MOVE3**	**Regression**	**MOVE3**
ratio *U*	0.988	0.989	0.994	1.002
BIASr	-0.012	-0.011	-0.006	0.002
RMSEr	0.639	0.747	0.600	0.702

Table 5. Average error measures for record-extension techniques

Error measure	**Regression**	**MOVE3**	***t***	**2-tailed sig.**
MME	1.014	1.030	-10.77	5.1E-27
BIASr	12.56	14.97	-6.23	0
RMSEr	47.92	55.92	-8.78	0
***BIASr* (ln values)**	4.12	3.04	7.54	4.7^{E}-14
***RMSEr* (ln values)**	18.14	19.41	-5.75	8.5^{E}-9

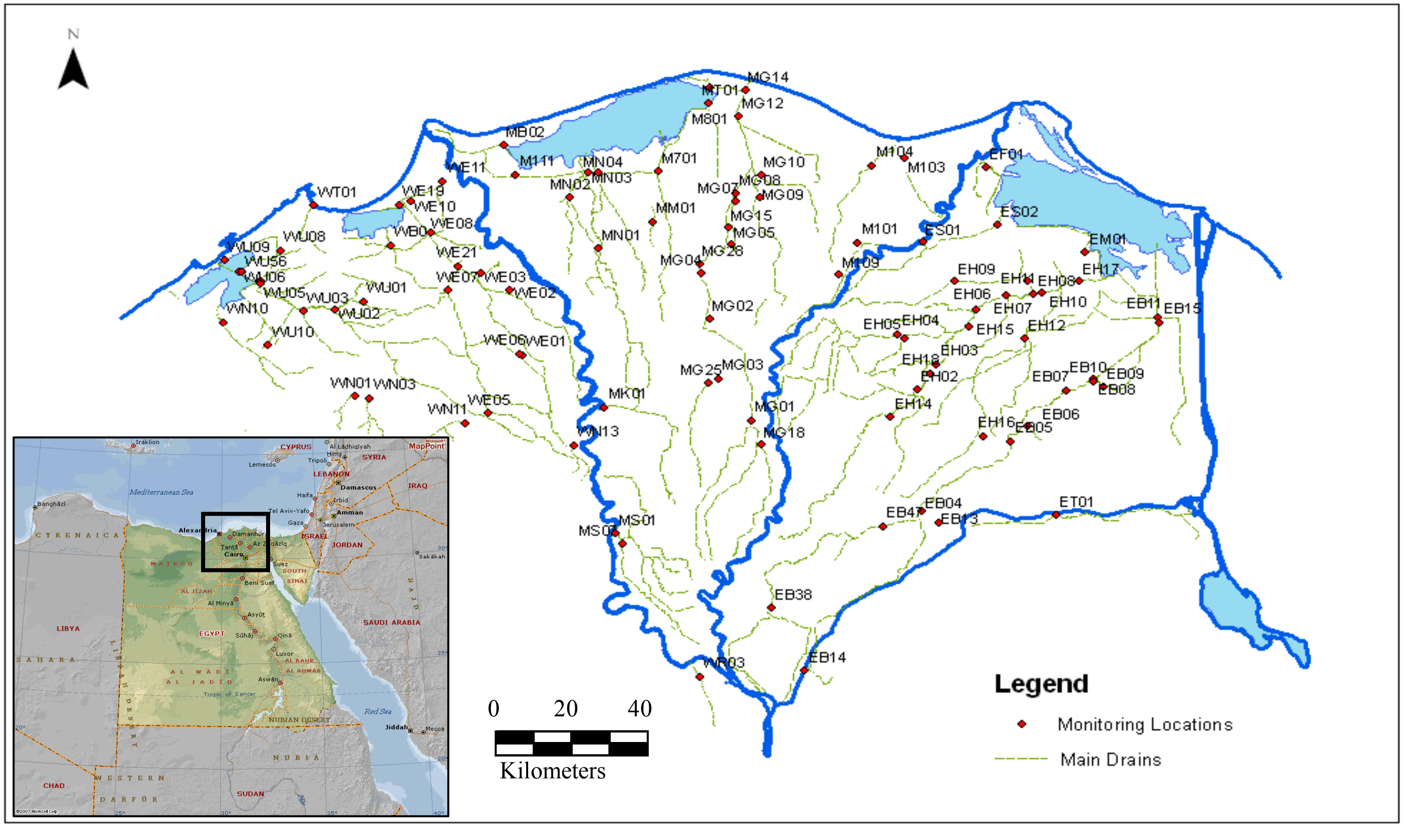


Figure 1. The Nile Delta surface WQM sites (source: NWRC)

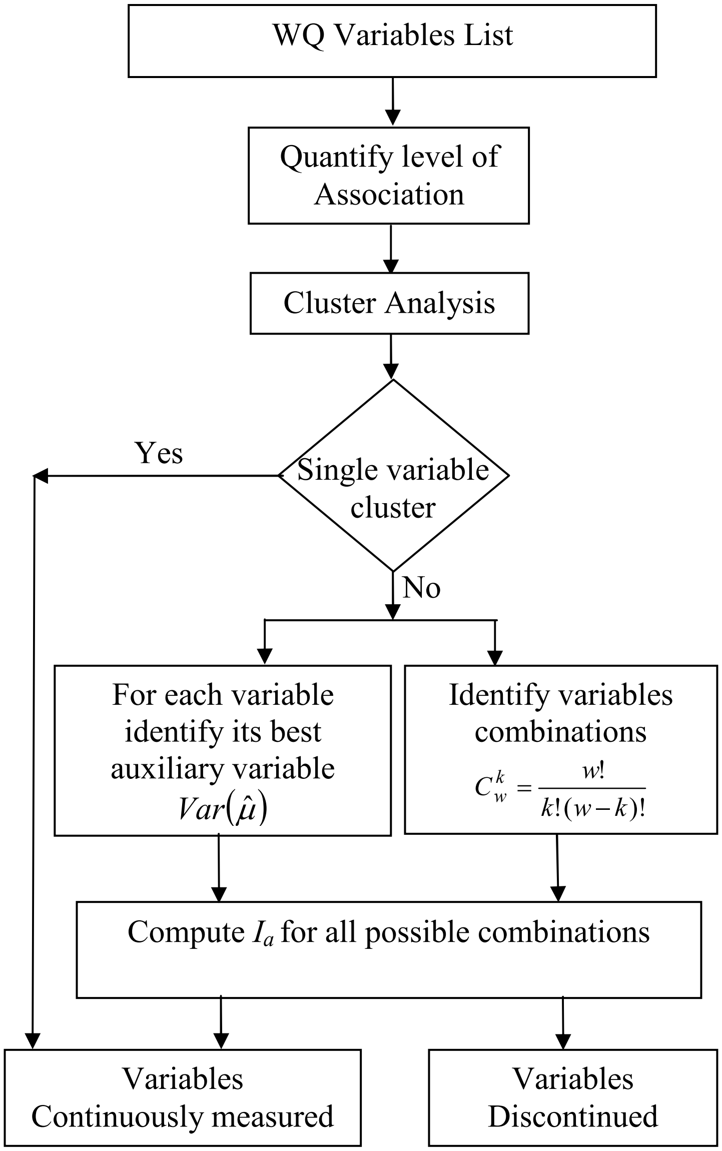

WQ Variables List
Quantify level of Association
Cluster Analysis
Single variable cluster
Yes
No
For each variable identify its best auxiliary variable $Var(\hat{\mu})$
Identify variables combinations $C_w^k = \frac{w!}{k!(w-k)!}$
Compute I_a for all possible combinations
Variables Continuously measured
Variables Discontinued

Figure 2. Flow chart of the proposed rationalization approach

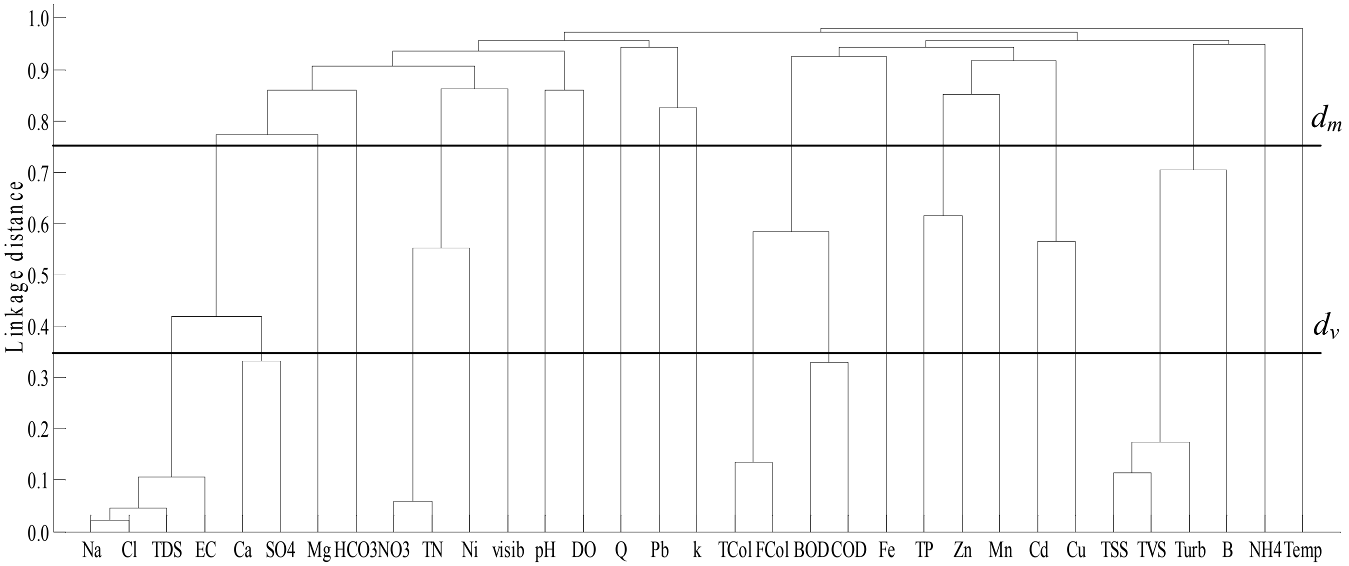


Figure 3. Cluster tree for water quality variables at the Arin drain monitoring site (EH16)

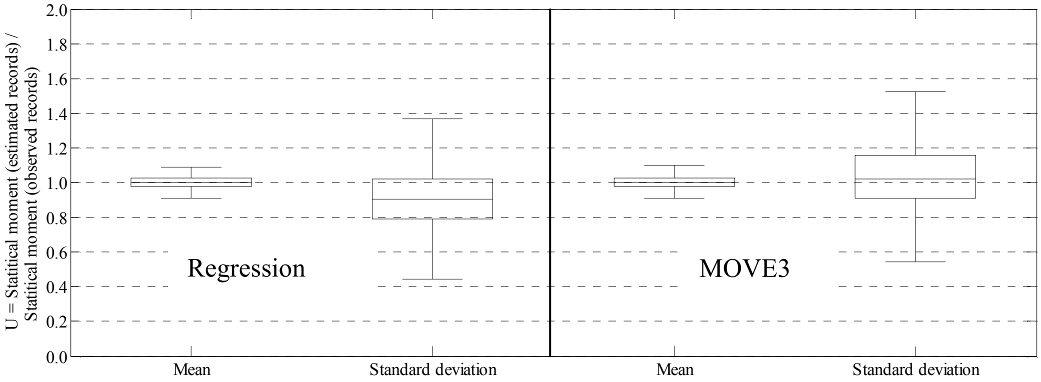


Figure 4. Box plots of the ratio *U* for the mean and standard deviation

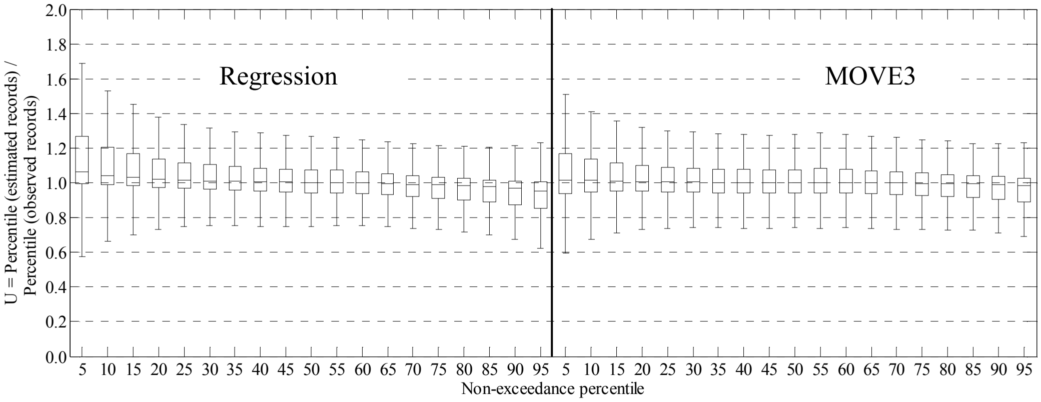


Figure 5. Box plots of the ratio *U* for different percentiles

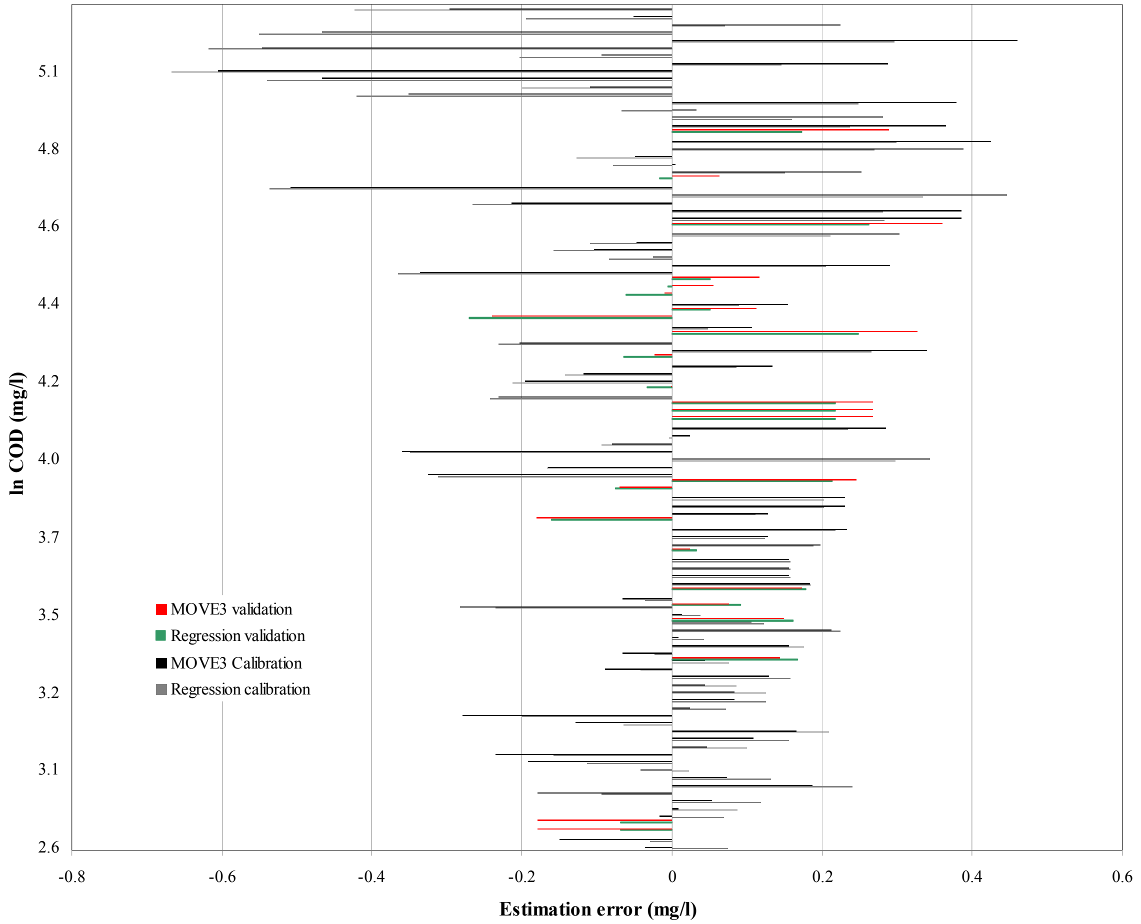


Figure 6. Estimation error vs. logarithm of *COD* values at EH16

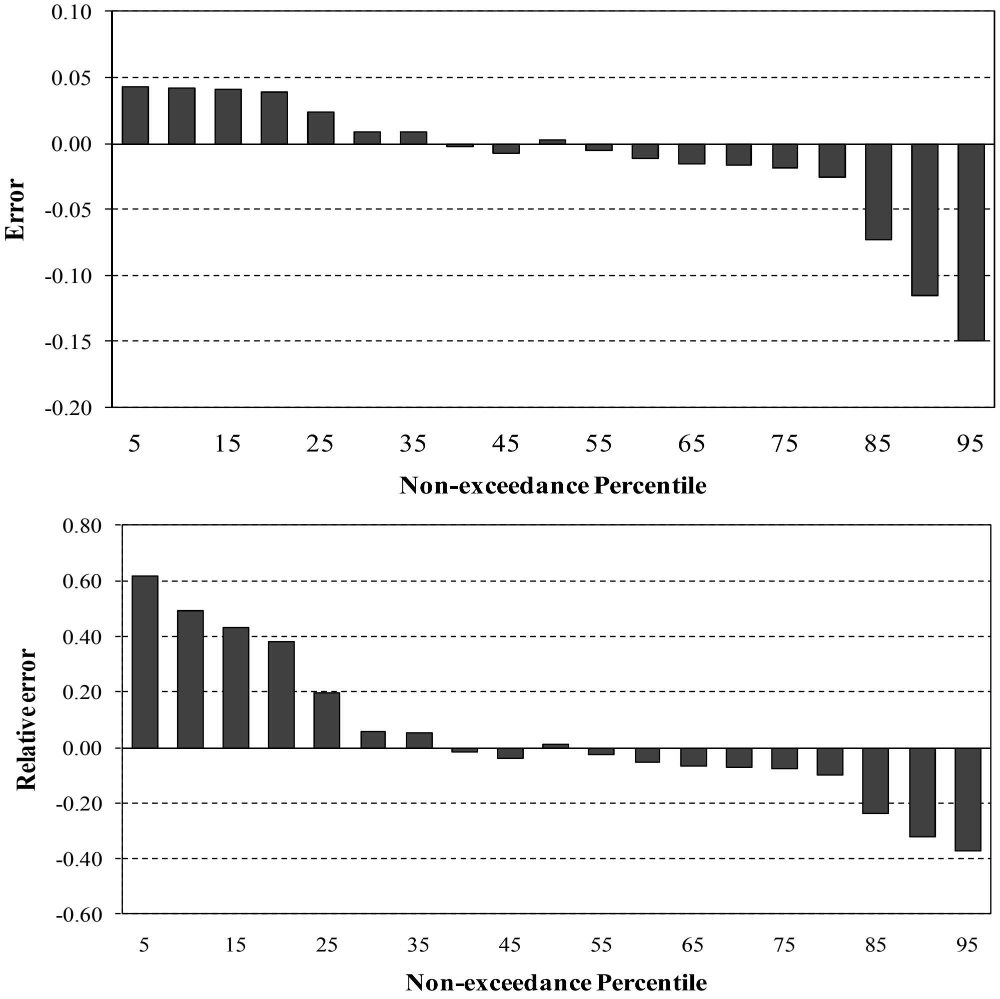


Figure 7. Error and relative error on non-exceedance percentile estimation

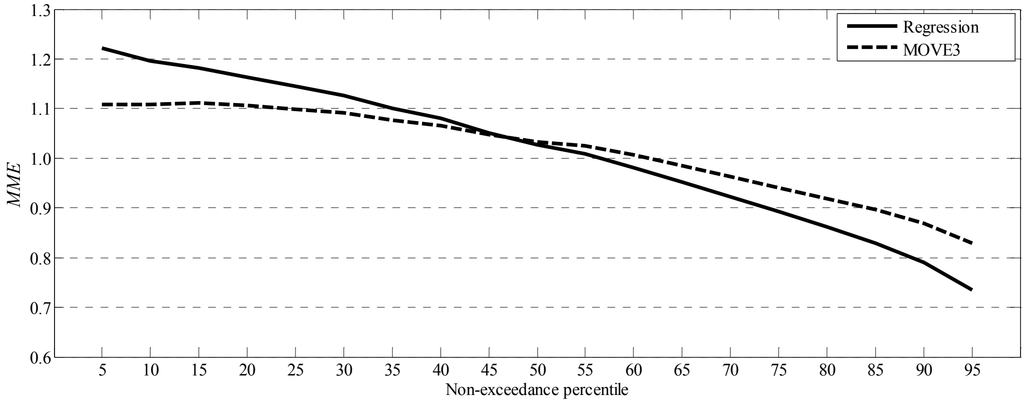


Figure 8. *MME* of the tested extension techniques in estimating various percentiles

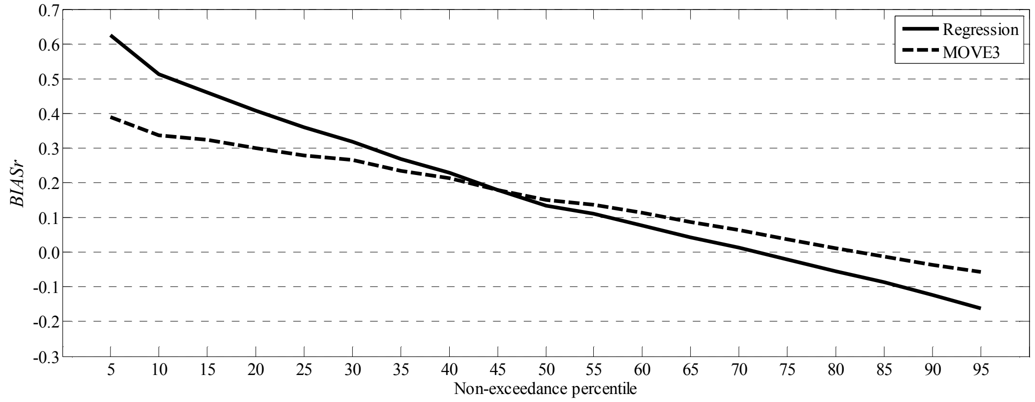


Figure 9. *BIASr* of the tested extension techniques in estimating various percentiles

Article III. Comparison of record-extension techniques for water quality variables

Comparison of record-extension techniques for water quality variables

Bahaa Khalil[1,2], Taha, B.M.J. Ouarda[1] and André St-Hilaire[2]

[1] Irrigation and Hydraulics department, Faculty of Engineering, Helwan University, Cairo, Egypt

[2]Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE,
Québec City, Canada.

Submitted to Journal of Hydrology

January 2011

Abstract

The extension of records at monthly, weekly or daily time steps at a short-record gauge from another continuously measured gauge is termed “record extension”. Ordinary least squares regression (OLS) of the flows, or any hydrological variable, is a traditional and still common record-extension technique. However, its purpose is to generate optimal estimates of each daily (or monthly) record, rather than the population characteristics, for which the OLS tends to underestimate the variance. The line of organic correlation (LOC) was developed to correct this bias. On the other hand, the Kendall-Theil robust line (KTRL) method has been proposed as an analogue of OLS, its advantage being its robustness in the presence of extreme values. In this study, four record-extension techniques are described, and their properties are explored. These techniques are OLS, LOC, KTRL and a new technique (KTRL2), which includes the advantage of LOC in reducing the bias in estimating the variance and the advantage of KTRL in being robust in the presence of extreme values. A Monte-Carlo study is conducted to examine these four techniques for bias, standard error of moment estimates and full range of percentiles. An empirical examination is made of the preservation of historic water quality concentration characteristics using records from the Nile Delta water quality monitoring network in Egypt. The Monte-Carlo study showed that the OLS and KTRL techniques are shown to have serious deficiencies as record-extension techniques, while the LOC and KTRL2 techniques show results that are nearly similar. Using real water quality records, the KTRL2 is shown to lead to better results than the other techniques.

Key words –Record extension; Ordinary least squares; Line of organic correlation; Kendall-Theil Robust Line; Monitoring network.

1 Introduction

Water quality monitoring network rationalization consists of reducing the number of monitoring sites, the number of samples and/or the number of water quality variables measured. The reduction in water quality variables is usually based on correlation analysis. Consideration should be given to reduce the number of variables sampled without substantial loss of information. Fewer variables are easier and less costly to analyze. The establishment of dependencies or correlations between various water quality variables saves time and effort (Strobl and Robillard, 2008).

A review of the literature reveals that correlation and regression analyses are commonly used to reduce the number of water quality variables being measured (e.g., Sanders et al., 1983; Yevjevich and Harmancioglu, 1985; Harmancioglu and Yevjevich, 1986, 1987). A correlation analysis is used to assess the level of association among the variables. If two variables show high correlation, this correlation is then an indication that some of the information produced may be redundant. Consequently, in practice, one may discontinue the monitoring of one variable and only continuously measure the other. Then, the ordinary least squares regression (OLS) technique is used to reconstitute information about the discontinued variable, using the continuously measured one as an explanatory variable.

The use of OLS regression often results in underestimation of the variance in extended records (Hirsch, 1982; Alley and Burns, 1983). Additionally, if the technique used for record extension introduces a bias into the value of the more extreme order statistics, this will lead to bias in the estimates of the probability of exceedance of selected extreme values or, conversely, to a bias in the estimation of distribution percentiles (Hirsch, 1982). In water quality, one may be interested in, not only the statistical moments, but also percentiles, which can be used to assess compliance with standards or objectives (Khalil et al., 2010).

The line of organic correlation (LOC) was proposed as a linear fitting procedure in hydrology by Kritskiy and Menkel (1968). This line has also been called the "maintenance-of-variance extension" or MOVE (Hirsch, 1982). The LOC is widely applied to streamflow record extension at short-gauged stations (e.g., Hirsch, 1982; Vogel and Stedinger, 1985, Moog et al., 1999). The main advantage of the LOC is that the cumulative distribution function of the predictions, including the variance and probabilities of extreme events such as floods and droughts, approximates the distributions of the actual records they are intended to represent (Helsel and Hirsch, 2002).

The Kendall-Theil robust line (KTRL) has the desirable properties of a nonparametric estimator. It is almost as efficient as the parametric estimator (e.g. OLS) when all assumptions of normality are met and is even much more efficient when those assumptions are not met (Helsel and Hirsch, 2002). When the data or their transforms exhibit a linear pattern, constant variance and near-normality of residuals, the KTRL and OLS will give nearly identical results (Hirsch et al., 1991). However, when extreme values exist, the KTRL will produce a line with greater efficiency than OLS (Hirsch et al., 1991; Helsel and Hirsch, 2002).

The main goal of this study is to assess the usefulness of four record-extension techniques in reconstituting information about discontinued water quality variables. These techniques are OLS, LOC, KTRL and a new technique that combines the advantage of LOC in preserving the cumulative distribution function of predictions as well as the advantage of the KTRL in being robust in the existence of extreme values. This new technique will be referred to hereafter as KTRL2.

In the following section, the theoretical background is provided. Section 3 details the Monte-Carlo simulations and empirical experiments. The results obtained are presented and discussed in section 4. Finally, conclusions are presented in section 5.

2 Theoretical background

Assume that the measured variable y has n_1 years of data and that the measured variable x has $n_1 + n_2$ years, of which n_1 are concurrent with the data observed for y, illustrated as follows:

$$\begin{array}{l} x_1, x_2, x_3, \ldots\ldots, x_{n_1}, x_{n_1+1}, x_{n_1+2}, \ldots\ldots, x_{n_1+n_2} \\ y_1, y_2, y_3, \ldots\ldots, y_{n_1} \end{array}$$

For water quality variables reduction, one can consider that the initial assessment and variable selection occurred in year n_1. After n_1 years, the measurement of variable y is discontinued, and the variable x continues to be measured. Assume that after n_2 years, we desire to reconstitute information about the variable y. To estimate records of the discontinued variable y for the period $n_1 + 1$ through n_2 years, a record extension technique can be used.

In this section, four record extension techniques are described. These are the ordinary least squares regression (OLS), line of organic correlation (LOC), Kendall-Theil robust line (KTRL) and a new technique Kendall-Theil robust line 2 (KTRL2).

2.1 Ordinary Least Squares (OLS) regression

Ordinary Least Squares (OLS), commonly referred to as linear regression, is an important tool for the statistical analysis of water quality data. It is used to describe the co-variation between a variable of interest and one or more other variables. OLS of y on x can be illustrated as follows:

$$\hat{y}_i = a + bx_i \tag{1}$$

where $\hat{y}_i$ is the estimated value of y for $i = n_1 + 1, ..., n_2$, a and b are the intercept and slope of the regression equation, respectively. The parameters a and b are the values that minimize the squared error in the estimated y values. The solution of a and b is found by solving the normal equations (Draper and Smith, 1966). The optimal solution to Equation 1 is:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y_1} / s_{x_1})\,(x_i - \bar{x}_1) \tag{2}$$

where $\bar{y}_1$ and $\bar{x}_1$ are the mean values of y_1 and x_1: the series for the concurrent records ($i = 1,...,n_1$), s_{y_1} and s_{x_1} are the standard deviations of y_1 and x_1 and r is the product-moment correlation coefficient between y_1 and x_1.

2.2 Line of Organic Correlation (LOC)

An alternative to OLS is to specify that the extension equation be of the form given in Equation 1, while *a* and *b* are to be set not to minimize the squared error, but rather to maintain the sample mean and variance. The idea that led to the development of LOC was that of finding the values of *a* and *b* in Equation 1 that satisfy the following equations (Hirsch, 1982):

$$\sum_{i=1}^{n_1} \hat{y}_i = \sum_{i=1}^{n_1} y_i \tag{3}$$

$$\sum_{i=1}^{n_1} (\hat{y}_i - \bar{y}_1)^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 \tag{4}$$

One such solution is:

$$\hat{y}_i = \bar{y}_1 + (s_{y_1} / s_{x_1})\,(x_i - \bar{x}_1) \tag{5}$$

The LOC has been called by various names, e.g., “reduced major axis” (Kermack and Haldane, 1950), “allometric relation” (Teisser, 1948), “geometric mean functional regression” (Halfon, 1985) and “maintenance-of-variance extension” or MOVE (Hirsch, 1982). Hirsch (1982) carried out a Monte-Carlo experiment to examine the OLS and LOC for bias and standard error of moments estimates and order statistics. Results showed that OLS cannot be expected to provide records with the appropriate variability, and that the LOC is an effective technique in terms of producing time series with properties (such as variance and extreme order statistics) most like the properties of the records they are intended to present.

2.3 Kendall–Theil Robust Line (KTRL)

The Kendall-Theil robust line (*KTRL*) is based on the Kendall's rank correlation coefficient (*tau*), which is used to test for any monotonic, and not necessarily linear, dependence of *y* on *x*. Related to *tau* is a robust nonparametric line applicable when *y* is linearly related to *x*. This line will not depend on the normality of residuals for the validity of significance tests, nor will it be strongly affected by extreme values, in contrast to OLS (Helsel and Hirsch, 2002). The robust estimate of slope for this nonparametric fitted line was first described by Theil (1950). The Theil slope estimate is computed by comparing each data pair to all others in a pair-wise fashion. An *n*-element data set of (*x*, *y*) pairs will result in *n (n-1)/2* pair-wise comparisons. For each of these comparisons, a slope $\Delta y / \Delta x$ is computed. The median of all possible pair-wise slopes is taken as the nonparametric slope estimate (b_K):

$$b_K = median\,\frac{y_j - y_i}{x_j - x_i} \qquad (6)$$
$$\forall\, i < j \qquad i = 1,2,......n-1 \qquad j = 2,3.........,n$$

The intercept (a_K) is defined as follows:

$$a_K = median\,(y) - b_K * median\,(x) \qquad (7)$$

This formula assures that the fitted line goes through the point (median (*x*), median (*y*)). This is analogous to OLS, where the fitted line always goes through the point (mean (*x*), mean (*y*)). b_K is an unbiased estimator of the slope of a linear relationship, and b from OLS is also an unbiased estimator. However, the variances of the estimators differ. When the residuals from the true linear relationship are normally distributed, OLS is slightly more efficient (has a lower variance) than KTRL. When residuals depart from normality (i.e., are skewed or prone to extreme values), then b_K can be much more efficient than the OLS slope (Hirsch et al., 1991; Helsel and Hirsch, 2002).

2.4 Kendall-Theil Robust Line 2 (KTRL2)

Similarly, the new record-extension technique (KTRL2) proposed in this paper follows KTRL but with a modification of the intercept (a_q) and slope (b_q). Its objective is to produce records with sample cumulative distribution functions (CDFs) that are close approximates of the CDFs of actual records. KTRL2's developmental goal is to find the values of a_q and b_q in Equation 1 that minimize the error in estimating the y percentiles; a_q and b_q are defined as follows:

$$b_q = median\,\frac{q(y)_j - q(y)_i}{q(x)_j - q(x)_i} \qquad (8)$$
$$\forall\, i < j \qquad i = 5^{th}, 10^{th}, \ldots\ldots 90^{th} \qquad j = 10^{th}, 15^{th} \ldots\ldots, 95^{th}$$

The intercept is defined as follows:

$$a_q = median(y) - b_q * median(x) \qquad (9)$$

where $q(y)$ and $q(x)$ are the percentiles of y and x estimated during the period of concurrent records. Percentiles are obtained for the range of the 5th, 10th,… to the 95th percentile. Thus, a set of 19 (x, y) pairs of percentiles will result in 171 ($n\ (n\text{-}1)/2$ = 19 (19-1)/2) pair-wise comparisons. For each of these comparisons, a slope $\Delta y / \Delta x$ is computed and the median of the 171 possible pair-wise slopes is taken as the slope estimate. Thus, the objective is to minimize the error in estimating the y percentiles rather than minimizing the error in estimating the y records.

3 Evaluation experiments

In order to evaluate the four record-extension techniques, Monte-Carlo and empirical experiments are conducted. Monte-Carlo experiments allow for a comparison and evaluation of different record-extension techniques using records with predefined distributional and statistical properties. The empirical experiment allows for an evaluation of the four record-extension techniques using real data. In both the Monte-Carlo and empirical experiments, the four record extension techniques are evaluated under different levels of correlation between x and y as well as different sizes of concurrent records (n_1).

3.1 Monte-Carlo experiment

In the Monte-Carlo experiment, the experimental design carried out by Hirsch (1982) for the comparison of OLS and LOC is followed. The x and y variable sequences of 120 cases were generated from a bi-variate normal distribution with $\mu_x = \mu_y = 0$ and $\sigma_x^2 = \sigma_y^2 = 1$. Three cross-correlation coefficients and different combinations of the number of records during the concurrent period (n_1) and the period to be estimated (n_2) are considered. Monte-Carlo experiments are carried out for (n_1, n_2) values of (96, 24), (72, 48), (48, 72) and (24, 96) and for correlation coefficient values (ρ) of 0.5, 0.7 and 0.9. Monte-Carlo experiments are conducted with each of the twelve different combinations of ρ and (n_1, n_2) to evaluate the ability of the four record-extension techniques in reproducing the various statistical properties of the population considered in the extended series. The extended series are evaluated based on the estimation of the mean, standard deviation and over the full range of percentiles (from the 5th to the 95th percentile).

3.2 Empirical experiment

The main objective of the empirical experiment is to evaluate the four record extension techniques using real water quality records under different levels of correlation as well as different sizes of concurrent records (n_1). In the empirical experiment, water quality records from the Nile Delta drainage system water quality monitoring network are used. The monitoring network consists of 94 monitoring locations (Figure 1), at which monthly samples have been taken since August 1997. Monthly water quality records, available for electric conductivity (EC), total suspended solids (TDS) and Chlorine (Cl) at the 94 monitoring locations from August 1997 to July 2007, are used in the empirical experiment. These three variables are chosen in order to be able to evaluate the record-extension techniques under different levels of correlation. The preliminary analysis indicates that the correlation between the EC and TDS is higher than that between the EC and Cl in most of the 94 monitoring locations. The EC is used as an explanatory variable to estimate the TDS and Cl using the four record-extension techniques.

(Figure 1)

An empirical experiment is designed to examine the utility of the four record-extension techniques in reproducing records that preserve the statistical characteristics of the observed TDS and Cl records. Three different combinations of the size of records during the concurrent period (n_1) and the extension period (n_2) are considered for each of the two estimation models (TDS and Cl). Empirical experiments are carried out for (n_1, n_2) values of (108, 12), (84, 36) and (60, 60). In order to evaluate the performance of the four record-extension techniques, a cross-validation (jackknife) method is applied. In the cross-validation, when (n_1, n_2) equals (108, 12), one year of monthly records is removed from the available ten years of data. The monthly values for the removed year are then estimated using the four record-extension techniques calibrated with the nine remaining years. Similarly, for (n_1, n_2) values of (84, 36) and (60, 60), the cross-validation method is applied to estimate three and five years of monthly data respectively.

At each of the 94 monitoring sites, the four record-extension techniques are applied to estimate the TDS and Cl using the EC as an explanatory variable. Thus, for each of the two estimation models considered, when (n_1, n_2) equals (108, 12), 940 (94 locations × 10 different sample combinations) different realizations of extended water quality variable records are considered. For (n_1, n_2) values of (84, 36) and (60, 60), possible combinations of successive or non-successive three and five years are considered respectively. Consequently, from the available ten years of monthly records $C(10,3)$ = 120 and $C(10,5)$ = 252 possible combinations are considered. Thus, 11280 (94 locations × 120 different sample combinations) different realizations of extended water quality variable records are considered when (n_1, n_2) equals (84, 36) and 23688 (94 locations × 252 different sample combinations) are considered when (n_1, n_2) equals (60, 60).

In general, water quality variables exhibit a positive skew (Lettenmaier, 1988; Berryman et al., 1988), which is also confirmed in our case by a preliminary analysis of the Nile Delta data. Consequently, in the empirical experiment, the record-extension techniques are applied to the logarithms of the EC, TDS and Cl.

3.3 Evaluation criteria

In the Monte-Carlo experiments, the utility of the four record-extension techniques in producing extended data series $\{y_1, y_2, ...y_{n_1}, \hat{y}_{n_1+1}, ...\hat{y}_{n_1+n_2}\}$ that preserve the population mean, standard deviation and the full range of percentiles is evaluated. In the empirical experiments, the utility of the techniques in reproducing the statistics of the observed data series $\{y_1, y_2, ...y_{n_1}, y_{n_1+1}, ...y_{n_1+n_2}\}$ is evaluated. Two metrics are used to evaluate the performances of the four record-extension techniques. These are the bias (*BIAS*) and the root mean square error (*RMSE*), which can be defined as follows:

$$BIAS = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i - y_i \qquad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2} \qquad (11)$$

where $\hat{y}_i$ and y_i are, respectively, the estimated and the measured statistic of the dependent variable for $i = 1,...n$, where *n* is the number of trials in the Monte-Carlo or empirical studies. In addition to the two metrics, a ratio U of the value of each statistic (mean, standard deviation or percentiles) for the extended series to that for the observed series is computed. The ratio *U* is used to assess the performance of a record-extension technique in preserving the observed records characteristics. If the ratio *U* for a given statistic is larger than 1, then the applied technique overestimates this statistic. If it is less than 1, the technique underestimates the target statistic.

4 Results

The results are divided into two subsections: the first subsection presents the results of the Monte-Carlo experiment, and the second subsection presents the results of the empirical experiment.

4.1 Monte-Carlo experiment results

In the Monte-Carlo experiment, 5000 trials are generated. This number was chosen based on pre-analysis examination of the convergence of the error in estimating different statistics. Table 1 shows the *BIAS* values for the estimation of the mean and standard deviation from the extended data series. The hypothesis that the mean estimation *BIAS* value is equal to zero could not be rejected at the 0.05 significance level for any of the extension techniques under any of the twelve designed combinations. In addition, results show that there is no significant difference in the *BIAS* values for the estimation of the mean between the four record-extension techniques, under any of the twelve different combinations considered.

(Table 1)

Using the OLS or KTRL, the hypothesis that the *BIAS* value for the estimation of the standard deviation is equal to zero is rejected at the 0.05 significance level for all of the 12 combinations being considered. On the other hand, this hypothesis is rejected for 3 out of the 12 combinations considered when LOC is used and for 5 out of the 12 combinations when KTRL2 is used. The OLS and KTRL techniques result in a systematic underestimation of the standard deviation with *BIAS* values ranging between -0.366 and -0.020. The LOC *BIAS* values range between -0.001 and 0.012, while KTRL2 *BIAS* values range between 0.000 and 0.015. Table 1 shows also that the *BIAS* values decrease with an increase in the correlation coefficient or an increase in the number of concurrent records. These results may indicate that, given the availability of a sufficient number of concurrent records with a high level of association, both the LOC and KTRL2 techniques will estimate the standard deviation with high precision. In general, Table 1 shows that *BIAS* values corresponding to the LOC and KTRL2 are closer to zero than those corresponding to the OLS and KTRL under any of the twelve designed combinations.

Table 2 shows the *RMSE* values for the estimation of the mean and standard deviation using the four record-extension techniques for each of the twelve designed combinations of n_1, n_2 and ρ. The *RMSE* values for the estimation of the mean and standard deviation decrease with high correlation coefficients and large sizes of concurrent records. *RMSE* values for the estimation of the mean using the OLS are lower than those corresponding to other techniques, especially for small sizes of concurrent records. *RMSE* values for the estimation of the standard deviation using either the LOC or KTRL2 are equivalent and lower than those corresponding to OLS and KTRL.

(Table 2)

The values of the ratio *U* of the estimation of the standard deviation are presented in Table 3. The results show that the ratio *U*'s mean values that correspond to either the LOC or KTRL2 are closer to 1 than those that correspond to OLS or KTRL. When using the OLS or KTRL, the hypothesis that the ratio *U*'s mean value is equal to 1 is rejected at the 0.05 significance level for all of the 12 combinations considered. On the other hand, for the LOC or KTRL2, this hypothesis cannot be rejected for most of these combinations. Table 3 also shows that the ratio *U* becomes closer to 1 with high correlation coefficients and with large sizes of concurrent records. An increase in the number of concurrent records will ensure more precise estimates of the intercept and slope of the record-extension technique.

(Table 3)

Figure 2 shows the *BIAS* values for the estimation of the full range of non-exceedance percentiles (from the 5th to the 95th) using the four record-extension techniques. In Figure 2, six frames representing the two extreme cases (n_1 = 96 and n_1 = 24) under each of the three correlation coefficients considered are presented. Figure 2 shows that the *BIAS* values for the estimation of extreme percentiles decreases with high correlation coefficients and large sizes of concurrent records. When n_1 is equal to 24, the *BIAS* ranges between -0.6 and 0.6 when the correlation coefficient is equal to 0.5. The *BIAS* ranges between -0.4 and 0.4 when the correlation coefficient is 0.7 and between -0.17 and 0.17 when the correlation coefficient is 0.9.

When n_1 is equal to 24, the LOC and KTRL2 tend to underestimate low percentiles and overestimate high percentiles, while when n_1 equals to 96, this bias in estimating extreme percentiles is reduced. For instance, when the correlation coefficient is equal to 0.7 and n_1 is equal to 24, the *BIAS* ranges between -0.4 and 0.4, while it ranges between -0.1 and 0.1 when n_1 is equal to 96.

(Figure 2)

Similarly, Figure 3 shows the *RMSE* for the same six combinations. Figure 3 indicates that, the *RMSE* values for the estimation of high or low percentiles using any of the record-extension techniques decrease when the correlation coefficient or the size of concurrent period (n_1) increase. When n_1 is equal to 24, results show that the *RMSE* values for the estimation of high or low percentiles reach 0.4 for KTRL2 when the correlation coefficient is equal to 0.5. Meanwhile, when the correlation coefficient is equal to 0.7, the maximum *RMSE* is 0.35 and is only 0.25 when the correlation coefficient is equal to 0.9. For a correlation coefficient of 0.5, the *RMSE* values for the estimation of high or low percentiles reach 0.4 for KTRL2 when n_1 is equal to 24, while its maximum is only 0.17 when n_1 is equal to 96. Similar results are shown when the correlation coefficient is 0.7 and 0.9. *RMSE* values for the estimation of the percentiles using the LOC are lower than those corresponding to the KTRL2, for the case of small size of concurrent records.

(Figure 3)

In summary, the Monte-Carlo experiment results indicate that OLS and KTRL substantially reduce variability, while LOC and KTRL2 tend to preserve variability. The OLS and KTRL techniques overestimate low percentiles and underestimate high percentiles. The LOC and KTRL2 techniques tend to reduce the bias in the estimation of extreme percentiles. In addition, results show that error measures for the estimation of the standard deviation and extreme percentiles decrease with high correlation coefficients and large sizes of concurrent records.

The main advantage of LOC and KTRL2 over OLS and KTRL is that they allow a proper estimate of the probabilities of extreme events corresponding to the actual records. Thus, LOC and KTRL2 are both preferable methods when the probability distribution of the estimates, rather than just the mean or an individual estimate, is to be interpreted and used.

4.2 Empirical experiment results

A preliminary analysis of the Nile Delta data shows that, for most of the 94 locations, the EC, TDS and Cl records exhibit a positive skew. Consequently, in the empirical experiment, the record-extension techniques are applied to the logarithms of the three variables considered and the performance measures are applied to the reverse transformed records. The distribution of the correlation coefficient values of the TDS and EC, and the Cl and EC monthly log-transformed records are shown in Figure 4. For the TDS and EC, the correlation coefficients range between 0.52 and 0.98 with a median value of 0.93, while for Cl and EC, they range between 0.39 and 0.97 with a median value of 0.8. Extreme values detected in the EC and TDS time series occur at the same time, which leads to high correlation coefficients. On the other hand, the number of extreme values is larger in EC than in Cl and these extreme values do not occur simultaneously, which leads to relatively low correlation coefficients.

(Figure 4)

The first estimation model considered deals with the estimation of TDS using EC as an explanatory variable. Figures 5 and 6 summarize the results obtained for the ratio *U* for the mean and standard deviation (Figure 5) and for the full range of percentiles (Figure 6). The box plots in Figures 5a and 6a are plotted using the 940 cross-validation records applied at the 94 monitoring locations. Figures 5b and 6b are plotted using 11280 records and Figures 5c and 6c are plotted using 23688 records. The cross-validation method shows that the ratio *U* median values corresponding to the estimation of the mean values are equal to 1 for the four record extension techniques and for the three (n_1, n_2) combinations considered. The boxes corresponding to the four record extension techniques are symmetric around 1 and have almost the same dispersion.

Figure 5 shows also that the OLS and KTRL techniques lead to the underestimation of the standard deviations. Approximately 75% of the OLS and KTRL standard deviations are lower than those calculated from the observed values. The ratio *U* median values for standard deviations are 0.994, 0.972 and 0.948 for OLS when (n_1, n_2) equals (108, 12), (84, 36) and (60, 60) respectively. As for KTRL the median values are 0.996, 0.983 and 0.967 respectively. The LOC median values are closer to 1 than those corresponding to OLS or KTRL (1.000, 0.999 and 0.995 respectively), while for KTRL2 the ratio *U* median value is 1.001 for the three (n_1, n_2) combinations considered. The box plots corresponding to LOC and KTRL2 are more symmetric around 1 and show almost the same dispersion as shown by those corresponding to the OLS and KTRL. However, for both the mean and standard deviation, the box plots dispersion increases with the decrease in the size of concurrent records. This indicates that, with a large number of concurrent records, estimates of the intercept and slope of the record-extension techniques are more precise.

(Figure 5)

Figure 6 shows that the median values of the ratio U for low percentiles are higher than 1, while those corresponding to high percentiles are lower than 1 when using either the OLS or KTRL record-extension techniques. However, when using either the LOC or KTRL2, the median values of the ratio U are very close to 1. The box plots corresponding to the LOC and KTRL2 are symmetric around 1 and show almost the same dispersion as shown by those corresponding to the OLS and KTRL.

Under any of the three (n_1, n_2) combinations, the ratio U median values corresponding to LOC or KTRL2 are closer to 1 than those corresponding to OLS or KTRL respectively. Using any of the four record extension techniques, the median values become closer to 1 as n_1 increases. These results suggest that OLS and KTRL tend to overestimate low percentiles and underestimate high percentiles. LOC and KTRL2 reduce the bias exhibited by OLS and KTRL. Given that the correlation coefficient is almost the same in the three (n_1, n_2) combinations, results indicate that the bias in estimating extreme percentiles is reduced by increasing the number of concurrent records.

(Figure 6)

Similarly, for the second estimation model considered, to estimate Cl using EC as an explanatory variable, Figures 7 and 8 (a, b and c) summarize the results obtained for the ratio *U* for the mean and standard deviation (Figure 7) and for the full range of percentiles (Figure 8). Figures 7 and 8 show also that the box plots dispersion increases with the decrease in the number of concurrent records. Figure 7 confirms that LOC and KTRL2 are superior in preserving the standard deviation of the observed records. However, comparing the LOC and KTRL2, the dispersion in the box plot corresponding to KTRL2 is relatively smaller than that corresponding to the LOC. This indicates that KTRL2 is relatively better than LOC in preserving the standard deviation of the observed records.

For the estimation of percentiles (Figure 8), when using OLS or KTRL, the ratio *U* median values are higher than 1 for low percentiles (e.g. 5th, 10th, 15th percentiles) and lower than 1 for high percentiles (e.g. 85th, 90th, 95th percentiles). Results show that 75% of the OLS and KTRL low percentiles are higher than those calculated from the observed values. On the other hand about 75% of the OLS and KTRL high percentiles are lower than those calculated from observed values. However, when using either LOC or KTRL2, the median values are very close to 1 under any of the three (n_1, n_2) combinations considered. For any of the four record extension techniques, the ratio *U* median values become closer to 1 as n_1 increases. These results confirm that LOC and KTRL2 reduce the bias exhibited by OLS and KTRL in the estimation of extreme percentiles.

(Figures 7 and 8)

The comparison of Figure 5 with Figure 7 and Figure 6 with Figure 8, shows that the box plots presented in Figures 7 and 8 have a larger dispersion than those of Figures 5 and 6 respectively. The higher dispersion in Figures 7 and 8 is due to the relatively low correlation coefficients between Cl and EC in the second estimation model.

Figure 9 illustrates the *BIAS* and *RMSE* exhibited by the four extension techniques in estimating the TDS percentiles for the three (n_1, n_2) combinations considered. Results indicate that *BIAS* values for the estimation of the extreme percentiles using LOC and KTRL2 are closer to zero than those corresponding to OLS and KTRL, respectively. OLS and KTRL tend to overestimate low percentiles and underestimate high percentiles, which is the consequence of underestimating the standard deviation. Figure 9 shows also that, for the estimation of extreme percentiles, KTRL is much better than OLS as *BIAS* values corresponding to KTRL are closer to zero than those corresponding to OLS. However, the comparison of LOC and KTRL2 *BIAS* values indicates that there is no significant difference. Results of the three (n_1, n_2) combinations indicate that the *BIAS* values for the estimation of extreme percentiles is reduced significantly by increasing the number of concurrent records.

(Figure 9)

Figure 9 shows also the *RMSE* exhibited by the four record-extension techniques for the estimation of the TDS percentiles. Results show that *RMSE* values corresponding to the estimation of high percentiles are higher than those corresponding to low percentiles. Figure 9 shows that the *RMSE* values corresponding to the estimation of extreme percentiles are lower for KTRL or KTRL2 than

OLS or LOC respectively, while those corresponding to KTRL are slightly lower than those corresponding to KTRL2. In addition, the *RMSE* results confirm that the error on the estimation of extreme percentiles is reduced significantly by increasing the number of concurrent records.

Figure 10 illustrates the *BIAS* and *RMSE* values exhibited by the four extension techniques for the estimation of the Cl percentiles for the three (n_1, n_2) combinations considered. Figure 10 shows that, for the estimation of extreme percentiles, KTRL and KTRL2 are much better than OLS and LOC respectively. The *RMSE* results indicate that KTRL and KTRL2 lead to better results than OLS and LOC respectively for the estimation of extreme percentiles.

(Figure 10)

Figures 9 and 10 indicate that LOC and KTRL2 lead to better results than OLS and KTRL respectively, except for the slightly lower *RMSE* for the estimation of TDS percentiles when KTRL is used. The comparison of LOC and KTRL2 indicates that both techniques will lead to comparable results when a strong linear relationship exists between the dependent and independent variables. However, KTRL2 is more robust to the presence of extreme values.

The presence of extreme values affects the calculation of the statistical parameters upon which the OLS and LOC depend for the estimation of the intercept and slope (Equations 2 and 5). The mean is very sensitive to the presence of extreme values, while the median is not affected by the magnitude of a single extreme observation. In addition, the sample variance is strongly influenced by outlying values. If extreme values are present, the variance will indicate a much greater spread than is indicated by the majority of the data. This is the main reason why KTRL and KTRL2 show better results than OLS and LOC when using real water quality records.

Figures 6, 8, 9 and 10 clearly illustrate that both OLS and KTRL overestimate low concentrations and underestimate high concentrations, as one would expect from these methods tendency to produce an extended record with a lower standard deviation than that of the observed record. From the two empirical estimation models considered, results of the three (n_1, n_2) combinations indicate that the *BIAS* values for the estimation of extreme percentiles is reduced significantly by increasing the number of concurrent records.

Thus, in summary, results indicate that OLS and KTRL substantially reduce variability observed in the dependant variable, while LOC and KTRL2 tend to preserve variability. The OLS and KTRL techniques underestimate high concentration values and overestimate low values. On the other hand, the LOC and KTRL2 techniques tend to reduce the bias in the estimation of both high and low concentration values. Both techniques produce extended records that preserve extreme percentiles relatively well. Using real water quality data, KTRL2 shows a better performance than LOC in the estimation of extreme percentiles. The better performance shown by KTRL2 is due to the fact that real water quality data do not follow a normal distribution due to presence of extreme values. Even after transformation, some deviation from normality may exist. This slight deviation from normality and/or the presence of extreme values makes KTRL2 preferable. This relatively better performance shown by KTRL and KTRL2 over OLS and LOC, respectively, is mainly due to the robustness of the slope estimator towards the presence of extreme values and/or deviation from normality.

5 Conclusions

OLS, LOC, KTRL and the new technique KTRL2 were compared using a Monte-Carlo study of samples from bi-variate normal populations and an empirical analysis, which used water quality records from the Nile Delta water quality monitoring network. Different statistical performance measures were used to assess the performance of each of the extension techniques and their ability to maintain the statistical characteristics of the water quality records.

The Monte-Carlo study and the empirical analysis show that OLS and KTRL fall substantially short of achieving the desired result of creating a realistic extended record. They cannot be expected to provide records with the appropriate variability or the appropriate distribution shape.

The evaluation of the biases of moments and non-exceedance percentiles shows that LOC and KTRL2 are better than OLS and KTRL, respectively. OLS and KTRL estimates substantially underestimate the variance. Consequently, the frequency of extreme events, such as the exceedance of standards, would be underestimated by OLS or KTRL. On the other hand, by estimating the KTRL2 slope using percentiles of *y* and *x* rather than observed records, allow the estimates $\hat{y}_i$ from observed x_i to have distributional properties similar to those expected, had y_i been measured. However, when the size of the concurrent records is small, the LOC becomes slightly more precise than KTRL2 for the estimation of percentiles. The empirical experiment shows KTRL2 to have more desirable properties than LOC. The results show also that these differences are significant when extreme values are present.

When multiple estimates are to be generated and statements are to be made about probabilities of exceedance, such as probabilities of exceeding some water quality standard, inferences are made that depend on the probability distribution of the estimated data. In these cases, LOC and KTRL2, rather than OLS and KTRL, should be used to generate data. KTRL2 is superior when no transformation can produce near-normality due to heavy tails in the distribution and/or the presence of extreme values.

Acknowledgment: The authors wish to thank Prof. Shaden Abdel-Gawad, president of the National Water Research Center of Egypt, for providing the data used in this study. The financial support provided by Helwan University, Cairo, Egypt, and the Natural Sciences and Engineering Research Council of Canada (NSERC) is acknowledged.

6 References

[1]. Alley, W.M. and Burns, A.W. (1983). Mixed-station extension of monthly streamflow records. Journal of Hydraulic Engineering, 109 (10), 1272 - 1284.

[2]. Berryman, D., Bobée, B., Cluis D. and Haemmerli, J. (1988). Nonparametric Tests for Trend Detection in Water Quality Time Series. Water Resources Bulletin, 24(3), 545 - 556.

[3]. Draper, N.R. and Smith, H. (1966). Applied regression analysis, John Wiley, New York, p. 736.

[4]. DRI, Drainage Research Institute (2004). Drainage Water Status in the Nile Delta – Yearbook 2001/2002. Monitoring and analysis of Drainage water Quality project. Technical Report No. 72.

[5]. Halfon, E., (1985). Regression method in ecotoxicology: A better formulation using the geometric mean functional regression: Environ. Sci. and Technol. 19, 747-749.

[6]. Harmancioglu, N.B. and Yevjevich, V. (1986). Transfer of Information among Water Quality Variables of the Potomac River, Phase III: Transferable and Transferred Information. Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 1986, 81 p.

[7]. Harmancioglu, N.B. and Yevjevich, V. (1987). Transfer of hydrologic information among river points. Journal of Hydrology, 91, 103 - 118.

[8]. Helsel, D.R., and Hirsch, R.M., (2002). Statistical methods in water resources, Amsterdam, the Netherlands, Elsevier Science Publishers, 522 p.

[9]. Hirsch, R.M. (1982). A comparison of four streamflow record extension techniques. Water Resources Research, 18(4), 1081 - 1088.

[10]. Hirsch, R.M., Alexander R., and Smith R.A., (1991). Selection of methods for the detection and estimation of trends in water quality. Water Resources Research 27, 803-813.

[11]. Kermack, K. A. and J. B. S. Haldane, (1950). Organic correlation and allometry: Biometrika 37, 30-41.

[12]. Khalil, B., T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana (2010). A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks. Journal of Hydrology, 386, 173-185.

[13]. Kritskiy, S.N. and Menkel, J.F. (1968). Some statistical methods in the analysis of hydrologic data: Soviet Hydrology Selected Papers 1, 80-98.

[14]. Lettenmaier, D.P. (1988). Multivariate nonparametric tests for trend in water quality, AWRA, Water Resources Bulletin (24)3, 505 - 512.

[15]. Moog, D.B., Whiting P.J. and Thomas R.B. (1999). Streamflow record extension using power transformations and application to sediment transport, Water Resources Research, 35 (1), 243 - 254.Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V. (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[16]. Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V. (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[17]. Strobl, R.O. and Robillard, P.D. (2008). Network design for water quality monitoring of surface freshwaters: A review, Journal of Environmental Management, 87, 639 - 648.

[18]. Teissier, G., (1948). La relation d'allometrie sa signification statistique et biologique, Biometrics 4, 14-53.

[19]. Vogel, R.M. and Stedinger, J.R. (1985). Minimum variance streamflow record augmentation procedures. Water Resources Research, 21(5), 715 - 723.

[20]. Yevjevich, V. and Harmancioglu, N.B. (1985). Modeling Water Quality Variables of Potomac River at the Entrance to its Estuary, Phase II (Correlation of Water Quality Variables within the Framework of Structural Analysis). Report to D.C. Water Resources Research Center of the University of the District of Columbia, Washington, D.C., 59p.

[21]. Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis, 1, 2, and 3: Ned. Akad. Wentsch Proc., 53, 386-392, 521-525, and 1397-1412.

Table 1. *BIAS* values for estimating the mean and the standard deviation for the Monte-Carlo experiment

n_1	n_2	ρ	Mean				Standard deviation			
			OLS	**LOC**	**KTRL**	**KTRL2**	**OLS**	**LOC**	**KTRL**	**KTRL2**
96	**24**	**0.5**	0.001	0.002	0.001	0.001	-0.080*	0.000	-0.079*	0.001
78	**48**	**0.5**	0.001	0.002	0.000	0.001	-0.166*	0.001	-0.164*	0.002
48	**78**	**0.5**	0.002	0.002	-0.001	-0.001	-0.260*	0.004*	-0.257*	0.007*
24	**96**	**0.5**	0.000	0.002	-0.001	0.001	-0.366*	0.012*	-0.361*	0.015*
96	**24**	**0.7**	0.000	0.000	0.000	-0.001	-0.054*	-0.001	-0.053*	0.000
78	**48**	**0.7**	-0.001	-0.001	-0.001	-0.001	-0.110*	0.000	-0.108*	0.001
48	**78**	**0.7**	0.000	0.000	-0.001	-0.001	-0.170*	0.002	-0.168*	0.004*
24	**96**	**0.7**	-0.002	0.001	0.000	0.003	-0.232*	0.009*	-0.229*	0.012*
96	**24**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.020*	-0.001	-0.020*	0.000
78	**48**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.040*	0.000	-0.039*	0.001
48	**72**	**0.9**	0.000	0.000	0.000	0.000	-0.061*	0.000	-0.059*	0.001
24	**96**	**0.9**	-0.001	-0.001	-0.001	-0.001	-0.081*	0.003	-0.078*	0.005*

* The hypothesis that the *BIAS* is equal to zero is rejected at the 5% level

Table 2. *RMSE* values for the estimation of the mean and the standard deviation for the Monte-Carlo experiment

n_1	n_2	ρ	Mean				Standard deviation			
			OLS	**LOC**	**KTRL**	**KTRL2**	**OLS**	**LOC**	**KTRL**	**KTRL2**
96	**24**	**0.5**	0.099	0.101	0.101	0.104	0.106	0.077	0.106	0.077
78	**48**	**0.5**	0.111	0.117	0.117	0.127	0.184	0.093	0.183	0.093
48	**78**	**0.5**	0.134	0.144	0.150	0.166	0.278	0.119	0.276	0.120
24	**96**	**0.5**	0.184	0.204	0.224	0.257	0.396	0.183	0.394	0.183
96	**24**	**0.7**	0.096	0.097	0.097	0.098	0.088	0.071	0.088	0.072
78	**48**	**0.7**	0.104	0.106	0.111	0.115	0.136	0.083	0.135	0.083
48	**78**	**0.7**	0.121	0.125	0.140	0.150	0.196	0.102	0.196	0.103
24	**96**	**0.7**	0.160	0.171	0.208	0.228	0.273	0.153	0.275	0.153
96	**24**	**0.9**	0.095	0.095	0.096	0.096	0.072	0.068	0.072	0.068
78	**48**	**0.9**	0.098	0.098	0.104	0.105	0.084	0.073	0.084	0.074
48	**78**	**0.9**	0.106	0.107	0.124	0.126	0.104	0.083	0.104	0.083
24	**96**	**0.9**	0.123	0.125	0.168	0.172	0.136	0.108	0.139	0.108

Table 3. Ratio *U* values for the standard deviation for the Monte-Carlo experiment

n_1	n_2	ρ	OLS	LOC	KTRL	KTRL2
96	**24**	**0.5**	0.920*	1.000	0.921*	1.001
72	**48**	**0.5**	0.834*	1.001	0.836*	1.002
48	**72**	**0.5**	0.740*	1.004*	0.743*	1.007*
24	**96**	**0.5**	0.634*	1.012*	0.639*	1.015*
96	**24**	**0.7**	0.946*	0.999	0.947*	1.000
72	**48**	**0.7**	0.890*	1.000	0.892*	1.001
48	**72**	**0.7**	0.830*	1.002	0.832*	1.004*
24	**96**	**0.7**	0.768*	1.009*	0.771*	1.012*
96	**24**	**0.9**	0.980*	0.999	0.980*	1.000
72	**48**	**0.9**	0.960*	1.000	0.961*	1.001
48	**72**	**0.9**	0.939*	1.000	0.941*	1.001
24	**96**	**0.9**	0.919*	1.003	0.922*	1.005*

* The hypothesis that the ratio *U* is equal to 1 is rejected at the 5% level

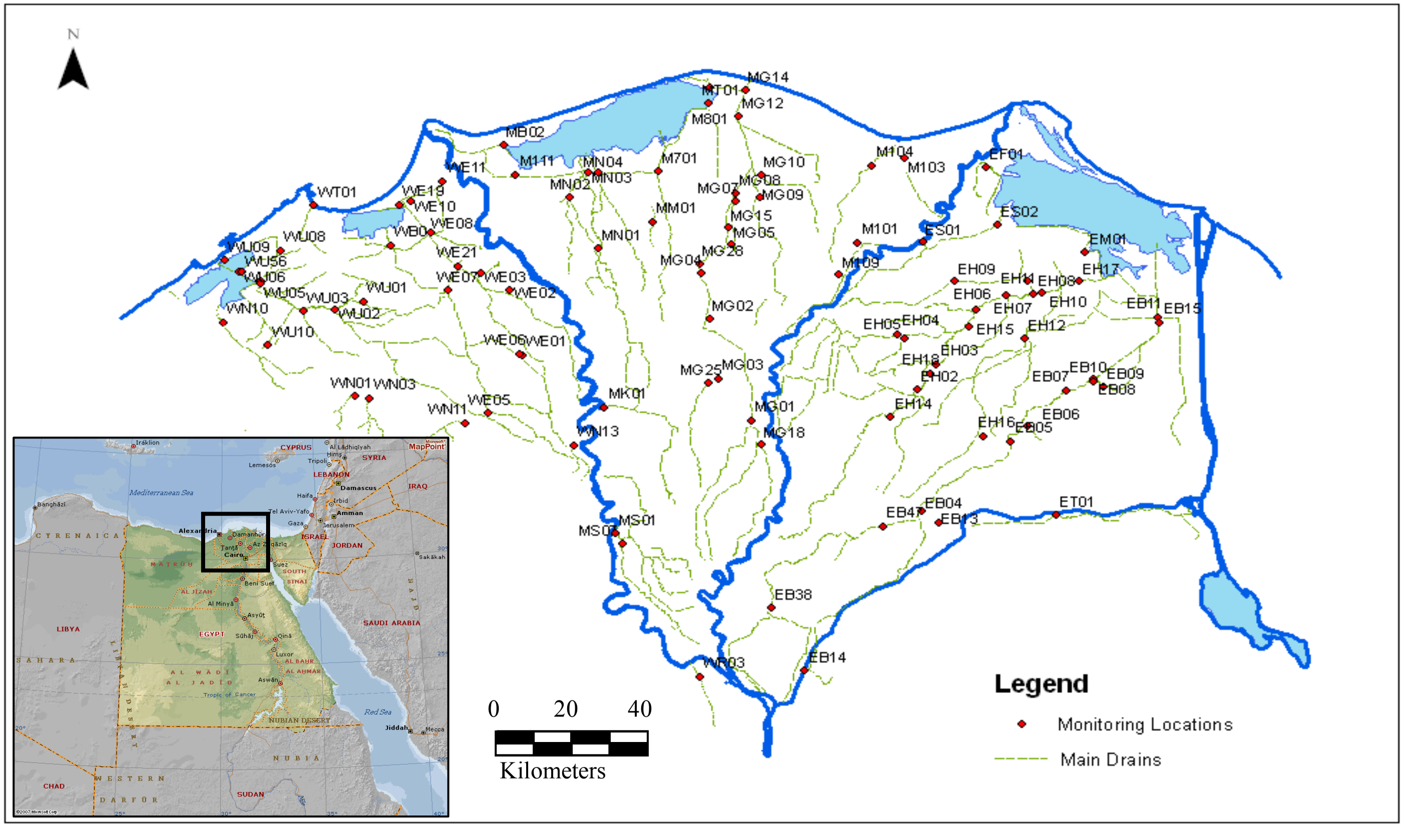


Figure 1. The Nile Delta surface WQM sites (source: DRI, 2004)

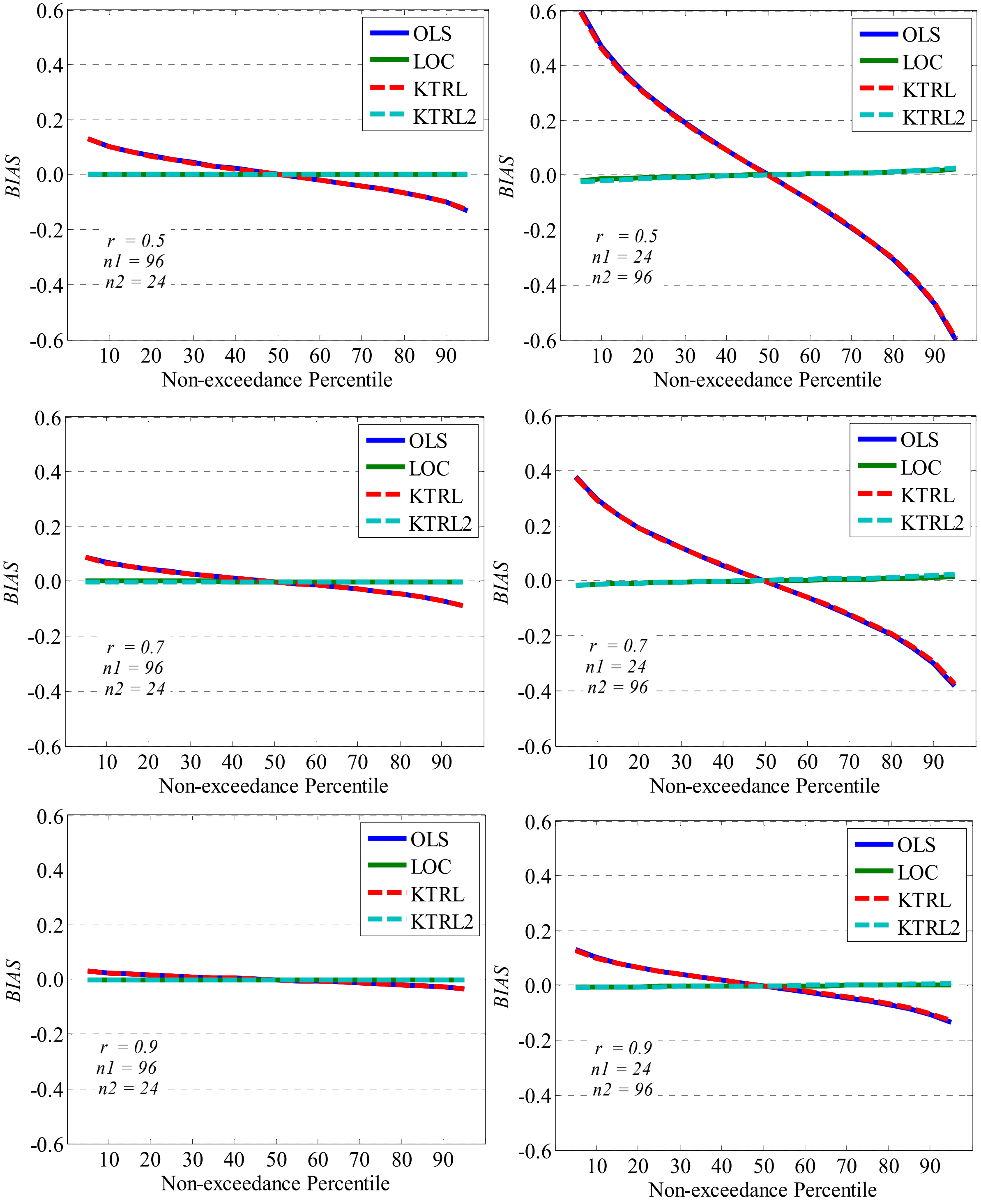


Figure 2. *BIAS* values for the estimation of the non-exceedance percentiles for the Monte-Carlo experiment

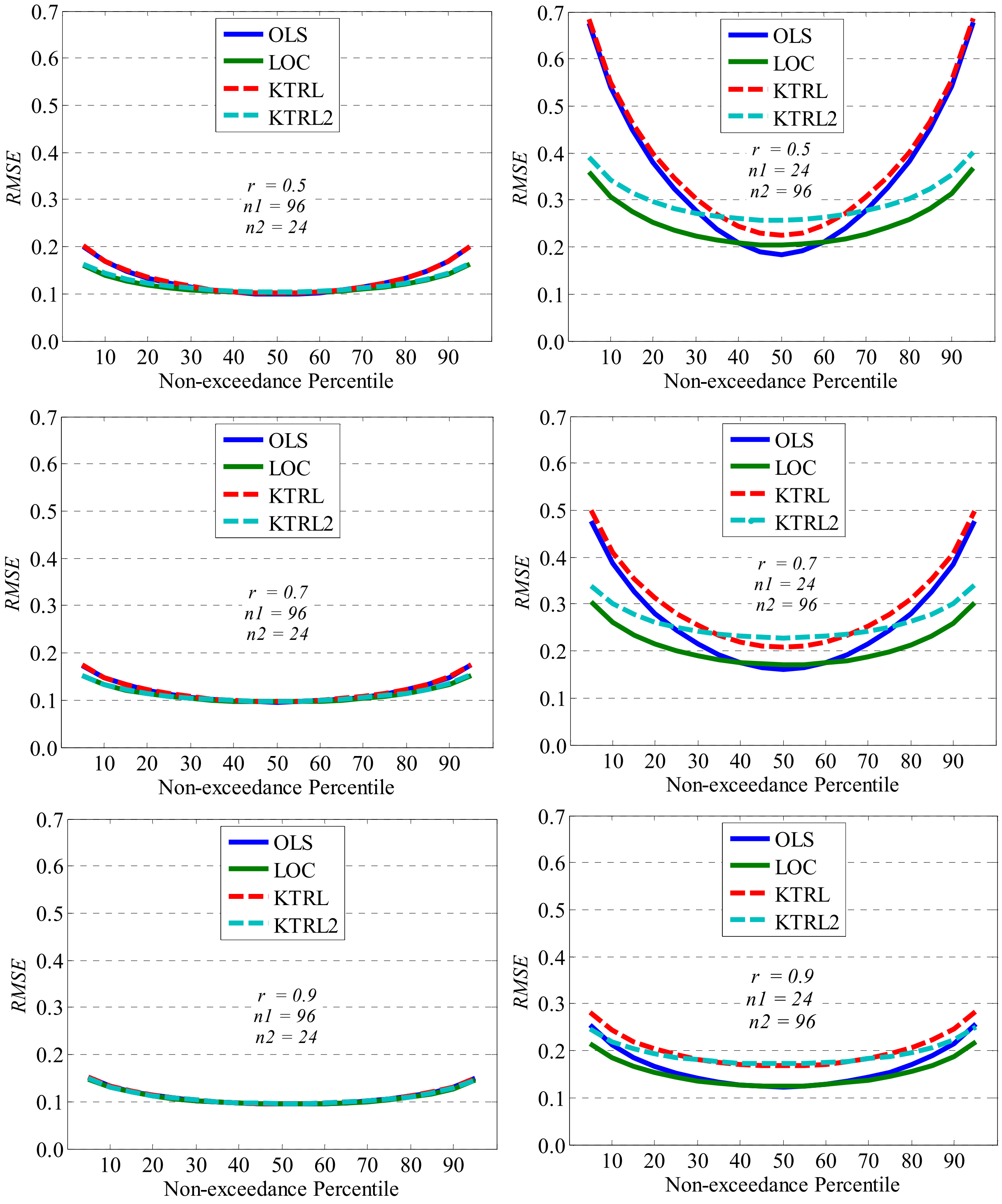


Figure 3. *RMSE* values for the estimation of the non-exceedance percentiles for the Monte-Carlo experiment

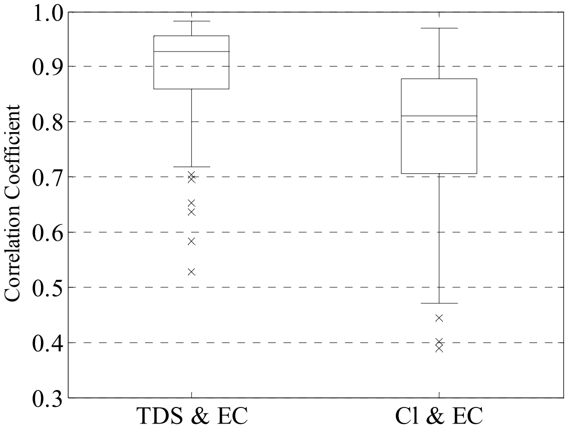


Figure 4. Box plot of correlation coefficients for the EC and TDS and Cl

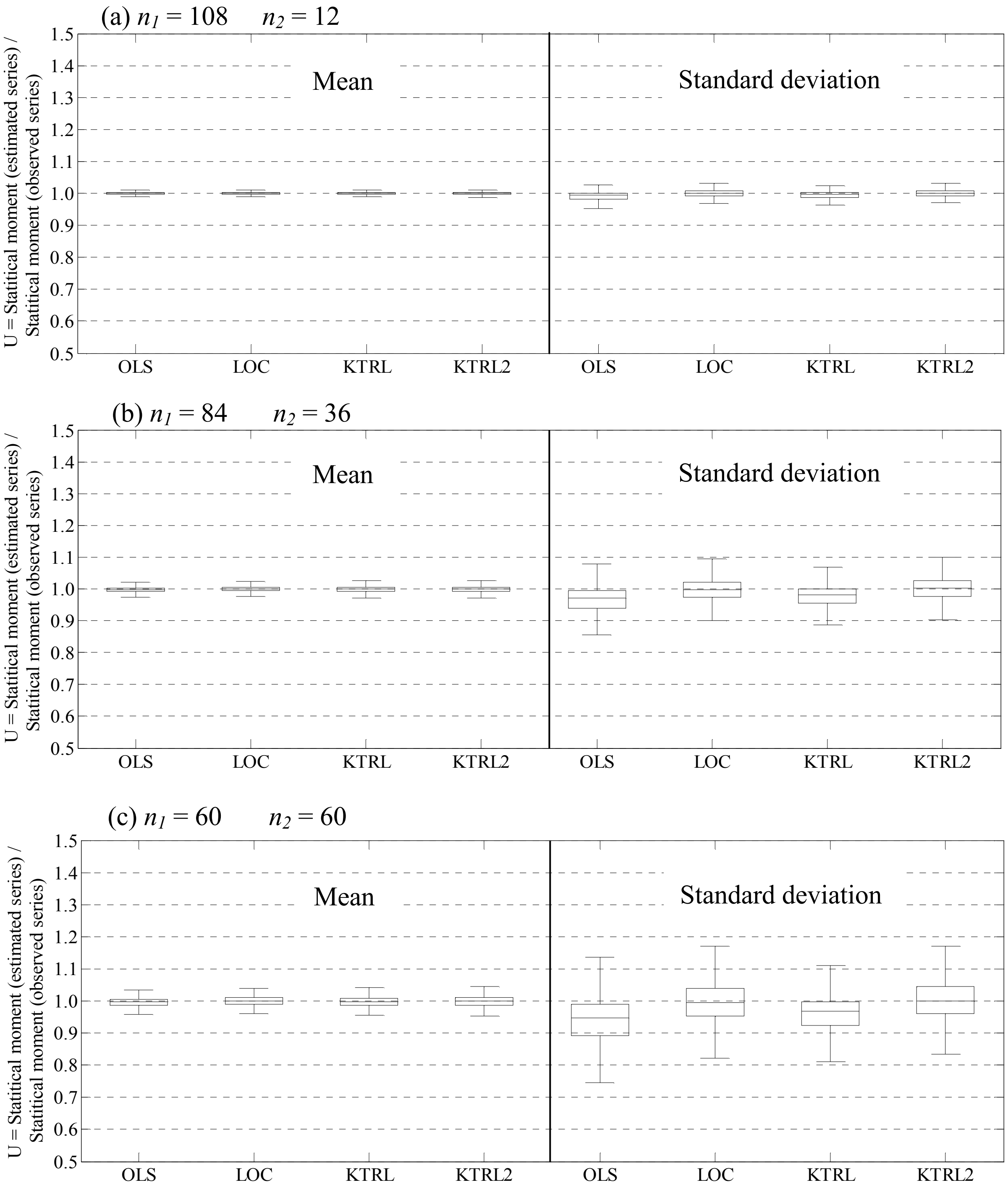


Figure 5. Box plots of the ratio *U* for the TDS mean and standard deviation (the y scale is the same as Figure 7)

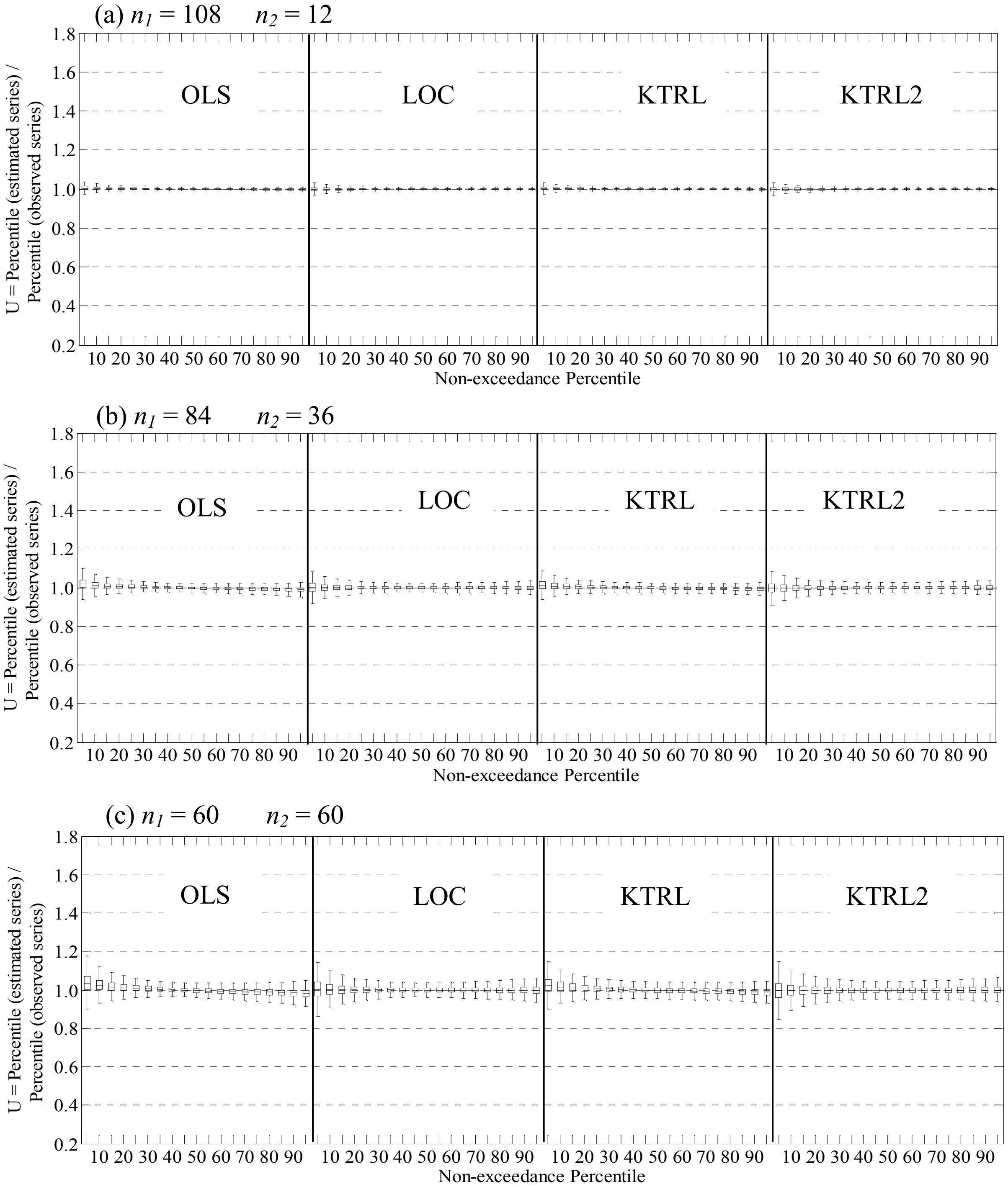


Figure 6. Box plots of the ratio *U* for different TDS percentiles (the y scale is the same as Figure 8)

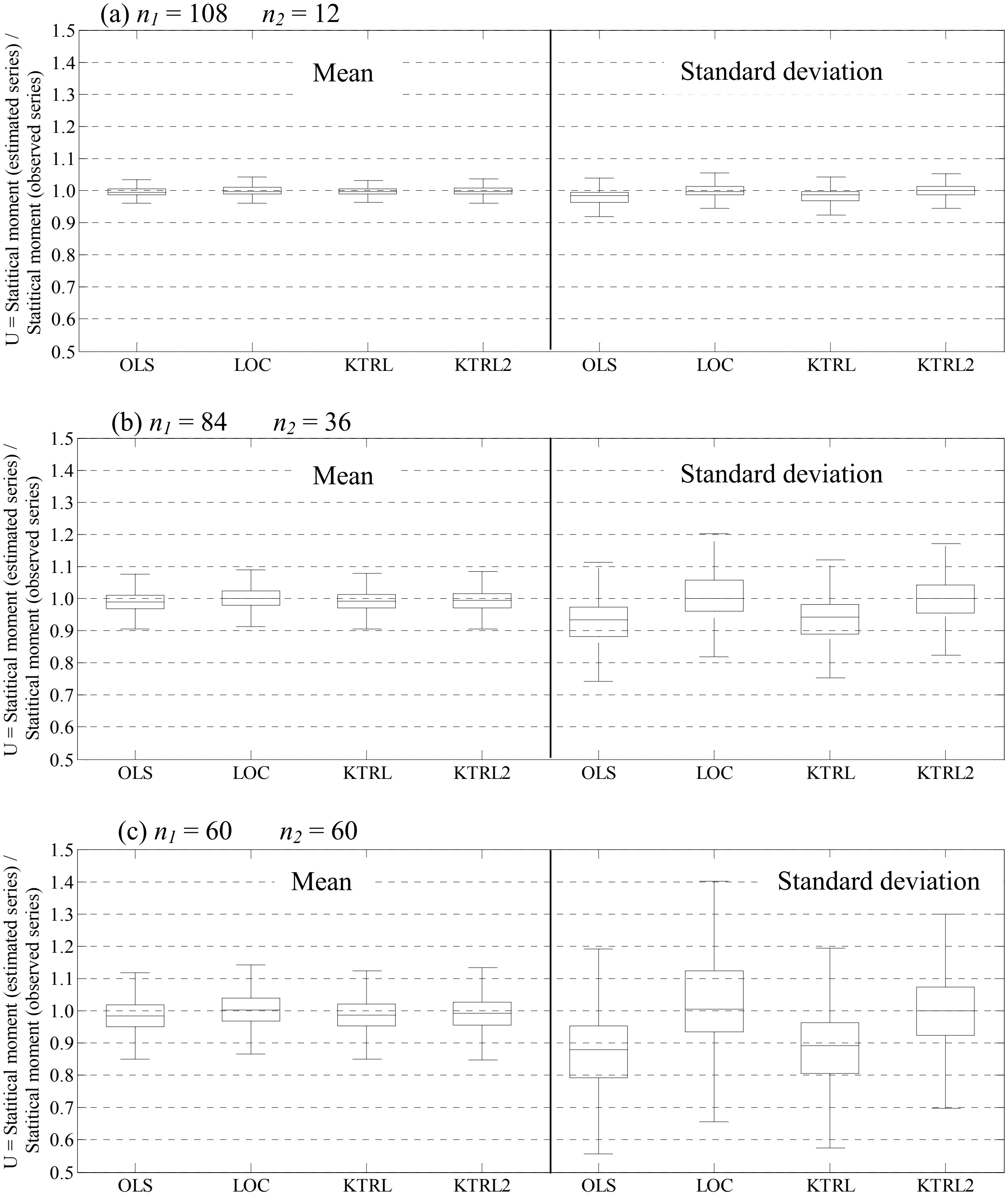


Figure 7. Box plots of the ratio *U* for the Cl mean and standard deviation

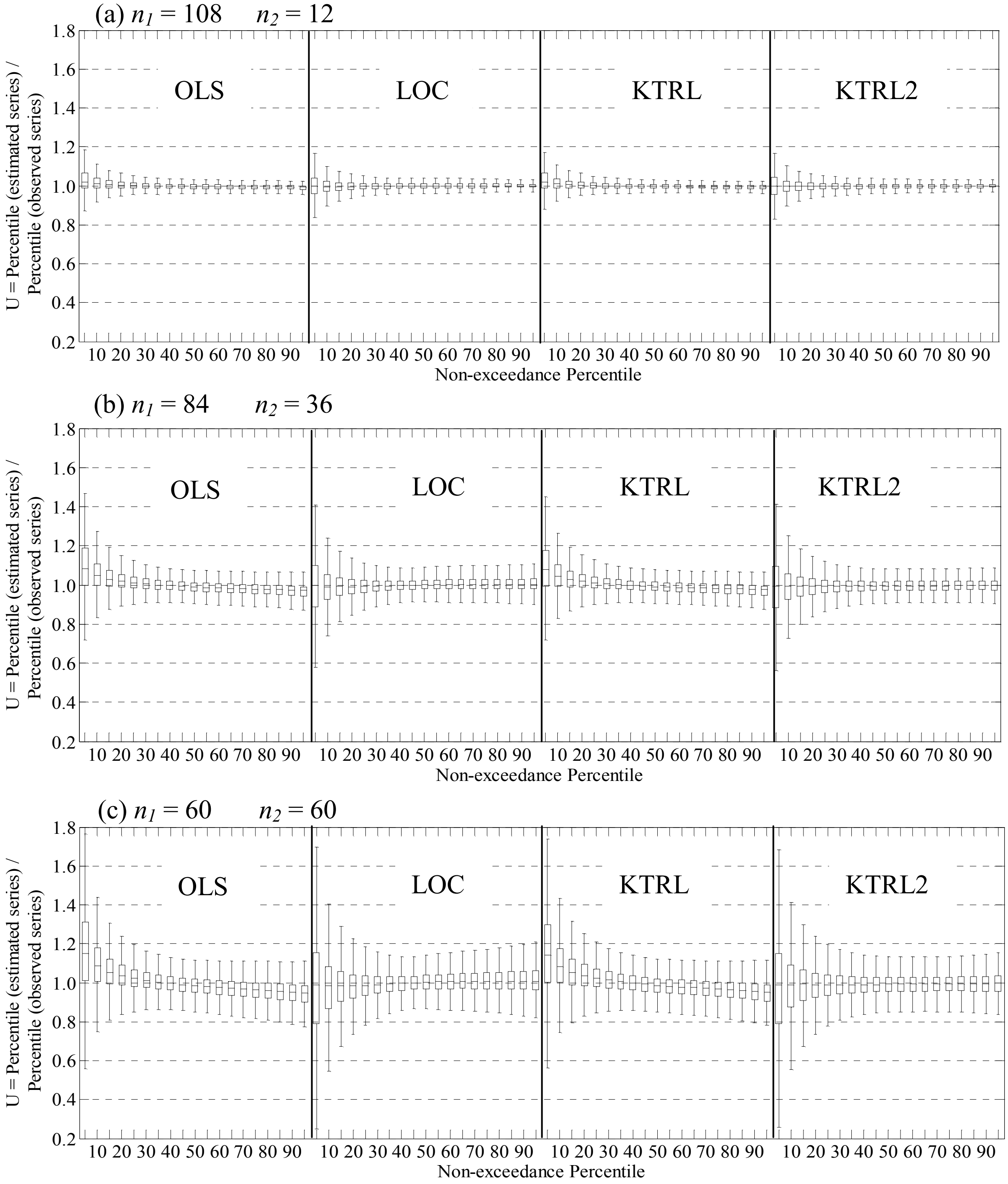
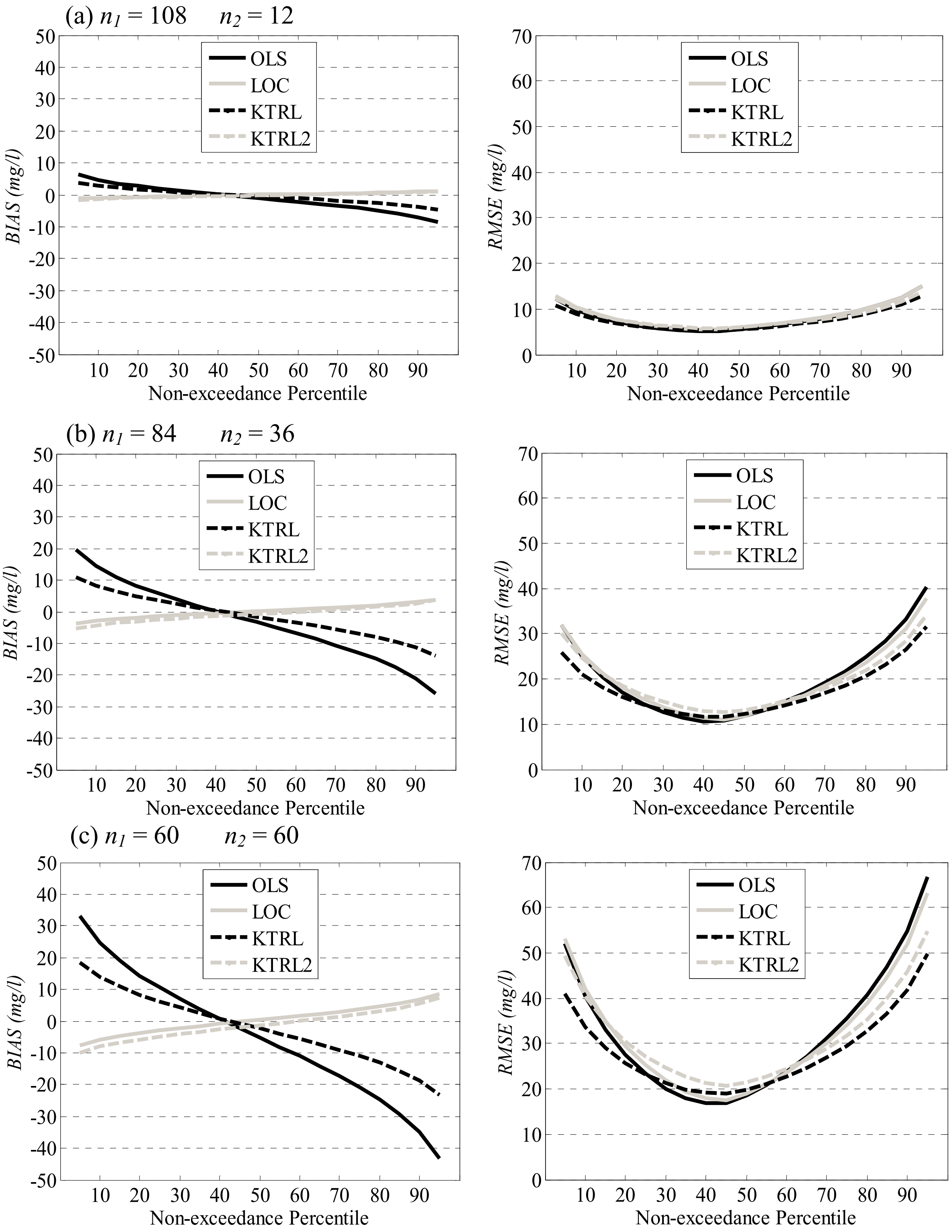


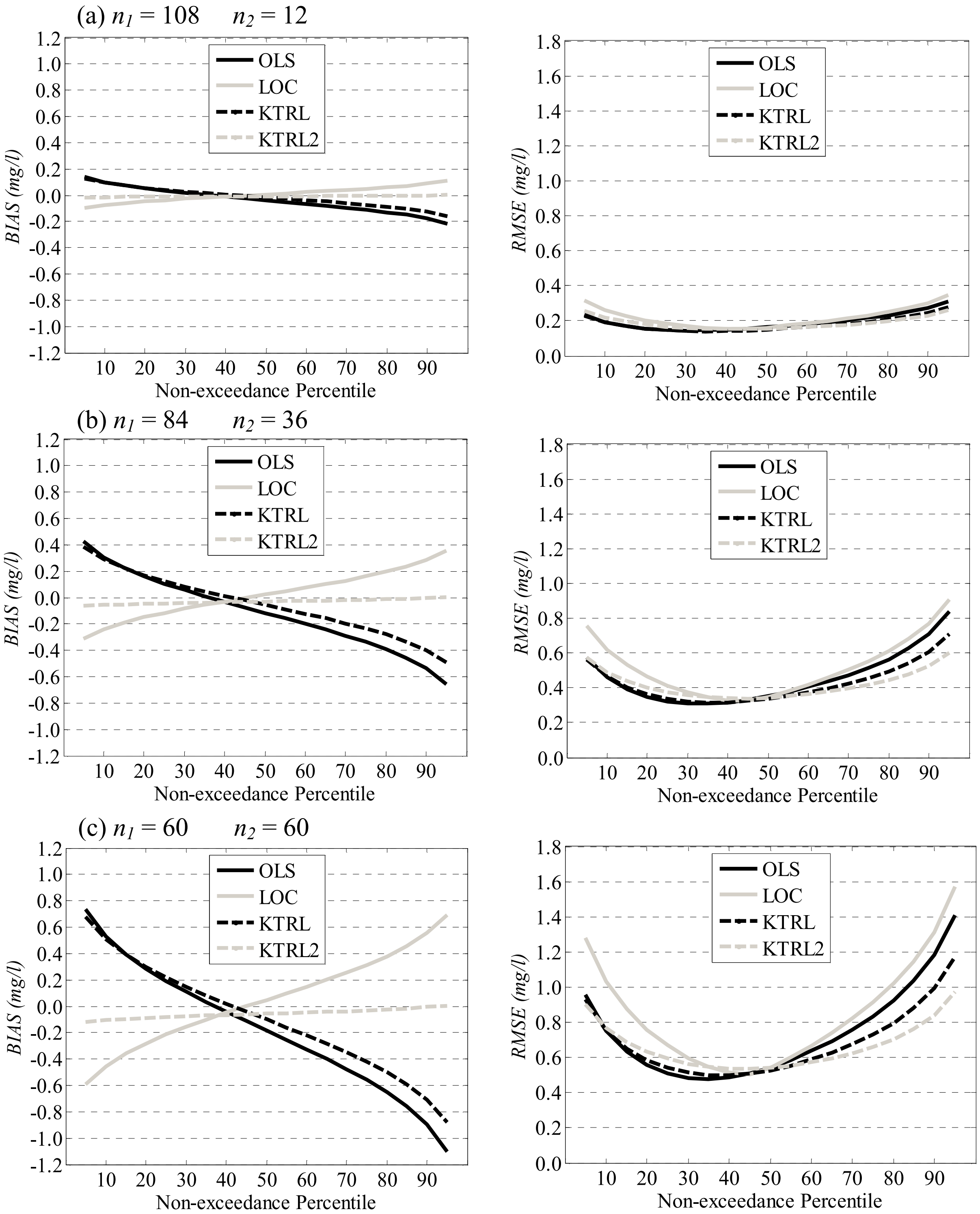
Figure 8. Box plots of the ratio *U* for different Cl percentiles


(a) $n_1 = 108$ $n_2 = 12$

(b) $n_1 = 84$ $n_2 = 36$

(c) $n_1 = 60$ $n_2 = 60$

Figure 9. *BIAS* and *RMSE* of the tested extension techniques for the estimation of TDS percentiles



(a) n_1 = 108 n_2 = 12

(b) n_1 = 84 n_2 = 36

(c) n_1 = 60 n_2 = 60

Figure 10. *BIAS* and *RMSE* of the tested extension techniques for the estimation of Cl percentiles

Article IV. A statistical approach for the assessment and redesign of the Nile Delta drainage water quality monitoring locations

A statistical approach for the assessment and redesign of the Nile Delta drainage water quality monitoring locations

B. Khalil[1,2], T. B.M.J. Ouarda[2] and A. St-Hilaire[2]

[1] Irrigation and Hydraulics department, Faculty of Engineering, Helwan University, Cairo, Egypt

[2]Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE, Québec City, Canada.

Submitted to Journal of Environmental Monitoring
November 2010.

Abstract

There are several deficiencies in the statistical approaches proposed in the literature for the assessment and redesign of surface water quality monitoring locations. These deficiencies vary from one approach to another, but generally include: (i) ignoring the attributes of the basin being monitored; (ii) handling multivariate water quality data sequentially rather than simultaneously; (iii) focusing mainly on locations to be discontinued; and (iv) ignoring the reconstitution of information at discontinued locations. In this paper, a methodology that overcomes these deficiencies is proposed. The basin being monitored is divided into sub-basins, and a hybrid-cluster analysis is employed to identify groups of sub-basins with similar attributes. A stratified optimum sampling strategy is then employed to identify the optimum number of monitoring locations at each of the sub-basin groups. An aggregate information index is employed to identify the optimal combination of locations to be discontinued. The proposed approach is applied for the assessment and redesign of the Nile Delta drainage water quality monitoring locations in Egypt. Results indicate that the proposed methodology allows the identification of (i) the optimal combination of locations to be discontinued, (ii) the locations to be continuously measured and (iii) the sub-basins where monitoring locations should be added. To reconstitute information about the water quality variables at discontinued locations, regression, artificial neural network (ANN) and maintenance of variance extension (MOVE) techniques are employed. The MOVE record extension technique is shown to result in a better performance than regression or ANN for the estimation of information about water quality variables at discontinued locations.

Keywords: water quality; monitoring locations; hybrid-cluster; stratified sampling; record extension; artificial neural networks.

1 Introduction and review

The selection of monitoring locations is an important aspect in the design of water quality monitoring networks. For instance, if the location is not representative of the water body, the data interpretation and presentation become inconsequential (Sanders et al., 1983; Ward et al., 1990). Early practices in water quality sampling focused on sites with easy access, without any systematic approach to the selection of monitoring locations (Harmancioglu et al., 1999). Over time, the number of sites has increased to include locations at points of interest such as those located upstream and downstream of highly industrialized or highly populated areas, areas with point-pollution sources or areas of intensive land use (Tirsch and Male, 1984). Later, various methodologies were proposed for the selection of both the number and location of monitoring locations.

The stream-order hierarchical approach is the most frequently used approach for designing a monitoring network when no water quality data are available. The stream-ordering procedure (Horton, 1945) assigns each un-branched small tributary the order of one, assigns a stream made up of only first-order tributaries the order of two, and so on. Sharp (1970; 1971) used stream-ordering procedures to measure the uncertainty involved in locating the source of pollutants observed at the outlet of a network. Sanders et al. (1983) followed Sharp's procedure (1970; 1971) by selecting monitoring locations on the basis of the number of contributing tributaries, pollutant discharges or loads. This approach systematically locates monitoring locations so as to divide the stream network into sections that are equal with respect to the number of contributing tributaries, the discharge or the pollutant loading. As discussed by Khalil and Ouarda (2009) assigning each un-branched small tributary the order of one assumes that each of these tributaries has the same contribution to the system. This assumption may only be valid if all tributaries drain the same area and have the same activities within these areas. In addition, in the presence of a point source of pollution (such as industrial effluent), assigning an order of one to the source assumes that it has the same weight as an un-branched tributary, yet it is completely different in the type and loading of pollution.

Tirsch and Male (1984) proposed an approach based on multivariate linear regression. Where a regression model is performed in which each monitoring location is considered the dependent variable and different combinations of the remaining locations are considered the independent variables. The adjusted coefficient of determination is obtained for each model, and the monitoring precision changes with the addition or deletion of some locations within the network. A high coefficient of determination indicates that a high degree of redundancy exists and that the location selected as the dependent variable might not be needed. Tirsch and Male (1984) applied this approach using daily specific conductance records from the network monitoring the Shoshone River basin in the USA. The main advantage of the regression approach is that it allows the reconstitution of information at discontinued locations. However, the main deficiency is that the approach is applied using only one water quality variable. In reality, the assessment and redesign of the water quality monitoring locations are more reliable when they are based on several water quality indicators. Multivariate data should be handled simultaneously rather than sequentially.

Harmancioglu and Alpaslan (1992) proposed the use of the entropy theory to decide upon the number and locations of monitoring sites. Using this method, the amount of transinformation (mutual information) between monitoring locations is determined based on the degree of uncertainty (Harmancioglu and Alpaslan, 1992). Dependence between monitoring locations results in reduced entropy, or uncertainty, between the locations. If the dependence is consistent over time, one or more of the monitoring locations may be discontinued with a minimal loss of information. This method was applied to select monitoring locations in the Gediz and Sakarya River basins in Turkey by Harmancioglu et al. (1994) and in the Mississippi River (Louisiana) in the USA by Ozkul (1996). Similar to the case for the regression approach, the main deficiency here is that the entropy approach is applied using only one water quality variable. A second deficiency is that the reconstitution of information at discontinued locations is not considered.

Different multivariate data analysis techniques have been employed for the redesign of water-quality-monitoring locations. These include principal component analysis (PCA), cluster analysis (CA) and discriminant analysis (DA). For example, Odom (2003) used PCA, CA and DA to assess and redesign the water-quality-monitoring network in the Great Smoky Mountains National Park, Tennessee, USA. Odom (2003) used the average pollutant values at each location in the PCA. Then, the first three principal components were used in a cluster analysis to define similar locations, and, finally, the DA was applied to verify the groups found in CA. Multivariate data analysis approaches overcome the deficiency in the regression and entropy approaches by employing several water quality variables simultaneously. However, the reconstitution of information at discontinued locations is not considered.

The common disadvantage of the proposed approaches is that they mainly focus on identifying monitoring locations to be discontinued. However, the optimum design may consist in the discontinuation of a number of existing monitoring locations while adding other locations at un-gauged sites. This downfall arises from the excessive focus on assessment using the water quality data already obtained, while ignoring the attributes of the basins being monitored (Khalil and Ouarda, 2009). Different basin attributes may affect the spatial allocation, such as the climatic region, land use, geology, the existence of point and non-point sources of pollution and the human activities within the basin being monitored (Khalil and Ouarda, 2009).

Thus, deficiencies in the current approaches can be summarized as follows: (i) ignoring attributes of the basin being monitored in the assessment procedure (regression, entropy, multivariate statistics); (ii) using only one water quality variable (stream order, regression, entropy); (iii) not considering the reconstitution of information at discontinued locations (stream order, entropy, multivariate statistics); and (iv) focusing only on the locations to be discontinued (regression, entropy, multivariate statistics).

The main goal of the present study is to develop a methodology for the assessment and redesign of monitoring locations that overcomes the deficiencies in the current applied approaches. In the following section, a description of the study area is provided. In section 3, the methodology is presented. In section 4, the obtained results are presented and discussed. Finally, the conclusions from this work are presented in section 5.

2 Nile Delta Water Quality Monitoring Network

Egypt is a semi-arid country with rainfall rarely exceeding 200 mm/year along its north coast. The rain intensity quickly decreases away from the coastal areas, and scattered showers can hardly be depended upon for agricultural production (Abu-Salama, 2007). According to an agreement between Egypt and Sudan (1959), the Nile water allocation is 18.5 billion m^3 to Sudan and 55.5 billion m^3 to Egypt (Dijkman, 1993). About 97% of Egypt's water resources are from the Nile River. The distribution of Egypt’s share of the Nile water to its population is near the water poverty threshold and will fall well below this threshold in the years to come (MWRI, 1997). The per capita share of fresh water resources in Egypt is about 800 m^3 per person per year (Abdel-Gawad et al., 2004; Frenken, 2005). It is expected to drop to 450 m^3 per person by the year 2025 (Abdel-Gawad et al., 2004).

One of the applied solutions to stretching the limited Egyptian water resources is the reuse of agricultural drainage water in irrigation after mixing it with fresh irrigation water. The drainage system in the Nile Delta is composed of 22 catchment areas. Depending on their quality, effluents are either discharged into the northern lakes or pumped into irrigation canals at 21 sites along the main drains to augment the freshwater supply (DRI-MADWQ, 1998). In 1997, the National Water Research Center (NWRC) started a national water-quality-monitoring program. The monitoring program of the Nile Delta drainage system aims to (i) assess its compliance with the national standards, (ii) estimate mass transport, and (iii) identify temporal and spatial trends (NAWQAM, 2001). The Nile Delta drainage system monitoring network (Figure 1) consists of 94 monitoring locations, through which 33 water quality variables are measured on a monthly basis.

(Figure 1)

Twenty-one of the 94 monitoring locations are located at drainage water reuse locations (mixing points), which are used to identify the optimal mixing rate of the drainage water with fresh irrigation water. Ten monitoring locations are located in the drainage system main streams. These locations are used as checkpoints for the assessment of the water and salt balance. Thirteen monitoring locations at the drainage system outfalls to the northern lakes and the Mediterranean are used to assess the amount of pollution discharged. Fifty monitoring locations are located at tributaries that serve sub-catchments and deliver water to the main drainage systems. These 50 locations are the focus of the present study.

3 Methodology

The methodology proposed in the present study is developed to overcome the four main deficiencies in the currently applied statistical approaches. In order to incorporate the attributes of the monitored region into the assessment and redesign of monitoring locations, the Nile Delta is divided into spatial units (SUs). A spatial unit (SU) is an area that is drained by only one point in the drainage system. Measuring the water quality at this point describes the effect of the SU attributes on the water quality conditions. Thus, for each SU, attributes that explain different natural and anthropogenic effects are identified. It is assumed that the SU attributes provide information concerning the main source of pollution in the drainage systems. The measured water quality variables that better explain the variability in water quality in the Nile Delta drainage system are selected based on PCA. These data preparation steps are described in subsection 3.1.

To reconstitute information about the variables measured at discontinued monitoring locations, record extension techniques are employed. A short review of record extension techniques is provided in subsection 3.2. The proposed methodology for spatially distributing water-quality-monitoring locations is described in subsection 3.3.

3.1 Data preparation

Data preparation consists of three steps, dividing the Nile Delta into SUs, identifying SU attributes and selecting the water quality variables that best describe the variation in the drainage system water quality. The Nile Delta is divided into 94 SUs (Figure 2), from which 50 are gauged. The SU attributes are selected to describe different natural and anthropogenic effects within each SU. However the selection is restricted by data availability.

(Figure 2)

Water quality variables that better explain the variability in water quality are selected using PCA. PCA is one of a number of factor extraction methods. PCA transforms a set of correlated variables into a smaller set of uncorrelated variates called principal components (Jobson, 1992). The first component explains most of the variance in the data, and each successive component explains less of the variance (Tabachnick and Fidell, 1996). Correlations between the variables and principal components are called component loadings. The component loading matrix obtained from PCA reflects the characteristics of the extraction procedure, which maximizes the variance in each successive component. Once the loading matrix is extracted, rotation can take place. Rotation is ordinarily used after extraction to maximize high correlations and minimize low ones, which facilitates interpretation of the components. Numerous methods of rotation are available. The one applied in this study is varimax (variance maximizing procedure). The interpretation of the components is in terms of the variables related to these components and their significance to the physical processes. The water quality variable most related to each component (the variable with the highest absolute factor loading) is then selected as an indicator of the water quality variables (one indicator from each component).

3.2 Information reconstitution at discontinued locations

Assume that the variable y measured at a discontinued monitoring location Y has n_1 years of data and that the variable x measured at a continuously monitored location X has $n_1 + n_2$ years of which n_1 are concomitant with the data observed at Y, illustrated as follows:

$$
\begin{aligned}
&Location\ (X): x_1, x_2, x_3, \ldots\ldots, x_{n_1}, x_{n_1+1}, x_{n_1+2}, \ldots\ldots., x_{n_1+n_2} \\
&Location\ (Y):\ y_1, y_2, y_3, \ldots.., y_{n_1}
\end{aligned}
$$

Consider that year n_1 is the year when the assessment and redesign took place. After n_1 years, a decision is made to stop monitoring at location Y and continue measuring at location X. Assume that after n_2 years, our interest is to reconstitute information about water quality variables at discontinued locations. Matalas and Jacobs (1964) developed a procedure for obtaining unbiased estimators of the mean (μ_y) and the variance (σ_y^2), showing that the mean value ($\hat{\mu}_y$) of the extended *Y* series can be determined using Equation 1:

$$\hat{\mu}_y = \bar{y}_1 + \frac{n_2}{n_1 + n_2} \hat{\beta}(\bar{x}_2 - \bar{x}_1) \tag{1}$$

where $\bar{y}_1$ and $\bar{x}_1$ are the mean values of y_i and x_i, respectively, based on the short records $i = 1,\ldots,n_1$, $\bar{x}_2$ is the mean value of x_i observed during the period $i = n_1 + 1,\ldots.,n_2$ and the parameter $\hat{\beta}$ is the estimated regression coefficient. Based on this formulation, it is possible to show (Cochran, 1953) that the variance of $\hat{\mu}_y$ is given by Equation 2:

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

where σ_y^2 is the population variance of y and ρ is the population correlation coefficient between x and y. For practical use, these values may be replaced by their estimates based on the n_1 years of data (Ouarda, et al., 1996). For the variance estimator ($\hat{\sigma}_y^2$), Matalas and Jacobs (1964) obtained the following expression:

$$\hat{\sigma}_y^2 = \hat{\beta}^2 s_x^2 + \left[1 - \frac{n_1 + n_2 - 3}{(n_1 - 3)(n_1 + n_2 - 1)}\right]\frac{n_1 - 1}{n_1 - 2}(s_{y1}^2 - \hat{\beta}\, s_{x1}^2) \tag{3}$$

where s_x^2 is the variance estimate based on the entire x-series and s_{y1}, s_{x1} are the standard deviations of y and x based on the short records $i = 1,\ldots,n_1$. Moreover, Matalas and Jacobs (1964) showed that the variance of the variance estimator ($Var\{\hat{\sigma}_y^2\}$) is given by Equation 4:

$$Var\{\hat{\sigma}_y^2\} = \frac{2\sigma_y^4}{n_1 - 1} + \frac{n_2 \sigma_y^4}{(n_1 + n_2 - 1)^2 (n_1 - 3)} (A\rho^2 + B\rho + C) \tag{4}$$

where A,B and C are constants that depend on n_1 and n_2 (see e.g. Vogel and Stedinger, 1985). Thus, by using the formulas provided by Matalas and Jacobs (1964), one can estimate the mean and the variance of the discontinued variables during the period $i = n_1 + 1,....,n_2$. However, to extend the records of the selected water quality variables at a discontinued location, a record extension technique should be applied.

3.2.1 Record extension techniques

Linear regression is one of the commonly used record extension techniques. To extend records of the discontinued variable y for the period $n_1 + 1$ through n_2 years, one can use simple linear regression of y on x, which leads to the following formula:

$$\hat{y}_i = a + bx_i \tag{5}$$

where $\hat{y}_i$ are the estimated values of y for $i = n_1 + 1,...n_2$ and a and b are the constant and slope of the regression equation, respectively. The parameters a and b are the values that minimize the sum of the squared difference between the estimated and measured y values. The solutions of a and b are found by solving the normal equations (Draper and Smith, 1966, p. 59). The optimal solution to Equation 5 is:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \tag{6}$$

where r is correlation coefficient between x and y computed from records of the concurrent period. The use of regression analysis often results in underestimation of the variance in the extended records (Alley and Burns, 1983). Hirsch (1982) suggested two other methods referred to as MOVE1 and MOVE2 (Maintenance of Variance Extension, Types 1 and 2). In MOVE1, Hirsch (1982) chose the estimators of a and b so that if Equation 5 is used to generate an entire sequence $\hat{y}_i$, for $i = 1,\ldots,n_1 + n_2$, the short sample moments $\bar{y}_1$ and s_{y1}^2 would be estimated. In MOVE2, Hirsch (1982) chose a and b so that if Equation 5 is used to generate an entire sequence $\hat{y}_i$ for $i = 1,\ldots,n_1 + n_2$, the unbiased estimates $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ would be estimated. Hirsch (1982) evaluated the MOVE1, MOVE2 and regression methods using both a Monte-Carlo study and empirical analysis, both of which showed that regression cannot be expected to provide records with the appropriate variability.

In practice, one uses Equation 5 to generate the $\hat{y}_i$ only for $i = n_1 + 1,\ldots,n_1 + n_2$. This suggests that Hirsch used estimators of a and b that did not achieve what he intended (Vogel and Stedinger, 1985). MOVE3 was then proposed by Vogel and Stedinger (1985). In MOVE3, the main goal is to select a and b in Equation 5 so that the resultant sequence of $n_1 + n_2$ values $\{y_1,\ldots,y_{n1},\hat{y}_{n1+1},\ldots,\hat{y}_{n1+n2}\}$ has mean $\hat{\mu}_y$ and variance $\hat{\sigma}_y^2$ (the Matalas and Jacobs estimators (Equations 1 and 4). Estimates of a and b for the MOVE3 method are obtained by rewriting Equation 5 as:

$$\hat{y}_i = a + b(x_i - \bar{x}_2) \tag{7}$$

where estimates of a and b are obtained from Equations 8 and 9:

$$a = [(n_1 + n_2)\hat{\mu}_y - n_1\bar{y}_1]/n_2 \tag{8}$$

$$b^2 = \left[(n_1 + n_2 - 1)\hat{\sigma}_y^2 - (n_1 - 1)s_{y1}^2 - n_1(\bar{y}_1 - \hat{\mu}_y)^2 - n_2(a - \hat{\mu}_y)^2\right]\left[(n_2 - 1)s_{x2}^2\right]^{-1} \tag{9}$$

Vogel and Stedinger (1985) carried out a Monte-Carlo experiment that indicated the MOVE2 and MOVE3 techniques to be nearly indistinguishable with respect to the mean square error of the estimators of the mean and variance of the complete extended record. Khalil et al. (2010) compared the regression and MOVE3 techniques for the reconstitution of information about discontinued water quality variables in the Nile Delta water quality monitoring network in Egypt. They recommended the use of MOVE3 in the case of reconstituting information about discontinued variables, while the regression technique is recommended for the reconstitution of missing values. Khalil et al. (2010) compared record extension techniques for information transfer between water quality variables measured at the same monitoring locations, while in this study information transfer among locations is considered.

An artificial neural network (ANN) is an information processing system that is designed to mimic certain structures and functions of the biological neural networks of the human brain. Given sufficient parameters, an ANN can be used to create nonlinear mathematical models for general approximation. Among the various types of ANNs, multilayer perceptrons (MLPs), originally proposed by Rumelhart and McClelland (1986), are the most commonly used and well-researched class of ANNs (Ouarda and Shu, 2009). This type of ANN implements a feed-forward supervised paradigm. A MLP consists of an input layer, one or more hidden layers and an output layer. The input layer receives the values of the input variables. The output layer provides the ANN prediction and represents the model output. The layers lying between the input and output layer are called hidden layers. Nodes in each layer are interconnected through weighted acyclic arcs from each preceding layer to the following without lateral or feedback connections.

In this study, a MLP having one input layer, one hidden layer and one output layer is used. Inputs are the selected water quality variables at the continuously monitored location(s), and the outputs are the same types of variables at the discontinued location. The tan-sigmoid transfer function is used for nodes in the hidden layer. The use of a nonlinear transfer function extends the nonlinear approximation ability of the ANN (Shu and Ouarda, 2007). A linear transfer function is used for the output nodes. A linear transfer function for the output nodes has the advantage of potentially unbounded outputs (Shu and Burn, 2004).

Determining the number of neurons in the hidden layer is an important task when designing an ANN (Shu and Ouarda, 2007). Too many hidden nodes may lead to the problem of overfitting. Too few nodes in the hidden layer may cause the problem of underfitting. As a rule of thumb, the number of nodes in the hidden layer should be less than twice the input layer size (Shu and Ouarda, 2007). In this study, a sensitivity analysis is performed to identify the optimal number of hidden nodes. By varying the number of hidden neurons from three to eight, ANNs with seven hidden neurons are identified as providing the most accurate estimation. Thus, seven hidden neurons are ultimately used in the hidden layer. In this study, the linear regression, MOVE3 and ANN record extension techniques are employed to reconstitute information about variables at the discontinued location(s).

Hirsch (1982), Vogel and Stedinger (1985) and Moog and Whiting (1999) applied the regression and MOVE techniques to the logarithm of the streamflow records rather than the raw data. This transformation tends to improve the normality of discharge histograms, which generally exhibit a strongly positive skew (Hirsch, 1982). Similarly, water quality variables generally exhibit a positive skew (Lettenmaier, 1988; Berryman et al., 1988), which is also confirmed in our case by a preliminary analysis of the Nile Delta data. Consequently, in this study, record extension techniques are applied to the logarithms of the water quality variables.

3.2.2 Identification of the best auxiliary location(s)

For each of the selected water quality variables at a discontinued location, a best auxiliary variable to be used as a predictor in the record extension techniques is selected from other continuously monitored locations using Equation 2. The continuously measured variable that minimizes the variance of the estimated mean of the discontinued variable after n_2 years is selected as the predictor. Thus, for a discontinued location, the best auxiliary variables could be selected from several continuously measured locations. Consequently, for each discontinued location, one or more continuously monitored locations could be identified as the auxiliary location(s). By using Equation 2, the choice of the best auxiliary variable is based on the number of concurrent years of measurements (n_1), the correlation coefficient and also the number of years after the assessment and reselection took place (n_2).

One can assess the precision of the variance of the mean value estimator (Equation 2) after a certain number of years, assuming that σ_y^2 and the correlation coefficient (ρ) remain unchanged and equal their estimates based on n_1 years of data (Ouarda et al., 1996; Khalil et al., 2010). In this study, n_2 is assumed to be two years, thus, one would like to reconstitute the information about the discontinued variables after two years from when assessment and reselection took place.

3.2.3 Empirical experiment

In the case of reconstituting information about water quality variables at a discontinued location, the objective is to estimate the monthly records of these variables for the n_2 years $\left\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\right\}$, while maintaining the main statistical characteristics of the historical records. In the assessment of water quality, one may be interested not only in the statistical moments but also in the extreme values. If the technique used for record extension introduces a bias in the value of the more extreme-order statistics, this will lead to bias in the estimates of the probability of exceeding selected extreme values or, conversely, bias in the estimation of distribution percentiles (Hirsch, 1982).

An empirical experiment is designed to examine the utility of the simple linear regression, MOVE3 and ANN techniques for preserving the statistical characteristics of the water quality variables measured at discontinued locations. In order to evaluate the performance of the three record extension techniques, a cross-validation (jackknife) is conducted. In the cross-validation, two years of monthly records are in turn removed from the available ten years of data. All possible combinations of successive or non-successive two-year periods are considered. Thus, from the available ten years of monthly records C(10,2) = 45 possible combinations are considered. The values for these two years of monthly observations are then estimated using the three record extension techniques calibrated with the remaining eight years.

The experimental design is as follows. Each of the 50 monitoring locations is assumed to be discontinued, and, for each of the selected variables at a discontinued location, the best auxiliary variable is identified from other continuously monitored locations using Equation 2. For each pair of water quality variables identified as the variables to be discontinued and their best auxiliary, the three record extension techniques are applied. Thus, different realizations of the extended water quality variable records (i.e. the number of selected variables at each location × 50 considered locations × 45 different combinations) are generated for cross-validation. Important characteristics of the observed and generated records during the extension period are computed for the three record extension techniques. Evaluation of the records generated by the extension techniques involves determining the ability of the techniques to estimate records with the various statistical properties of the observed records.

The extended records $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$ are compared to the observed records $\{y_{n_1+1},....,y_{n_2}\}$ based on the estimation of the mean, standard deviation and over the full range of percentiles (from the 5th to the 95th percentile). Different performance measures are applied. First, the ratio U of each statistic estimated from the extended records over that estimated from the observed records is computed. The ratio U is used to assess the performance of the record extension techniques in preserving the variable characteristics. If the ratio U for a given statistical parameter is larger than 1, this means that the applied record extension technique overestimates this statistical parameter. If it is less than 1, the record extension technique underestimates the statistical parameter. Concurrently, two performance measures are used to assess the three record extension techniques. They are the relative bias (*BIASr*) and the relative root mean square error (*RMSEr*), which are defined as follows:

$$BIASr = \frac{1}{nt}\sum_{i=1}^{nt}\frac{\hat{st}_i - st_i}{st_i} \tag{10}$$

$$RMSEr = \sqrt{\frac{1}{nt}\sum_{i=1}^{nt}\left[\frac{\hat{st}_i - st_i}{st_i}\right]^2} \tag{11}$$

where $\hat{st}_i$ and st_i are the estimated statistical parameters from the extended records $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$ and the observed records $\{y_{n_1+1},....,y_{n_2}\}$, respectively, and nt stands for the number of trials generated in the empirical experiment. The three performance measures are applied to the reverse-transformed records (original records).

3.3 Assessment and redesign of monitoring locations

An approach to spatially distributing monitoring locations, when the objective is either to establish a new water-quality-monitoring network or to assess and redistribute monitoring locations for an operating monitoring network is proposed in this section. The assessment and redesign of the water quality monitoring locations consists of two main steps. The first step is to group the SUs into clusters of similar SUs with respect to their attributes. In the second step, stratified optimum sampling is applied to identify the optimal number of gauged SUs in each of the identified clusters.

3.3.1 Hybrid-cluster algorithm

To build clusters of SUs with similar attributes, cluster analysis (CA) can be employed. CA is an exploratory data technique used to group similar observations into clusters, where the within-cluster variance is minimized and the between-cluster variance is maximized (Peck et al., 1989; Jobson, 1992). In earlier years, the validity of clustering techniques was questioned because of the lack of inferential tests offered by many other statistical techniques (Baker and Hubert, 1975; Wong, 1982). However, using additional tests such as combining hierarchical and partitional (non-hierarchical) clustering techniques can provide validation for the chosen clusters (Jobson, 1992).

Hierarchical methods most commonly use agglomerative techniques where each observation starts in a cluster by itself (*N* SUs = *N* clusters). As the algorithm progresses, observations are joined based on a linkage distance until there is only one cluster composed of all the observations (*N* clusters = 1). K-means clustering is a partitioning method based on the Euclidean distance where the optimal cluster centers are obtained in an iterative fashion by minimizing the sum of the squared distances between the cluster means and the cluster members. The objective of this technique is to divide *N* observations (SUs) with *P* dimensions (attributes) into *K* clusters so that the within-cluster sum of squares is minimized (Hartigan, 1975). The algorithm separates the data into clusters by identifying a set of cluster centers, assigning each SU to a cluster, determining new cluster centers, and repeating this process until no change occurs in the cluster centers and members. It is an iterative procedure in which the SU attribute vectors move from one cluster to another to minimize the value of the objective function (*F*) :

$$F = \sum_{k=1}^{K}\sum_{j=1}^{P}\sum_{i=1}^{N_k} d^2 \left(S_{ijk} - C_{jk}\right) \tag{12}$$

where *K* denotes the number of clusters, *P* is the number of attributes, N_k represents the number of SUs in cluster *k*, *d* is the linkage distance, S_{ijk} denotes the value of attribute *j* in the SU vector *i* assigned to cluster *k* and C_{jk} is the mean value of attribute *j* for cluster *k* defined by the following equation:

$$C_{jk} = \frac{1}{N_k}\sum_{i=1}^{N_k} S_{ijk} \tag{13}$$

One of the benefits of using hierarchical methods is the generation of the dendrogram, which can be very useful in visualizing “good” cluster partition points based on the linkage distance. The question of the ideal number of clusters is addressed by Jobson (1992). Jobson (1992) suggested graphing the linkage distance for each number of clusters and picking the number of clusters prior to an obvious increase in the linkage distance. This point identifies a large increase in the linkage distance that is needed to join additional observations.

While hierarchical clustering procedures are not influenced by initialization and local minima, partitional clustering procedures are influenced by the initial guesses (e.g., the number of clusters and cluster centers) (Rao and Srinivas, 2006). The partitional clustering procedures are dynamic in the sense that a SU can move from one cluster to another to minimize the objective function (Equation 12). In contrast, the SU committed to a cluster in the early stages cannot move to another one under hierarchical clustering procedures. The relative merits of the hierarchical and partitional clustering methods spurred the development of hybrid-clustering methods that are a blend of these methods (Srinivas et al., 2002; Rao and Srinivas, 2006).

K-means are initialized to *K* random points from the data or are input by the user. The initial seeds (cluster centers) may be obtained from a previous clustering method (Jobson 1992). Cluster centres developed based on hierarchical method provide the non-hierarchical method with an initial partition but do not force the non-hierarchical method to strictly match the hierarchical results (Srivivas et al., 2002). Seeding merely provides a rational starting point that increases confidence in the final results and decreases the time in which the partitional method takes to arrive at an optimum solution (Rao and Srinivas, 2006).

Hosking and Wallis (1997) adjusted the clusters obtained by Ward's minimum variance algorithm using the K-means algorithm. They used one at-site attribute (the drainage basin area) and three geographic location attributes (the gauge elevation, latitude and longitude) in cluster analysis to identify homogeneous watersheds based on the four selected attributes. Recently, Rao and Srinivas (2006) compared three hybrid-clustering algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures, for the regionalization of watersheds for flood-frequency analysis. The hierarchical clustering algorithms considered for hybridization were single linkage, complete linkage and Ward's minimum variance algorithms, while the partitional clustering algorithm used was the K-means. The relative performances of the three hybrid-cluster algorithms, the three hierarchical clustering algorithms and the K-means were evaluated by using annual maximum flow data from watersheds in Indiana, USA. The overall performance of the hybrid models in optimizing the objective function was found to be better than that of the hierarchical and K-means clustering algorithms. Of the three hybrid models presented, the combination of Ward's and K-means algorithms consistently provided good estimates of the groups of watersheds.

In this study, a hybrid-clustering algorithm is employed to identify clusters of similar SUs with respect to their attributes. Each of the considered attributes was standardized prior to the cluster analysis in order to remove the dimensionality and scale effects. The Euclidian distance is used as a proximity measure to define the distance between the SUs, and Ward's algorithm is used to define the various clusters in an agglomerative hierarchical algorithm. From the hierarchical cluster algorithm, different numbers of clusters are considered, and cluster centers are computed and then used as seeds for the K-means clustering algorithm.

The Euclidean distance between the attributes of two SUs can be used to quantify the distance (dissimilarity) between two SUs or between the SU and the cluster center. The Euclidean distance between the SUs $S_1 = (s_{11}, s_{12}, ..., s_{1P})$ and $S_2 = (s_{21}, s_{22}, ..., s_{2P})$ in the Euclidean P-space is defined as:

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{P} (s_{1i} - s_{2i})^2} \qquad (14)$$

where s_{1i} stands for the i^{th} attribute in the first SU and P is the number of attributes. As for Ward's algorithm, the objective function (*W)* (Ward, 1963) minimizes the sum of the squares of the deviations of the SU attribute vectors from the centers of their respective clusters:

$$W = \sum_{k=1}^{K} \sum_{j=1}^{P} \sum_{i=1}^{N_k} \left(S_{ijk} - C_{jk} \right)^2 \tag{15}$$

Ward's algorithm starts with single-SU clusters. At this point, the cluster centers are the same as the attribute vectors. Therefore, the value of *W* is zero. At each step in the analysis, the union of every possible pair of clusters is considered and two clusters whose combination results in the smallest increase in *W* are merged. The change in the value of the objective function, *W*, due to the merger depends only on the relationship between the two merged clusters and not on the relationships with other clusters. Ward's algorithm is good at recovering the cluster structure, and it tends to form clusters of equal size. This characteristic of Ward's algorithm makes it useful for the identification of homogeneous regions (SUs, in our case) (Hosking and Wallis, 1997; Rao and Srinivas, 2006; Ouarda et al., 2008).

3.3.2 Stratified optimum sampling

The methodology proposed in this section can be applied for both expansion and reduction of the number of monitoring locations. If the monitoring budget allows increasing the number of monitoring locations by *u*, then the total number of locations to be distributed among clusters becomes (*M* + *u*), where *M* is the current number of monitored locations. Similarly, if the budget requires reducing the number of monitoring locations by *u*, the total number of locations to be distributed becomes (*M* - *u*).

From the first step, by applying the hybrid-cluster algorithm, the number of clusters, the number of SUs in each cluster, the cluster centers and the distance between each SU and its cluster center are identified. If the objective is to establish a new monitoring network, the question involves how to distribute *M* monitoring locations among obtained clusters? To distribute *M* monitoring locations among the obtained clusters, a stratified optimum sampling strategy is applied. The optimal allocation for each cluster of SUs is proportional to the standard deviation of the distribution of the SUs within the cluster. A larger number of locations is allocated in the group with the greatest variability as follows:

$$m_i = M \frac{N_i \sigma_i}{\sum_{i=1}^{k} N_i \sigma_i} \qquad (16)$$

where m_i is the number of SUs to be gauged in cluster (*i*), N_i is the number of SUs in cluster (*i*), σ_i is the standard deviation for cluster (*i*) (which is the standard deviation of distances between each of the cluster members and the cluster center), and *k* is the number of clusters, where:

$$M = \sum_{i=1}^{k} m_i \, . \qquad (17)$$

In the case where the objective is the assessment and redistribution of water quality monitoring locations of an operating monitoring network, one can proceed as with the design approach to identify the optimal number of gauged SUs at each cluster by applying Equation 16. Comparing the optimal number of gauged SUs for each cluster and the number of already gauged SUs, one may have three cases for the number of already-gauged SUs within the cluster: (i) equal to, (ii) less than, or (iii) greater than the optimal number of gauged SUs. In the first case, no action should be taken; in the second case, monitoring locations should be added at some of the un-gauged SUs; finally, in the third case, some of the already-gauged SUs may be discontinued.

When adding monitoring location(s) to a cluster that exhibits a shortage in the number of monitoring locations, the un-gauged SUs with the greatest distance from gauged SUs should have the first priority. These SUs are indicative of SUs within the cluster that explain a greater amount of variability, and therefore, they offer the greatest benefit in terms of information gained. If only one monitoring location is to be added, the minimum distance in the attribute space between each un-gauged SU and the gauged SUs is computed. Then the un-gauged SU with the maximum value of the minimum distance is selected to be gauged. When more than one monitoring location must be added, a similar procedure is followed to identify the second un-gauged SU, the minimum distance is computed between each of the remaining un-gauged SUs and each of the gauged SUs including the un-gauged SU selected in the first step (considered gauged) and so forth.

When discontinuing monitoring location(s) from over-monitored cluster(s), an aggregated information index is applied to identify the optimal combination of monitoring locations to be discontinued. For instance, consider the case where the assessment carried out in the previous steps requires h gauged SUs to be discontinued (h = the number of already-gauged SUs – the optimum number). Which h SUs among the v gauged SUs in the over-monitored cluster should be selected? The number of possible combinations of SUs to discontinue is given by the binomial coefficient, $C(v,h)$. For each combination, one may compute an information index, according to which the combinations may be ranked.

Such a procedure allows the identification of the best combination of gauged SUs to discontinue or provides the decision-maker with the rank of the best combinations to discontinue. For practical comparison of the combinations, the information index must be based on some kind of aggregated information (Ouarda et al., 1996). Khalil et al. (2010) used an aggregated performance index (I_a) to evaluate combinations of water quality variables to be discontinued based on the variance of the estimated mean value, which can be modified for the case of locations as follows:

$$I_a = \sum_{Locations} \sum_{\mathrm{var}iables\ V} \sqrt{Var\{\hat{\mu}\{V\}\}} \tag{18}$$

where V is the water quality variable and $Var\{\hat{\mu}\{V\}\}$ is the variance of the mean value estimator expected after n_2 years (Equation 2). The summation in I_a is carried out over all of the selected variables and over all of the gauged locations within the over-monitored cluster. For variables at a discontinued SU, the variance of the mean value estimator after n_2 years is estimated using Equation 2. The population parameters in Equation 2 are replaced by their estimates based on the

n_1 years of data. For variables at continuously measured SUs, the variance of the mean value after n_2 years is assumed to be equal to the variance of the mean after n_1 years multiplied by (n_1-1)/(n_1+n_2-1). The performance index is applied to the standardized variables to remove the dimensionality and scale effects from the variables.

By applying the aggregated performance index of Equation 18, one can evaluate each of the possible combinations. After having examined all possible combinations, one can identify the optimal combination that has the minimum I_a, or one can even provide the decision-makers with the rank of all possible combinations. Figure 3 illustrates the flow of the analyses described above.

(Figure 3)

Thus, the proposed approach to assess and redesign the water quality monitoring locations can identify, in a systematic and objective manner, the optimal combination of gauged SUs to discontinue, the locations to be continuously measured and the SUs to be gauged.

4 Results

This section is divided into three subsections, respectively representing the data preparation, the empirical experiment results and the monitoring location assessment and redesign.

4.1 Data preparation

The data preparation consists of three steps. The first step is to divide the Nile Delta into SUs, where each SU is drained by one point on the drainage system. Ninety-four SUs are hence obtained, from which 50 are gauged and 44 are un-gauged (Figure 2). The second step is to identify the SUs attributes. The following attributes are identified for each SU: (i) cultivated area (feddan = 4200 m^2); (ii) average soil salinity (ppm); (iii) average soil hydraulic conductivity (m/d); (iv) average total annual rainfall (mm/year); (v) drainage system total length (km); (vi) average total industrial effluent (m^3/day); (vii) waste water plants total capacity (m^3/day); (viii) number of livestock; and (ix) average annual applied fertilizers (tons/year). Although the population density is an important attribute, it is excluded from this study due to limited information and difficulty in identifying the population for each SU. The summary statistics of the identified attributes are presented in Table 1.

(Table 1)

The third step is to identify variables best describing the variation in the Nile Delta drainage water quality using PCA. The results show that the first seven principal components have eigenvalues greater than one and explain about 85.08% of the total variance in the original data set. Table 2 shows components loading matrix, where the percentage of variance explained by each principal component and the rotated correlation coefficients for each of the water quality variables and the considered components are presented. The values shown are the correlations between the variables and the principal components. Because of the large number of principal components extracted, 0.7 (70%) is considered the cut-off value for the correlation coefficients.

Thus, any water quality variable with an absolute component correlation coefficient value greater than or equal to 0.7 is considered to be an important variable contributing to variations of the Nile Delta drainage water quality.

(Table 2)

Based on the PCA results (Table 2), the first principal component represents mineral-related variables (Electric Conductivity (EC), Total Dissolved Solids (TDS), Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Bi-Carbonate (HCO_3), Sulphate (SO_4) and Chloride (Cl)). The second component consists of Total Coliform (TCol), Fecal Coliform (FCol), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Phosphorus (TP) and Water Temperature (Temp.), and it is negatively correlated with the pH and Dissolved Oxygen (DO). The third component represents turbidity with four most correlated variables: Total Suspended Sediments (TSS), Total Volatile Solids (TVS), Turbidity (turbid.) and Visibility (Visib.) (negatively correlated). The fourth component represents nitrates, including Nitrate (NO_3), Ammonium (NH_4) and Total Nitrogen (TN). The fifth to seventh components represent trace elements such as Cadmium (Cd), Copper (Cu), Iron (Fe), Manganese (Mn), Zink (Zn), Lead (Pb) and Nickel (Ni). The highly correlated variables in the first four components are selected as indicators: TDS, BOD, TVS and TN.

4.2 Empirical experiment

The results of the empirical experiment designed to examine the three record extension techniques are presented herein. Figures 4 and 5 summarize the results obtained for the ratio *U*.

Figure 4 shows the distribution of the ratio *U* for the estimation of the mean and standard deviation, while Figure 5 shows the distribution of the ratio *U* for the estimation of different non-exceedance percentiles. The box plots in Figures 4 and 5 are constructed from 9000 records. This number of records is based on four water quality variables selected at each of the 50 considered locations, for 45 different record combinations. The box plots in Figures 4 and 5 represent the distribution of the ratio *U* for a given statistic and record extension technique. The accuracy of each approach can be judged by the degree of dispersion in the box plots, by the closeness of the median to a value of 1 and by the symmetry of the box plot around the value of 1 (Hirsch, 1982; Vogel and Stedinger, 1985).

(Figures 4 and 5)

Figure 4 indicates that for the estimation of the mean, the regression, MOVE3 and ANN techniques lead to median values of *U* equal to 0.96, 1 and 0.97 respectively. The boxes are symmetric around 1 and have almost the same dispersion. Figure 4 shows also that the regression and ANN techniques tend to underestimate standard deviations. The cross-validation shows that more than 75 percent of the regression and almost 75% of the ANN standard deviations are lower than the values estimated from the observed records, with median values of 0.62 and 0.73, respectively. For the MOVE3 technique, the median value is 0.98. The MOVE3 standard deviation box plot is more symmetric around 1 and shows almost the same dispersion as ANN, while the regression box plot shows the least dispersion.

Figure 5 shows that the median values of the ratio *U* for low percentiles are higher than 1, while those corresponding to high percentiles are lower than 1, for the regression and ANN. In general, the median values of the ratio *U* for the MOVE3 percentiles are very close to 1. The ratio *U* median values range between 0.81 and 1.25 for the regression, between 0.85 and 1.24 for ANN and between 0.99 and 1.01 for MOVE3. In general, MOVE3 box plots are symmetric around 1 and show relatively less dispersion than those corresponding to the regression and ANN for low percentiles. For high percentiles, box plots corresponding to the three techniques show almost the same dispersion. These results suggest that the regression and ANN tend to overestimate low percentiles and underestimate high percentiles, while MOVE3 reduces the bias exhibited by the other two techniques.

Figure 6 illustrates the *BIASr* exhibited by the three extension techniques in estimating water quality concentration percentiles. Results indicate that MOVE3 *BIASr* values for extreme percentiles are closer to zero than regression and ANN *BIASr* values. The curves corresponding to the MOVE3, regression and ANN techniques intersect at the median as the regression and ANN techniques overestimate low percentiles and underestimate high percentiles. Figure 7 illustrates the *RMSEr* exhibited by the three extension techniques in estimating water quality concentration percentiles. Figure 7 shows that the MOVE3 *RMSEr* values are closer to zero than those corresponding to regression and ANN for low percentiles, while for large percentiles there is no significant difference between the three techniques. Figures 5 and 6 clearly illustrate the regression and ANN overestimation of low concentrations and underestimation of high concentrations, as would be expected from their tendency to produce an extended record with a lower variance than the observed record.

(Figures 6 and 7)

4.3 Monitoring location assessment and redesign

Based on the SUs attributes, Figure 8 shows the SU hierarchical clustering tree (dendrogram), where the x-axis indicates the SU code and the y-axis indicates the linkage distance between the clusters. Figure 9 shows the relationship between the linkage distance and the number of clusters. There is a sharp change of slope around a linkage distance of 5, corresponding to 11 clusters. Based on Figure 9, 11 clusters are selected, and the cluster centers are computed.

(Figures 8 and 9)

The K-means clustering algorithm is then employed, using the computed cluster centers (initial seeds) to identify 11 SU clusters. The K-means cluster analysis almost verified the 11 clusters obtained in Figure 8, with slight differences. Table 3 shows the K-means cluster members and the distance of each member from its cluster center. Confirming the K-means clusters, the 94 SUs are thus grouped into 11 clusters. Table 3 shows that two out of the eleven clusters are single-SU clusters. These are SU 3 (Cluster O) and SU 41 (Cluster Z).

SU 3 exhibits the highest WWTPs capacity in the Nile Delta. The drainage system of SU 3 receives around 1415000 m^3/day of treated sewage from the WWTPs serving eastern Cairo. The total WWTPs capacity of SU 3 is more than ten folds that of the second highest SU (SU 11). The drainage system of SU 11 receives around 133483 m^3/day of treated sewage from the WWTPs serving the Mansoura city in the Eastern Delta. SU 41 shows the highest industrial activities within the Nile Delta. It includes about 50 factories in the Alexandria industrial region with a total industrial effluent of around 123561 m^3/day. The total industrial effluents of SU 41 is almost four times that of the second highest SU (SU 3), where the drainage system of SU 3 receives around 34139 m^3/day of industrial effluent from factories in eastern Cairo. Consequently, these two SUs should be continuously measured.

In addition, cluster H includes two SUs (SUs 48 and 50) which are equal distance from the cluster center. Consequently, the distances standard deviation is equal to zero. Hence, by applying stratified optimum sampling, no gauged locations will be allocated to cluster H. Since both SUs are gauged, one of these SUs may be discontinued.

(Table 3)

Table 4 shows the number of SUs in each cluster, the number of gauged SUs and the standard deviation within each cluster. In the case, the objective is to assess and redistribute the 50 gauged locations, SUs 3 and 41, which form the single-SU clusters O and Z, respectively, should be continuously measured. One of the SUs 48 and 50, which are the members of cluster H should be discontinued. Thus, for the case of no change in the number of monitoring locations, 47 monitoring locations are distributed among the remaining eight clusters.

(Table 4)

By applying the stratified optimum sampling (Equation 16), the optimal number of gauged SUs in each of the 11 clusters is determined, as shown in Table 4. The optimal number of gauged SUs is equal to the number of already-gauged SUs at each of clusters A, O and Z. It is greater than the number of already-gauged SUs at each of clusters C, E, G and L. It is smaller than the number of already-gauged SUs at each of clusters B, D, F and H.

Thus, no action is taken for clusters A, O and Z, and the already-gauged SUs in these clusters should be continuously measured. For clusters C, E, G and L, three, three, two and three monitoring locations should be added, respectively. The addition of new monitoring locations at un-gauged SUs is based on the minimum distance between each un-gauged SU and the gauged SUs in the attributes space. The un-gauged SU with the maximum minimum distance is the first to be gauged. For instance, monitoring locations are added at SUs 51, 61 and 80 in cluster C. These SUs show the maximum distances calculated for un-gauged SUs in cluster C. The SUs marked by (*) in Table 3 are selected to be gauged in these four clusters.

For clusters B, D, F and H, five, one, four and one of the gauged SUs should be discontinued, respectively. For cluster D, where three SUs are already-gauged, I_a is applied to identify which SU should be discontinued. Assuming that SU 13 is to be discontinued, I_a is 1.4408, when SU 19 is assumed to be discontinued, I_a is 1.4410, and it is 1.4411 when SU 18 is assumed to be discontinued. Similarly for cluster H, in case SU 48 is assumed to be discontinued, I_a is 1.4411, while if SU 50 is assumed to be discontinued, I_a is 1.4412. Thus, based on I_a, SUs 13 and 48 may be discontinued from cluster D and H respectively. Table 5 shows different combinations of five gauged SUs to be discontinued from cluster B. Table 5 shows the first 9 combinations out of 252 possible combinations ranked based on I_a. Similarly, Table 6 shows the first 9 combinations out of 1001 possible combinations for the discontinuation of four gauged SUs from cluster F.

(Tables 5 and 6)

For instance, if the monitoring budget allows increasing the number of monitoring locations by 10%, stratified optimum sampling is used to distribute 55 monitoring locations among the 11 clusters. Given that two of the eleven clusters are single-SU clusters and one of the SUs in cluster H should be discontinued, stratified optimum sampling is used to distribute 52 locations among the remaining eight clusters. Table 4 shows the optimal number of gauged locations at each cluster for the case of increasing the number of locations by 10%. Similarly, if the monitoring budget requires decreasing the number of monitoring locations by 10 %, stratified optimum sampling is used to distribute 42 monitoring locations among the eight multi-SU clusters.

5 Conclusions

A methodology for the assessment and redesign of surface water quality monitoring location is presented in this paper. The proposed methodology is applied to assess the Nile Delta drainage water-quality-monitoring network. The proposed methodology incorporates basin attributes to cluster the Nile Delta SUs into groups of similar SUs using a hybrid-cluster algorithm. Stratified optimum sampling is then applied to identify the optimum number of gauged SUs at each cluster.

Dividing the monitored basin into clusters of spatial units with similar attributes and applying a stratified optimum sampling strategy allows to spatially distribute the monitoring locations for the establishment of a new water quality monitoring network. For an operating monitoring network, the proposed methodology can be applied to expand or contract the monitoring network. In the case of over-monitored clusters, an aggregate information index is used to identify the best combination of locations to be discontinued. For the under-monitored clusters, the distance between each of the un-gauged SUs and the gauged SUs in the attributes space is used to identify locations to be added. It can be concluded that the proposed approach can systematically and objectively assess and identify the monitoring locations to be continuously measured, the locations to be discontinued and the locations to be added.

In addition, an empirical experiment is conducted to examine the utility of three record extension techniques in reconstituting information regarding variables at discontinued locations. Simple linear regression, MOVE3 and ANN techniques are applied to reconstitute information about variables at discontinued locations using the data from the case study. Different statistical performance measures are used to assess each of the extension techniques and their ability to maintain the statistical characteristics of the water quality records. The MOVE3 technique showed better performance in preserving the statistical characteristics of the discontinued water quality variables. Consequently, the MOVE3 technique is recommended for the reconstitution of information about water quality variables at discontinued locations in the Nile Delta WQM network.

The main advantages of the proposed methodology are as follows: 1) it allows for the spatial distribution of monitoring locations when no water quality data is available and for the expansion or contraction of the number of monitoring locations in an operating monitoring network; and 2) it allows for the reconstitution of information about water quality variables at discontinued locations. Thus, the use of the proposed approach and MOVE3 allows to overcome the deficiencies inherent in the conventional approaches used to design the monitoring locations.

Parallel to the proposed methodology, a cost analysis can be introduced. This analysis would help to address the trade-off between the number of water quality monitoring locations on one hand and the sampling frequency and number of water quality variables to be measured on the other hand. Thus, the decision would be either to reduce the number of monitoring locations in favor of keeping more variables and increasing the sampling frequency or to increase the number of monitoring locations while reducing the number of variables to be measured and/or the sampling frequency.

Acknowledgments: The authors wish to thank Prof. Shaden Abdel-Gawad, president of the National Water Research Center of Egypt, for providing the water quality data used in this study. The authors also wish to acknowledge the support provided by Prof. Alaa Abdin, director of the Strategic Research Unit of the NWRC in the identification of the Nile Delta SUs and for providing the SU attribute data used in this study. The financial support provided by Helwan University, Cairo, Egypt, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs (CRC) program is acknowledged.

6 References

[1]. Abdel-Gawad, S.T., Kandil, H.M. and Sadek, T.M. (2004). Water scarcity prospects in Egypt 2000-2050, in: Marquina (ed.) Environmental Challenges in the Mediterranean 2000-2050, Dordrecht: Kluwer Academic Publishers, 187 - 203.

[2]. Abu-salama, M. S. M. (2007). Spatial and temporal consolidation of drainage water quality monitoring networks, Ph.D. dissertation, Universität Lüneburg, Fakultät III, Umwelt und Technik, Germany.

[3]. Alley, W.M. and Burns, A.W. (1983). Mixed-station extension of monthly streamflow records. Journal of Hydraulic Engineering, 109 (10), 1272 - 1284.

[4]. Baker, F.B. and Hubert, L.J. (1975). Measuring power of hierarchical cluster analysis. Journal of the American Statistical Association, 70 (349), 31 - 38.

[5]. Berryman, D., Bobée, B., Cluis D. and Haemmerli, J. (1988). Nonparametric Tests for Trend Detection in Water Quality Time Series. Water Resources Bulletin, 24(3), 545 - 556.

[6]. Cochran, W.G. (1953). Sampling Techniques, John Wiley, New York, p. 428.

[7]. Dijkman, J.P.M. (1993). Environmental Action Plan of Egypt, A Working Paper on Water Resources. Directorate of General International Cooperation, Ministry of Foreign Affairs, the Netherlands, 116 - 127.

[8]. Draper, N.R. and Smith, H. (1966). Applied regression analysis, John Wiley, New York, 736 p.

[9]. DRI (Drainage Research Institute) - MADWQ (1998). Monitoring and analysis of drainage water quality in Egypt, Interim Report, Cairo.

[10]. Frenken, K., (2005). Irrigation in Africa in Figures, Aquastat Survey (2005), Food & Agriculture Org, Rome, Italy, 88 p.

[11]. Harmancioglu, N.B. and Alpaslan, M.N. (1992). Water quality monitoring network design: a problem of multi-objective decision making, Water Resources Bulletin, 28 (1), 179 - 192.

[12]. Harmancioglu, N.B., Alpaslan, N., Alkan, A., Ozkul, S., Mazlum, S. and Fistikoglu, O. (1994). Design and Evaluation of Water Quality Monitoring Networks for Environmental Management (in Turkish). Report prepared for the research project granted by TUBITAK, Scientific and Technical Council of Turkey, Project code: DEBAG-23, 514 p.

[13]. Harmancioglu, N.B., Fistikoglu, O., Ozkul, S.D., Singh, V.P. and Alpaslan, M.N. (1999). Water Quality Monitoring Network Design. Kluwer Academic Publishers, Dordrecht, the Netherlands, 290 p.

[14]. Hartigan, J.A. (1975). Clustering Algorithms, New York: John Wiley & Sons, Inc.

[15]. Hirsch, R.M. (1982). A comparison of four streamflow record extension techniques. Water Resources Research, 18(4), 1081 - 1088.

[16]. Horton, R.E. (1945). Erosional Development of Streams. Geological Society Am. Bull., 56, 281 - 283.

[17]. Hosking, J.R.M., Wallis, J.R. (1997). Regional frequency analysis: an approach based on L-moments. Cambridge University Press, New York, USA.

[18]. Jobson, J.D. (1992). Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods. New York, Springer-Verlag, 768 p.

[19]. Khalil, B. and Ouarda, T.B.M.J. (2009). Statistical approaches used to assess and redesign surface water quality monitoring networks, Journal of Environmental Monitoring, doi: 10.1039/b909521g., 11, 1915 - 1929.

[20]. Khalil, B., T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana (2010). A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks. Journal of Hydrology, 386, 173-185.

[21]. Lettenmaier, D.P. (1988). Multivariate nonparametric tests for trend in water quality, AWRA, Water Resources Bulletin (24)3, 505 - 512.

[22]. Matalas, N.C. and Jacobs, B. (1964). A correlation procedure for augmenting hydrologic data, U.S. Geol. Surv. Prof. Pap., 434-E, E1-E7.

[23]. Moog, D.B. and Whiting P.J. (1999). Streamflow record extension using power transformations and application to sediment transport, Water Resources Research, vol. 35 (1), 243 - 254.

[24]. MWRI, Ministry of Water Resources and Irrigation (1997). Review of Egypt's Water Policies, Strengthening the Planning Sector Project, Ministry OF Water Resources and Irrigation, Cairo, Egypt.

[25]. NAWQAM, National Water Quality and Availability Management Project (2001). Evaluation and Design of Egypt National Water Quality Monitoring Network. Report no.: WQ-TE-0110-005-DR, NAWQAM, NWRC, Cairo, Egypt.

[26]. Odom, K.R. (2003). Assessment and Redesign of the Synoptic water quality monitoring network in the Great Smoky Mountains National Park. Ph.D. Dissertation, University of Tennessee, Knoxville, USA, 268 p.

[27]. Ouarda, T.B.M.J. and C. Shu (2009). Regional low-flow frequency analysis using single and ensemble artificial neural networks, Water Resour. Res., 45, W11428, doi:10.1029/2008WR007196.

[28]. Ouarda, T.B.M.J., K.M. Ba, C. Diaz-Delgado, A. Carsteanu, K. Chokmani, H. Gingras, E. Quentin, E. Trujillo and B. Bobee (2008). Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study, Journal of Hydrology, 348, 40-58.

[29]. Ouarda, T.B.M.J., Rasmussen, P.F., Bobée, B., and Morin, J. (1996) Ontario Hydrometric Network Rationalization, Statistical Considerations, Research Report No. R-470, National Institute for Scientific Research, INRS-ETE, University of Québec, Québec, Canada, 75 p.

[30]. Ozkul, S.D. (1996). Space / Time Design of Water Quality Monitoring Networks by the Entropy Method, Ph.D. Thesis on Civil Engineering, Dokuz Eylul University, Graduate School of Natural and Applied Sciences, Izmir, 196 p.

[31]. Peck, R., Fisher, L. and Van Ness, J. (1989). Approximate Confidence Intervals for the Number of Clusters. Journal of the American Statistical Association, 84 (405), 184 - 191.

[32]. Rao, A.R. and Srinivas, V.V. (2006). Regionalization of watersheds by hydbrid-cluster analysis. Journal of Hydrology, 318, 37 -56.

[33]. Rumelhart, D.E. and J.L. McClelland (Eds.) (1986). Parallel Distributed Processing: Explorations in the Microstructure of Congnition, vol. 1, Foundations, MIT Press, Cambride, Mass.

[34]. Sanders, T.G., Ward, R.C., Loftis, J.C., Steele, T.D., Adrian, D.D. and Yevjevich, V. (1983). Design of Networks for Monitoring Water Quality. Water Resources Publications, Littleton, Colorado, 328 p.

[35]. Sharp, W.E. (1970). Stream Order as a Measure of Sample Source Uncertainty. Water Resources Research, 6 (3), 919 - 926.

[36]. Sharp, W.E. (1971). A topologically optimum water sampling plan for rivers and streams. Water Resources Research, 7 (6), 1641 - 1646.

[37]. Shu, C. and D.H. Burn (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis, Water Resources Research, 40, W09301, doi: 10.1029/2003WR002816.

[38]. Shu, C. and Ouarda, T.B.M.J. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43, W07438, doi:10.1029/2006WR005142.

[39]. Srinivas, V.V., Rao, A.R., Govindaraju, R.S., (2002). A hybrid cluster analysis for regionalization. Proceedings of ASCE Environmental and Water Resources Institute (EWRI) Conference, Roanoke, VA, USA.

[40]. Tabachnick, B.G. and Fidell, L.S. (1996). Using Multivariate Statistics. Allyn and Bacon, Boston, London, 879 p.

[41]. Tirsch, F.S. and Male, J.W. (1984). River basin water quality monitoring network design: options for reaching water quality goals, in: T.M. Schad (ed.). Proceeding of Twentieth Annual Conference of American Water Resources Associations, AWRA Publications, 149 - 156.

[42]. Vogel, R.M. and Stedinger, J.R. (1985). Minimum variance streamflow record augmentation procedures. Water Resources Research, 21(5), 715 - 723.

[43]. Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

[44]. Ward, R.C., Loftis, J.O. and McBride, G.B. (1990). Design of Water Quality Monitoring systems, Van Nostrand Reinhold, New York, USA, 231 p.

[45]. Wong, M. A. (1982). A Hybrid Clustering Method for Identifying High-Density Clusters. Journal of the American Statistical Association, 77 (380), 841 - 847.

Table 1. Descriptive Statistics of SU attributes

SU Attributes	Minimum	Mean	Maximum	STDV
Cultivated area (fedd)	912.51	63323.11	294852.06	50902.16
Soil Salinity (ppm)	1280.00	2027.57	2560.00	281.77
Soil hydraulic conductivity (m/day)	0.08	0.19	1.00	0.15
Average Rainfall (mm/year)	37.50	87.43	150.00	35.07
Drains total length (km)	18.12	173.25	540.97	115.11
Total Industrial effluent (m3/day)	0.00	4420.42	123561.00	18314.81
Total WWTP capacity (m3/day)	0.00	37078.78	1415000.00	200098.72
Livestock (count)	1957.00	62470.27	206552.00	49447.63
Total annual fertilizers applied (tones/year)	444.64	11730.73	31047.36	7887.32

Table 2. PCA loading matrix for the Nile Delta water quality variables

Water Quality Variables	**Principal Components**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Discharge (Q)	0.427	-0.103	0.156	0.105	-0.116	0.094	0.602
Total Coliform (TCol)	-0.180	**0.931**	0.164	0.057	-0.076	0.138	-0.001
Fecal Coliform (FCol)	-0.182	**0.926**	0.160	0.058	-0.080	0.165	0.002
Biochemical Oxygen Demand (BOD)	-0.031	**0.961**	-0.032	0.014	0.111	-0.106	-0.057
Chemical Oxygen Demand (COD)	-0.021	**0.943**	-0.022	0.033	0.111	-0.113	-0.063
Total Suspended Solids (TSS)	0.125	0.012	**0.972**	0.001	0.049	0.003	-0.053
Total Volatile Solids (TVS)	0.115	0.006	**0.974**	0.009	0.047	-0.008	-0.067
Nitrate (NO_3)	0.210	-0.005	-0.015	**0.947**	-0.102	-0.054	0.030
Ammonium (NH_4)	0.134	0.536	-0.042	**0.781**	-0.013	-0.098	-0.034
Total Phosphorus (TP)	-0.162	**0.860**	0.209	0.130	-0.088	0.113	-0.109
Total Nitrogen (TN)	0.241	-0.078	-0.012	**0.954**	-0.050	0.023	0.023
Cadmium (Cd)	-0.183	-0.068	-0.083	-0.028	0.110	-0.129	0.693
Copper (Cu)	0.087	0.114	-0.038	-0.090	-0.641	0.179	-0.036
Iron (Fe)	-0.088	-0.060	0.070	-0.020	0.208	**0.888**	-0.140
Manganese (Mn)	0.136	0.301	0.385	-0.058	-0.170	0.616	0.071
Zinc (Zn)	0.164	-0.085	-0.130	-0.250	0.634	0.354	0.061
Lead (Pb)	0.165	0.226	0.037	0.029	0.569	0.110	0.285
Nickel (Ni)	0.130	0.036	0.007	-0.164	**0.769**	0.051	-0.206
Boron (B)	0.548	-0.042	-0.253	0.118	0.354	-0.040	-0.314
pH	0.331	**-0.700**	0.168	0.367	0.097	-0.124	-0.162
Electric Conductivity (EC)	**0.981**	-0.110	0.093	0.087	0.034	0.009	0.023
Total Dissolved Solids (TDS)	**0.982**	-0.116	0.076	0.085	0.035	0.004	0.016
Calcium (Ca)	**0.940**	-0.144	-0.130	0.130	0.085	-0.051	-0.029
Magnesium (Mg)	**0.969**	-0.116	0.134	0.090	0.001	0.036	0.023
Sodium (Na)	**0.978**	-0.107	0.121	0.067	0.027	0.017	0.033
Potassium (K)	**0.937**	-0.014	0.009	0.049	0.164	-0.013	-0.039
Bi-Carbonate (HCO_3)	**0.913**	-0.009	0.299	0.092	-0.115	0.049	-0.014
Sulphate (SO_4)	**0.935**	-0.176	-0.150	0.145	0.141	-0.061	-0.054
Chlorine (Cl)	**0.973**	-0.094	0.147	0.053	0.006	0.029	0.052
Water temperature (Temp.)	0.018	**0.732**	-0.102	0.129	0.213	-0.427	-0.167
Dissolved Oxygen (DO)	0.343	-0.682	-0.197	0.231	0.256	-0.231	-0.288
Turbidity (Turbid.)	0.117	0.024	**0.951**	0.028	-0.014	0.090	0.081
Visibility (Visib.)	0.030	-0.351	**-0.793**	0.103	0.135	-0.272	-0.099
% Variance exp.	31.83	19.14	12.17	8.35	5.59	4.39	3.61

Bold numbers indicate variables with an absolute factor correlation coefficient greater than or equal to 0.7. Shaded cells indicate the factor correlation coefficients for the selected variables

Table 3. Spatial Units Clusters

Cluster	SU	Dist.**	Cluster	SU	Dist.	Cluster	SU	Dist.
A	4	gauged	C	59	0.564	F	43	gauged
	10			61*^{1}	2.086		56	2.409
	33			62	1.134		74	2.676
	44			66	1.015	G	49	gauged
	47			80*^{2}	1.444		78	1.121
	52	1.755	D	13	gauged		81*^{2}	1.978
	60	1.583		18			82	1.544
	75	1.496		19			85*^{1}	2.555
	77	2.087		63	3.139		86	1.398
	88	0.705		73	2.266		87	0.750
	89	0.961	E	6	gauged	H	48	gauged
	90	0.488		17			50	
	92	0.795		25		L	22	
B	5	gauged		31			24	
	7			54*^{1}	2.580		38	
	8			55	0.985		39	
	9			58	1.286		40	
	11			64*^{3}	1.542		42	
	15			69*^{2}	1.568		45	
	27		F	2	gauged		46	
	29			12			70	0.264
	30			14			71	0.218
	35			16			72	0.482
	65	2.921		20			79*^{1}	1.761
	67	2.284		21			84*^{3}	0.583
	68	1.530		23			91	0.273
	76	1.925		26			93	0.391
	83	2.440		28			94*^{2}	0.959
C	1	gauged		32		O	3	gauged
	51*^{3}	1.431		34		Z	41	gauged
	53	0.832		36		* Dist. Stands for the minimum distance to gauged SU.		
	57	0.630		37				

*SUs selected to be gauged; *^{1} First SU to be gauged; *^{2} Second SU to be gauged

** Minimum distance to gauged SU.

Table 4. Stratified optimum sampling

Cluster	**No. SUs**	**Gauged SUs**	**Standard deviation**	**St-deviation × SU no.**	**Optimum no. gauged SUs**	**Expansion 10% (55)**	**Contraction 10% (45)**
A	13	5	0.075	0.975	5	5	4
B	15	10	0.058	0.87	5	5	4
C	9	1	0.089	0.801	4	5	4
D	5	3	0.097	0.485	2	3	2
E	9	4	0.157	1.962	7	8	6
F	16	14	0.127	2.032	10	11	9
G	7	1	0.087	0.609	3	3	3
H	2	2	0	0	1	1	1
L	16	8	0.137	2.192	11	12	10
O	1	1	0	0	1	1	1
Z	1	1	0	0	1	1	1

Table 5. Aggregate Information index to identify 5 locations to be discontinued from Cluster B

Location to discontinue					I_a
5	7	8	9	15	1.44056
5	7	8	9	27	1.44057
5	7	8	9	35	1.44059
5	7	9	15	27	1.44060
5	7	9	15	35	1.44061
5	7	8	15	27	1.44061
5	7	8	15	35	1.44062
5	7	9	27	35	1.44062
5	7	8	27	35	1.44064

Table 6. Aggregate Information index to identify 4 locations to be discontinued from Cluster F

Location to discontinue				***I_a***
12	23	26	32	1.44074
12	23	32	37	1.44077
2	12	23	32	1.44077
12	23	32	36	1.44078
12	23	26	37	1.44081
2	12	23	26	1.44081
12	16	23	32	1.44081
12	23	26	36	1.44082
2	12	23	37	1.44083

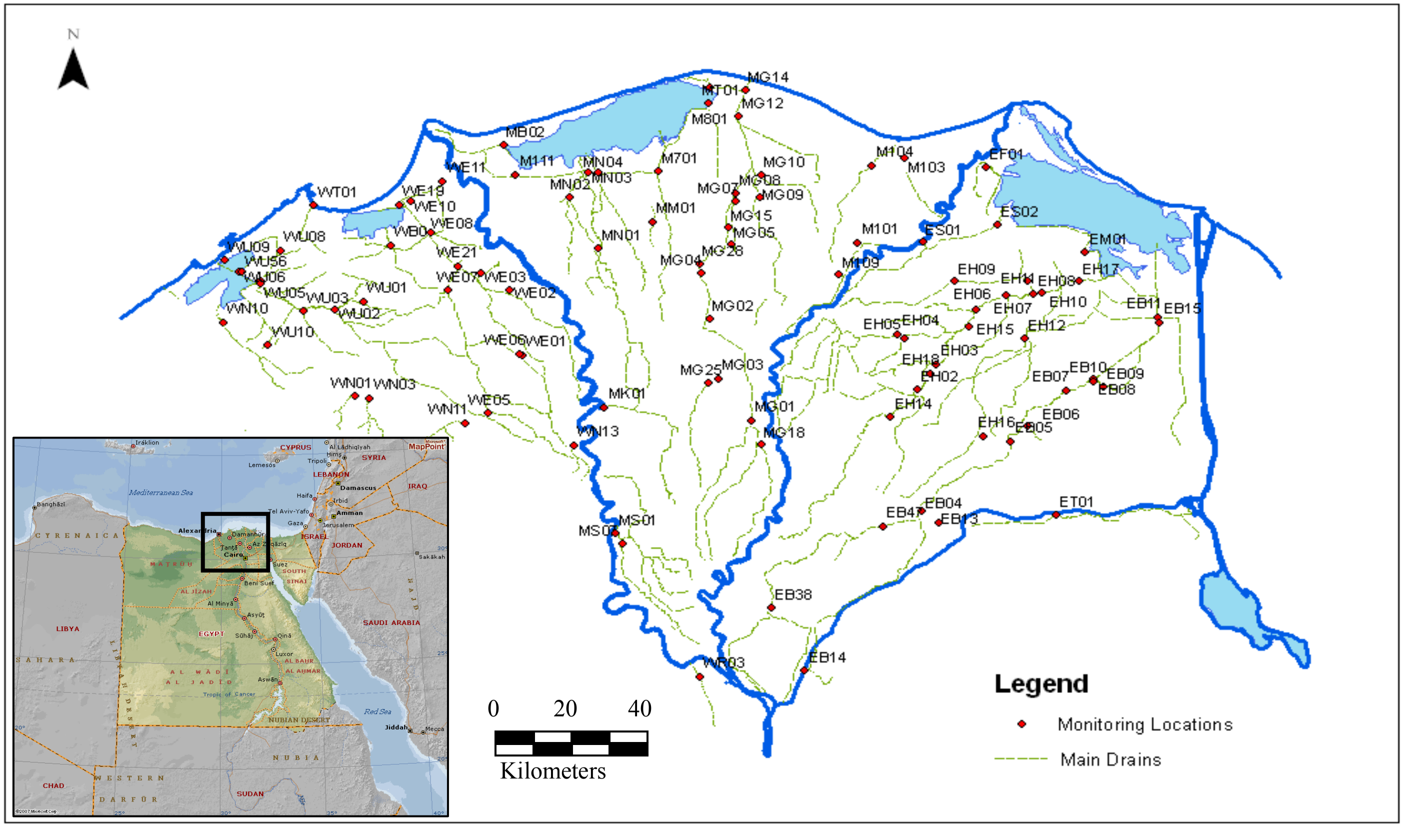


Figure 1. The Nile Delta surface WQM sites (source: NAWQAM, 2001)

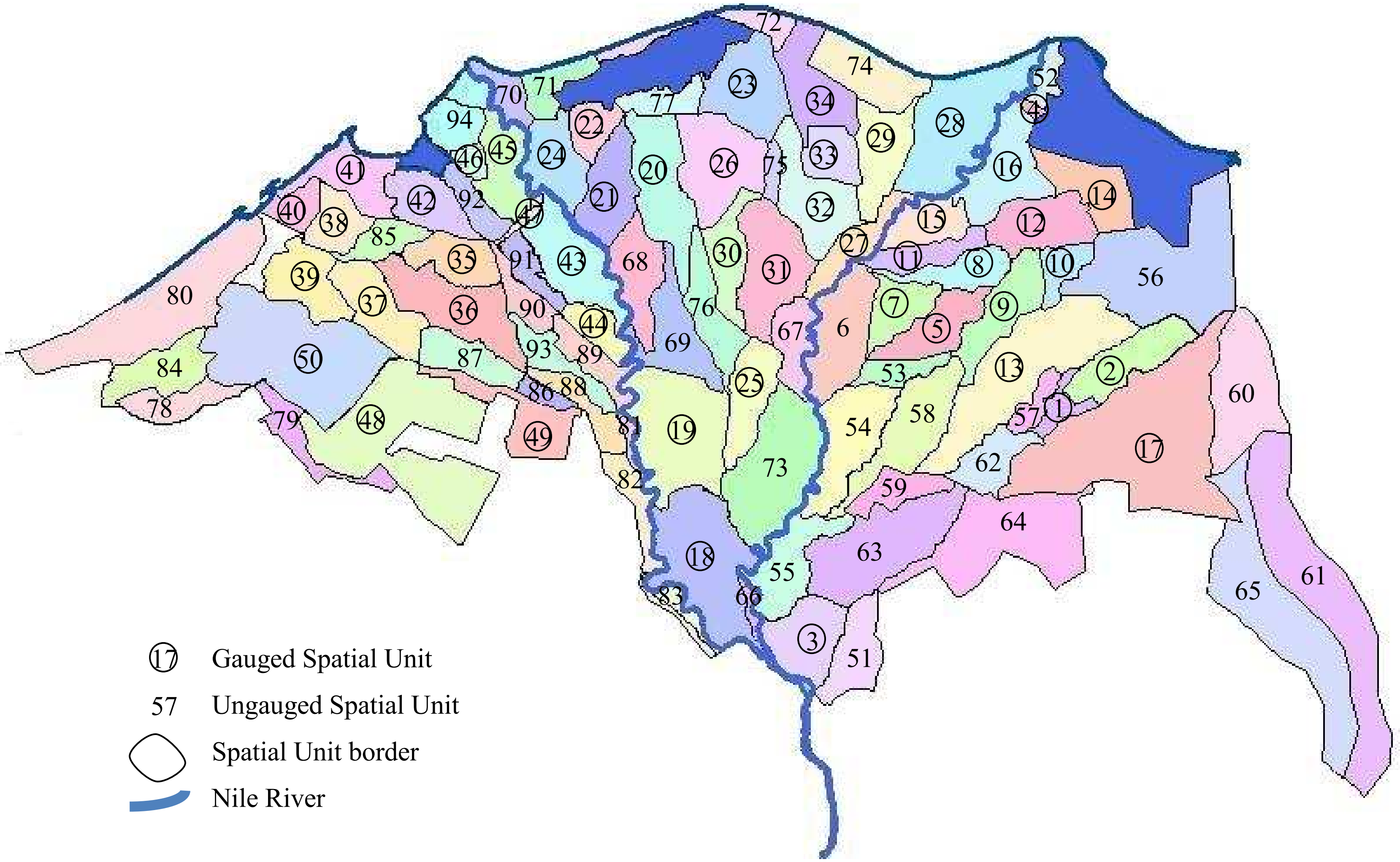


Figure 2. The Nile Delta Spatial Units

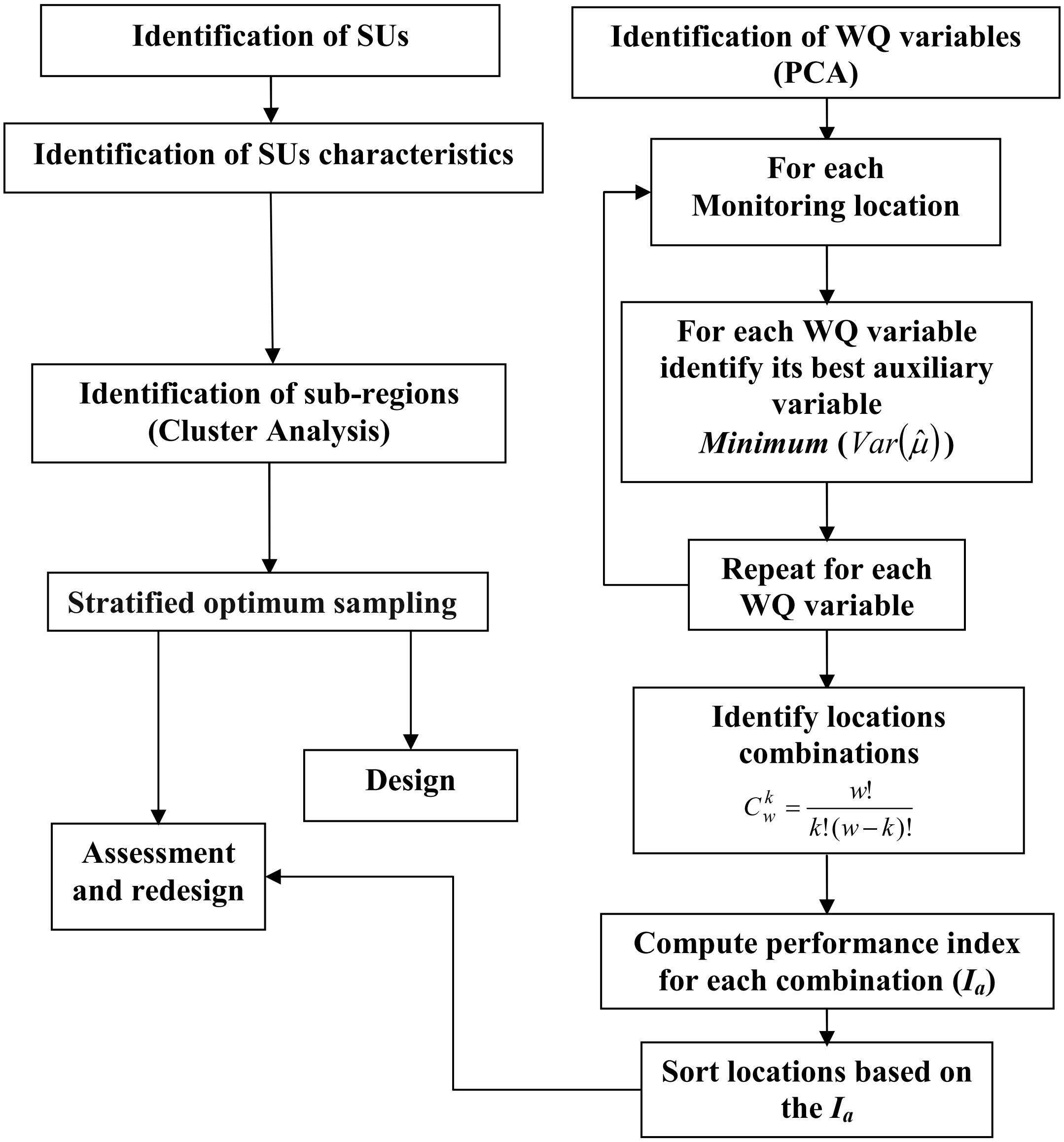

Identification of SUs
Identification of SUs characteristics
Identification of sub-regions (Cluster Analysis)
Stratified optimum sampling
Design
Assessment and redesign
Identification of WQ variables (PCA)
For each Monitoring location
For each WQ variable identify its best auxiliary variable
Minimum $(Var(\hat{\mu}))$
Repeat for each WQ variable
Identify locations combinations
$C_w^k = \frac{w!}{k!(w-k)!}$
Compute performance index for each combination (I_a)
Sort locations based on the I_a

Figure 3. Flow chart of the proposed approach for the monitoring location design

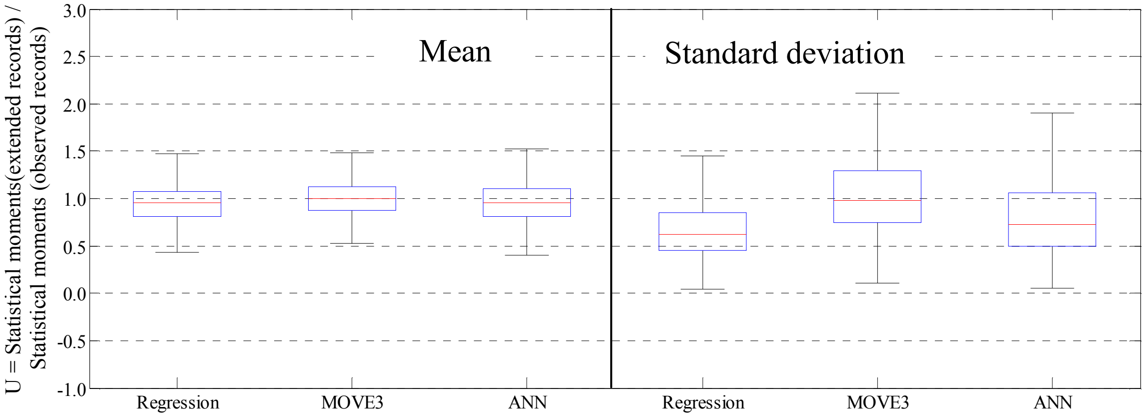


Figure 4. Box plots of the ratio *U* for the mean and standard deviation

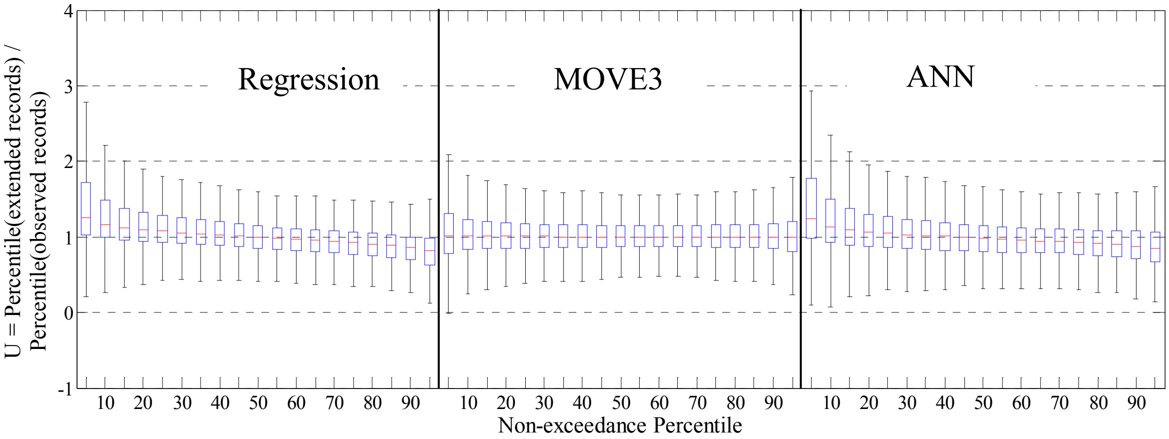


Figure 5. Box plots of the ratio *U* for different percentiles

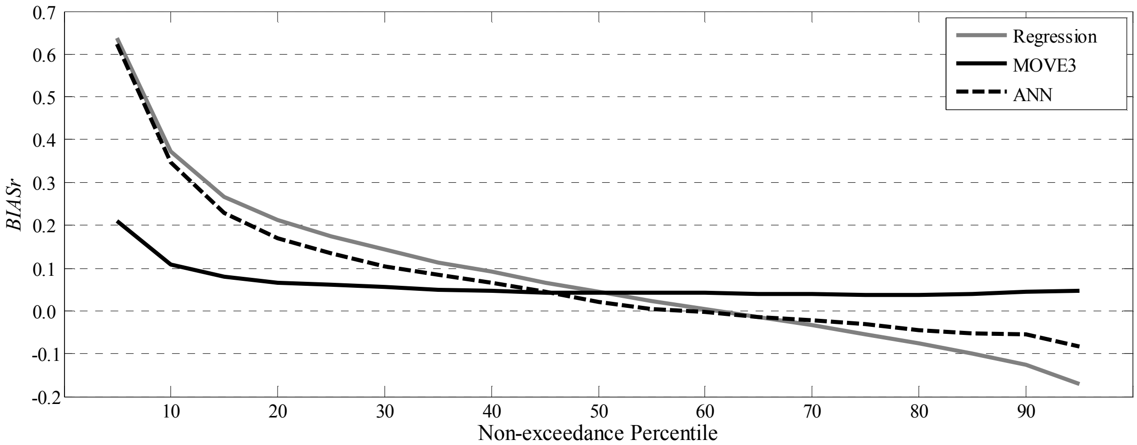


Figure 6. *BIASr* of the three extension techniques for the estimation of percentiles

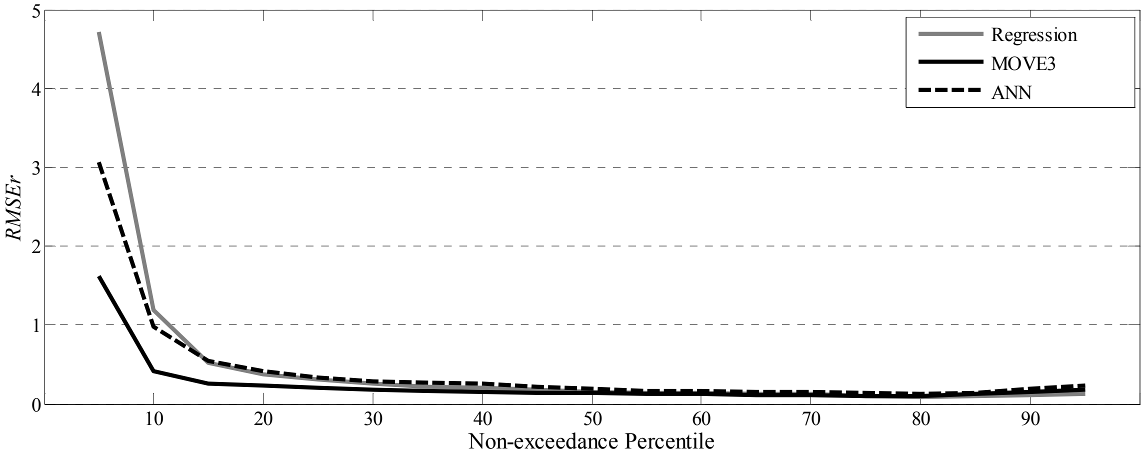


Figure 7. *RMSEr* of the three extension techniques for the estimation of percentiles

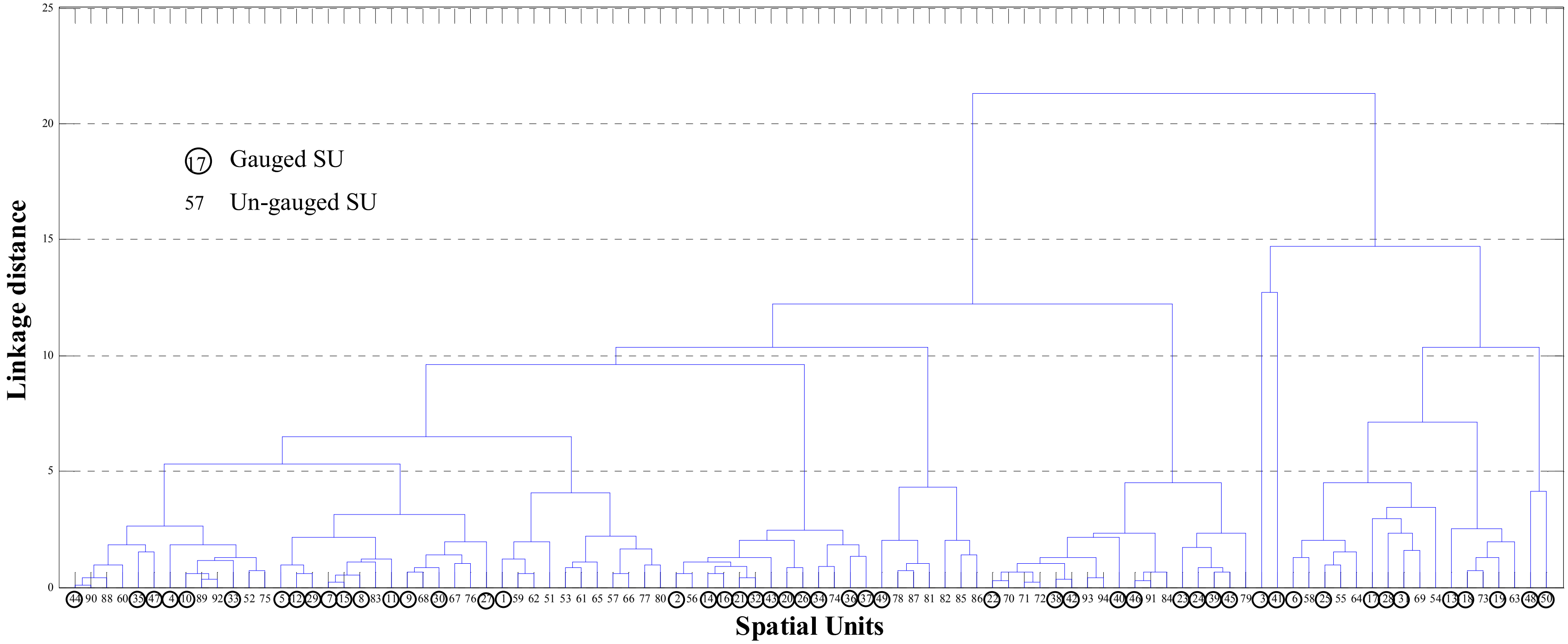


Figure 8. Spatial Units hierarchical clustering dendrogram

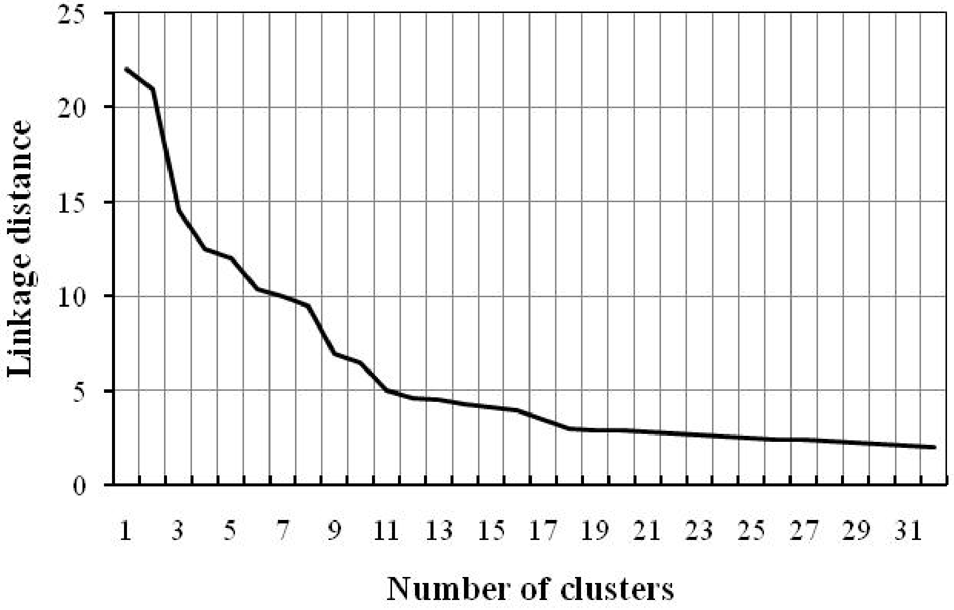


Figure 9. Number of clusters as a function of clusters linkage distance

Article V. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis

Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis

B. Khalil[1,2], T. B.M.J. Ouarda[2], and A. St-Hilaire[2]

[1] Irrigation and hydraulics department, Faculty of Engineering, Helwan University, Cairo, Egypt.

[2]Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE, Québec City, Canada.

In press, Journal of Hydrology

Abstract

Two models are developed for the estimation of water quality mean values at ungauged sites. The first model is based on artificial neural networks (ANNs) and the second model is based on canonical correlation analysis (CCA) and ANNs. An ensemble ANN model is developed to establish the functional relationship between water quality mean values and basin attributes. In the CCA-based ANN model, CCA is used to form a canonical attribute space using data from gauged sites. Then, an ANN is applied to identify the functional relationships between water quality mean values and the attributes in the CCA space. The two models are applied to 50 subcatchments in the Nile Delta, Egypt. A jackknife validation procedure is used to evaluate the performance of the two models. The results show that the developed models are useful for estimating the water quality status at ungauged sites. However, the CCA-based ANN model performed better than the ANN in terms of prediction accuracy.

Keywords: water quality; canonical correlation; artificial neural networks; jackknife; ungauged site; prediction.

1 Introduction

Water quality is a term used to describe the chemical, physical and biological characteristics of water with respect to its suitability for a particular use. Water quality is affected by a wide range of natural and anthropogenic influences. Natural processes (hydrological, physical, chemical and biological) may affect the characteristics and concentration of chemical elements and compounds in freshwater. There are also anthropogenic impacts that affect water quality, such as human-induced point and nonpoint pollution sources, introduction of xenobiotics and alteration of water quality due to water use and river engineering projects (Chapman, 1996). The assessment of water resources requires knowledge about the processes affecting both water quantity and water quality (Harmancioglu et al., 1999). To understand the process dynamics of a watershed, a well-designed water quality monitoring network is required.

In general, water quality data are needed to delineate the following (Whitfield, 1988; Harmancioglu et al., 1992): the general nature and trends in water quality; the effects of natural and anthropogenic factors on the general trends in water quality processes; the effectiveness of water pollution control measures; and the level of compliance with established quality standards. They are also needed to assess the general water quality conditions over a wide area and to model water quality processes. Thus, water quality monitoring programs help explain the various processes affecting water quality, as well as provide water managers with the necessary information for decision making (Khalil and Ouarda, 2009).

Normally, information is derived from water quality data at gauging sites. However, in many cases, water quality networks are characterized by their low density and their inconvenient spatial distribution. Even when water quality monitoring networks are well developed, historical water quality data may not be available at the sites of interest. In this study, the feasibility of using water quality data generated from gauging sites and the attributes of the basins being monitored to estimate water quality conditions at ungauged sites is investigated.

In hydrometric networks, regional flood frequency analysis has been widely used to estimate flood quantiles at sites where streamflow records are not available. Regional flood frequency analysis is a practical way of obtaining estimates of flood quantiles at ungauged or insufficiently gauged hydrological sites (Ouarda et al., 2000). A number of regionalization techniques have been developed for this purpose (see e.g., Wiltshire, 1986; Burn, 1990; Cavadias, 1990; Ouarda et al., 2001; Shu and Burn, 2004; Shu and Ouarda, 2007).

The first step in regional flood frequency analysis is the identification of homogeneous regions. This step aims to identify catchments that are hydrologically similar to the target site (ungauged site). Conventionally, homogeneous regions are identified based on geographic or administrative boundaries (Matalas, 1975; Beable and McKerchar, 1982). However, regions identified based on geographic continuity do not necessarily show hydrologic similarity (Cunnane, 1988). Alternatively, the hydrological neighborhood (Ouarda et al., 2001) or region of influence technique (Burn, 1990), in which each site has a potentially unique set of catchments forming the homogeneous region for the site, has received much attention because of its flexibility and effectiveness (Shu and Ouarda, 2007). The Groupe de Recherche en Hydrologie Statistique (GREHYS, 1996) conducted comparison studies, which indicate that the neighborhood technique performs better than the homogeneous region approach for the identification of hydrologically similar catchments.

Canonical correlation analysis (CCA) (Ouarda et al., 2000, 2001) is a frequently used approach for defining hydrological neighbourhoods (Shu and Ouarda, 2007). Originally, it was introduced by Cavadias (1990) for flood quantile estimation. Cavadias (1990) identified homogeneous regions based on visual judgments of clustering patterns. Ouarda et al. (2000) applied the CCA approach to jointly estimate extreme flood peak and volume quantiles in Québec, Canada.

Ouarda et al. (2001) further improved the CCA approach and proposed detailed algorithms for identifying homogeneous regions for gauged and ungauged sites using CCA (Shu and Ouarda, 2007).

Shu and Ouarda (2007) used CCA to construct a canonical physiographic space using the site characteristics from gauged sites. Then, artificial neural network (ANN) models are applied to identify the functional relationships between flood quantiles and the physiographic variables in the CCA space. Shu and Ouarda (2007) compared the original CCA model (Ouarda et al., 2001), the canonical kriging model (Chokmani and Ouarda, 2004), the original ANN models and the CCA-based ANN models for the estimation of flood quantiles at ungauged sites. Results showed that the CCA-based ANN models provide superior estimations compared with the original ANN models. In addition, results showed that the ANN ensemble approaches provide a better generalization ability than the single ANN models. From this work, it was concluded that the CCA-based ensemble model has the best performance among all models in terms of prediction accuracy. CCA was also used by Guillemette et al. (2009) to estimate water temperature at ungauged site. They interpolated maximum monthly temperature using kriging in CCA space, an approach initially developed by Chokmani and Ouarda (2004) for flood quantile estimation. Guillemette et al. (2009) showed that interpolating in CCA space provided better water temperature estimations than kriging in Principal Component Analysis (PCA) space.

In this study, the regional flood frequency analysis approach as applied for the estimation of flow quantiles at ungauged sites is adapted for the estimation of water quality mean values at ungauged sites. In the following section, a description of the study area is provided. In Section 3, the methodology is presented. In Section 4, results are presented and discussed. Finally, conclusions from this work are presented in Section 5.

2 Nile Delta water quality monitoring network

Egypt is a semiarid country with rainfall rarely exceeding 200 mm/year along the North Coast. The rain intensity decreases quickly away from the coastal areas, and scattered showers can hardly be depended upon for agricultural production (Abu-Salama, 2007). According to an agreement between Egypt and Sudan in 1959, the Nile water allocation is 18.5 billion m^3 to Sudan and 55.5 billion m^3 to Egypt (Dijkman, 1993). About 97% of Egypt's water resources are from the Nile River. The distribution of Egypt's share of the Nile's water to its population is near the water poverty threshold and will fall well below this threshold in the years to come (MWRI, 1997). The per capita share of freshwater resources in Egypt is about 800 m^3 per person per year (Abdel-Gawad et al., 2004; Frenken, 2005). It is expected to drop to 450 m^3 per person by the year 2025 (Abdel-Gawad et al., 2004).

One of the solutions applied to stretch limited Egyptian water resources is the reuse of any kind of drainage water (agricultural/industrial/municipal or most often a mixture thereof) for agricultural production. The drainage system in the Nile Delta is composed of 22 catchment areas. Depending on their quality, the effluents are either discharged into the northern lakes or pumped into irrigation canals at 21 sites along the main drains to augment the freshwater supply (DRI-MADWQ, 1998). Numerous programs have been developed in the past to monitor the quality of the Nile water and the agricultural drainage water in the Nile Delta. In 1977, the National Water Research Center (NWRC) started to monitor a few volumetric and qualitative water parameters (predominantly concerning salinity) in some of the main drains in the Nile Delta.

Since 1997, the NWRC has expanded its monitoring activities to include an ever-increasing number of sampling sites and water quality variables. The monitoring program of the Nile Delta drainage system has the following aims: assessing compliance with national standards, estimating mass transport and identifying temporal and spatial trends (NAWQAM, 2001). The Nile Delta drainage system monitoring network (Figure 1) consists of 94 monitoring locations, through which 33 water quality variables are measured on a monthly basis. Of the 94 monitoring locations, 21 locations are located at drainage water reuse locations, and 10 locations in the drainage system main streams are used as check points for assessing the water and salt balance. Thirteen monitoring locations are located at outfalls to the northern lakes and the Mediterranean and 50 monitoring locations are located at tributaries that serve small catchments and deliver water to the main drainage systems.

(Figure 1)

3 Methodology

The description of the methodology is divided into two main parts. The first part describes the data preparation and the second part describes the development of two water quality models designed to estimate the water quality conditions at ungauged sites. However, before data preparation, the Nile Delta is divided into subcatchments, which are spatial units (SUs) that drain to only one point on the drainage system. Measuring the water quality at one of these points describes the effect of the SU’s natural and anthropogenic impacts on the water quality conditions.

The data preparation consists of two main steps. In the first step, attributes that explain different natural and anthropogenic effects are identified for each SU. In the second step, water quality indicators are selected from the 33 measured water quality variables. In this step, the water quality variables that better explain the variability in the water quality in the Nile Delta drainage system are selected. These data preparation steps are described in sub-section 3.1.

In the second part, two models are developed to estimate water quality at ungauged sites. The models are based on the functional relationship between SU attributes and the selected water quality variables at gauged SUs. Model development is described in subsection 3.2, along with the evaluation procedure and criteria.

3.1 Data preparation

The data preparation consists of two steps: identifying SU attributes and selecting the water quality variables that best describe the variations in the water quality. The water quality measured at each SU reflects the impact of its unique natural and anthropogenic influences. The attributes are selected to describe the different natural and anthropogenic characteristics within each SU. However, selection is restricted by data availability. The following attributes are identified for each SU:

- Cultivated area (feddan [feddan = 4200 m^2]);

- Average soil salinity (ppm);
- Average soil hydraulic conductivity (m/d);
- Average total annual rainfall (mm/year);
- Drainage system total length (km);
- Average total industrial effluent (m^3/day);
- Wastewater plant total capacity (m^3/day);
- Number of livestock; and
- Average annual applied fertilizers (tons/year).

Although the population density is an important attribute, it is excluded from this study due to limited information and difficulty to identify the population for each SU.

The quality of a water body is usually described by sets of physical, chemical and biological variables that are mutually interrelated. Water quality can be defined in terms of one variable to hundreds of compounds (Khalil et al., 2010).The water quality variables that best explain the variability in the water quality are selected by using principal component analysis (PCA). PCA is one of a number of factor extraction methods. Since Hotelling (1933) introduced PCA, it has been used for data interpretation, pattern recognition, dimensional analysis, and multicollinearity detection. The PCA transforms a set of correlated variables into a smaller set of uncorrelated variates, called principal components (Jobson, 1992). The PCA can be applied to a set of water quality variables to identify variables that form coherent subsets that are relatively independent one another (Khalil and Ouarda, 2009). The principal components summarize the patterns of correlation among observed variables. Variables that are correlated with one another, but largely independent of other subsets, are linearly combined into one component (Khalil and Ouarda, 2009). The first component explains most of the variance in the data, and each successive component explains less of the variance (Tabachnick and Fidell, 1996).

Interpretation of the components is based on the variables related to them and their significance to physical processes. Correlations between variables and principal components are called component loadings. The component loading matrix obtained from PCA reflects the characteristics of the extraction procedure, which maximizes the variance in each successive component. Once the loading matrix is extracted, rotation can take place. Rotation is ordinarily used after extraction to maximize high correlations and minimize low ones. Numerous methods of rotation are available; the one applied in this study is varimax (variance maximizing procedure). Water quality variables most related to the main components (the ones that explain most of the data variance) are selected to represent water quality at the gauged SU. One water quality variable is selected as an indicator from each component (the variable with the highest absolute factor loading). This step is followed by calculating the mean values of the selected variables over the record period at the gauged SUs. These values are used as the output in the developed models.

3.2 Water quality models

The relationship between SU attributes and water quality is established using two different models. The first model is based mainly on ANNs. The second model uses CCA and ANN. The following subsections describe these two models as well as the model evaluation procedure and criteria.

3.2.1 ANN Ensemble (EANN) model

ANN is a computational model that attempts to imitate the way the human brain biological neural networks works. The ANN can be used to model complex and nonlinear relationships or to find patterns in data. Among the various types of ANNs, multilayer perceptrons (MLPs), originally proposed by Rumelhart and McClelland (1986), are the most commonly used and well-researched class of ANNs (Ouarda and Shu, 2009). This type of ANN implements a feed-forward supervised paradigm (Shu and Ouarda, 2007). A MLP consists of an input layer, which receives the values of the input variables, an output layer, which provides the model output, and one or more hidden layers. Nodes in each layer are interconnected through weighted acyclic arcs from each preceding layer to the following, without lateral or feedback connections (Shu and Ouarda, 2007).

Several studies have shown that the performance of a single ANN can be improved by using ensemble techniques (Sharkey, 1999; Dietterich, 2000; Shu and Burn, 2004; Ouarda and Shu, 2009; Zaier, et al., 2010). The ANN ensemble is a group of ANNs that are trained for the same problem, where results obtained by these ANNs are combined to produce the ANN ensemble output. As described by Merz (1998), the construction of ANN ensembles consists of two main tasks: (1) the generation of the component ANN constructing the ensemble; and (2) the combination of the multiple outputs from the ANN components to produce the ANN ensemble output. Several methods were proposed in the literature to generate ensemble ANNs include: (i) manipulating the set of initial random weights, (ii) using different network topology, (iii) training component networks using different training algorithms, and (iv) manipulating the training set

(Sharkey, 1999; Shu and Ouarda, 2007). The most frequently used techniques for the manipulation of the training data set are bagging (Breiman, 1996) and boosting (Schapire, 1990) techniques. The most frequently used approaches to combine outputs from ANN components are averaging and stacked generalization (Wolpert, 1992; Shu and Ouarda, 2007). In this paper, the ANN ensemble is used to establish the functional relationship between the SU attributes and the water quality mean values at gauged SUs. The abbreviation EANN stands for an ANN ensemble and will be used in the remainder of the paper to represent this model.

For each of the EANN components, a MLP having one input layer, one hidden layer and one output layer is used. Inputs are the SU attributes and outputs are the mean values of the selected water quality indicators. For the nodes in the hidden layer, the tan-sigmoid transfer function is used. The use of a nonlinear transfer function extends the nonlinear approximation ability of the ANN (Shu and Ouarda, 2007). For the output nodes, a linear transfer function is used. A linear transfer function for the output nodes has the advantage of potentially unbounded outputs (Shu and Burn, 2004).

Determining the number of neurons in the hidden layer is an important task when designing an ANN (Shu and Ouarda, 2007). Too many hidden nodes may lead to the problem of overfitting. Too few nodes in the hidden layer may cause the problem of underfitting. As a rule of thumb, the number of nodes in the hidden layer should be less than twice the input layer size (Shu and Ouarda, 2007). In this study, a sensitivity analysis is performed to identify the optimal number of hidden nodes. By varying the number of hidden neurons from three to fifteen, ANNs with seven hidden neurons are identified as providing the most accurate estimation when they are applied to estimate the mean values for the selected water quality indicators. Thus, seven hidden neurons are finally used in the hidden layer.

To identify the size of an ensemble, Hansen and Salamon (1990) and Agrafiotis et al. (2002) have suggested that using ten networks can achieve a significant reduction in classification error. Opitz and Maclin (1999) have shown that, when the ensemble size increases to ten or fifteen, the generalization ability of the ensemble can be noticeably improved. Shu and Burn (2004) have suggested that a network size of ten is necessary to attain sufficient generalization ability. The preliminary analysis of Shu and Ouarda (2007) showed that the estimation error of flood quantiles gradually decreases when the ensemble size increases from 5 to 14, while with a further increase of the ensemble size to 20, no improvement in the estimation error is observed. Different ensemble sizes ranging from 5 to 20 were applied in this study. The results indicated that estimation error gradually decreases when the ensemble size increases to 15. Beyond a size of 15, virtually no improvement in the estimation error is observed. Thus, an ensemble size of 15 is used in the present paper. Following Shu and Ouarda (2007) and Ouarda and Shu (2009), the bagging procedure is selected to generate the individual member networks, and simple averaging is used to combine the outputs from each individual ANN.

3.2.2 EANN-CCA model

In the second model, the EANN model in the CCA space is used to establish the functional relationship between water quality mean values and SU attributes at gauged SUs. The CCA is used to form a canonical attribute space using the SU attributes at gauged SUs. The ANN ensemble models are applied to identify the functional relationships between water quality mean values and SU attributes in the CCA space. The abbreviation EANN-CCA will be used in the remainder of the paper to represent this model.

The goal of CCA is to analyze the linear relationships between two sets of random variables. Consider *X* and *Y*, two random sets of variables, CCA can be defined as the problem of finding linear combinations $W = \alpha' X$ and $V = \beta' Y$ that are maximally correlated (where *W* and *V* are the canonical variables, α and β are the eigenvectors). The maximum number of canonical variable pairs is equal to the smallest dimensionality of the two variables *X* and *Y* (Ouarda et al., 2001). Let C be the total covariance matrix of variables *X* and *Y* defined as:

$$C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \qquad (1)$$

where C_{XX} is the *X* covariance matrix, C_{YY} is the *Y* covariance matrix, and $C_{XY} = C_{YX}{}^{T}$ is the between-sets covariance matrix. The correlation between *W* and *V* can then be calculated as (Ouarda et al., 2001):

$$\rho = \frac{\alpha' \, C_{XY} \, \beta}{\sqrt{\alpha' \, C_{XX} \, \alpha \, \beta' \, C_{YY} \, \beta}} \qquad (2)$$

The canonical correlations between *X* and *Y* can be found by solving the eigenvalues equations (Tabachnick and Fidell, 1996):

$$\begin{cases} C_{XX}{}^{-1}\, C_{XY}\, C_{YY}{}^{-1}\, C_{YX}\, \alpha = \rho^2 \alpha \\ C_{YY}{}^{-1}\, C_{YX}\, C_{XX}{}^{-1}\, C_{XY}\, \beta = \rho^2 \beta \end{cases} \qquad (3)$$

where the eigenvalues ρ^2 are the squared canonical correlation. The number of non-zero solutions to these equations are limited to the smallest dimensionality of *X* and *Y* (Tabachnick and Fidell, 1996). CCA requires all variables to be transformed for normality and standardized. In this study, all variables are transformed using a Box-Cox transformation, and standardized prior to CCA. Details concerning CCA are available in reference text books (e.g. Muirhead, 1982; Tabachnick and Fidell, 1996).

Suppose a set of SU attributes, *X*, and water quality mean values, *Y*, associated with each gauged SU. Using CCA, canonical variables *W* and *V* can be obtained as a linear combination of *X* and *Y*, respectively. The coefficients used for the combination are computed so that the correlation between the variables *W* and *V* is maximized. The goal of the EANN model is to approximate the functional relationship between the canonical attribute variables *W* and the water quality mean values *Y*, which act as the input and output of an EANN, respectively. To achieve this goal, the attributes and the water quality records available at gauged SUs are used to calibrate the CCA and to train the EANN. Knowing the CCA combination coefficients, the attributes variable for an ungauged site (*Xu*) can be easily projected into the CCA space to obtain the canonical attributes variable in the CCA space (*Wu*). Applying the projected attributes data to the EANN input layer, estimation can be obtained directly from the output layer.

The EANN models in the CCA space have the same structure, transfer function, number of neurons in the hidden layer and number of ANN components as defined for the EANN model. The component networks in the EANN-CCA models are also generated using the bagging approach, and the resulting networks are combined using simple averaging.

3.2.3 Evaluation procedure and criteria

A jackknife procedure is used to compare the relative performance of the EANN and EANN-CCA models. In this procedure, the water quality mean values at each gauged SU are temporarily removed, thus the SU is assumed to be ungauged. Then, each model is calibrated using the data of the remaining gauged SUs. Estimates for water quality variable means are obtained for the temporarily removed SU using the calibrated models, and then the estimations are evaluated against mean values calculated from the observed records. Evaluations are conducted using the following five indices: the Nash criterion (*NASH*), the root mean squared error (*RMSE*), the relative root mean squared error (*RMSEr*), the mean bias (*BIAS*) and the relative mean bias (*BIASr*). The metrics are computed according to the following equations:

$$NASH = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (5)$$

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \qquad (6)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \quad (7)$$

$$Biasr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \quad (8)$$

where *n* is the total number of SUs being modeled, y_i is the mean estimation from the observed records for site *i*, $\hat{y}_i$ is the mean estimation obtained from the model for site *i*, and $\bar{y}$ is the mean of the mean estimation from the observed records of the *n* sites.

4 Results

4.1 Data preparation

Dividing the Nile Delta into SUs leads to 94 SUs, of which 50 are gauged and 44 are ungauged. Figure 2 shows the Nile Delta gauged and ungauged SUs. Summary statistics for the identified attributes are presented in Table 1.

(Figure 2) and (Table 1)

PCA is employed to identify the water quality variables that best describe the variation in the water quality. Results show that the first seven principal components have eigenvalues greater than one and explain about 85.08% of the total variance in the original data set. Table 2 shows the percentage of variance explained by each principal component and the rotated correlation coefficients for the highly correlated water quality variables. The values shown are the correlations between the variables and the principal components (loadings). Because of the large number of extracted principal components, 0.7 (70%) is treated as the cutoff value for the correlation coefficients. Thus, any water quality variable with an absolute loading value greater than or equal to 0.7 is considered to be an important variable that contributed to variations in the Nile Delta drainage water quality.

(Table 2)

Based on the PCA results (Table 2), the first principal component represents mineral-related variables (Electric Conductivity (EC), Total Dissolved Solids (TDS), Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Bi-Carbonate (HCO_3), Sulphate (SO_4) and Chloride (Cl)). The second component consists of Total Coliform (TCol), Fecal Coliform (FCol), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Phosphorus (TP) and Water temperature (Temp.), and it is negatively correlated to pH and Dissolved Oxygen (DO). The third component represents turbidity with four highly correlated variables: Total Suspended Sediments (TSS), Total Volatile Solids (TVS), Turbidity (turbid.) and visibility (Visib.) (negatively correlated). The fourth component represents nitrates, Nitrate (NO_3), Ammonium (NH_4) and Total Nitrogen (TN). The fifth to seventh components represent trace elements such as Cadmium (Cd), Copper (Cu), Iron (Fe), Manganese (Mn), Zink (Zn), Lead (Pb) and Nickel (Ni). The highly correlated variables in the first four components are selected as indicators: TDS, BOD, TVS and TN. The summary statistics of these selected variables are presented in Table 1.

The selection of water quality variables is followed by a trend assessment of the annual mean values. The trend assessment is performed using the Mann-Kendall nonparametric trend test. The results indicate that for the four selected water quality variables, at almost all 50 gauged sites, the null hypothesis that there is no trend in the annual mean values can be accepted when the significance level is 0.05. Thus, an overall mean value is calculated for each water quality variable at each of the gauging sites for water quality models calibration and validation.

4.2 Water quality models

For the EANN and EANN-CCA models, results obtained using the jackknife validation method are presented in Table 3. A model can be claimed to produce a perfect estimation if the *NASH* criterion is equal to 1. Normally, a model can be considered accurate if the *NASH* criterion is greater than 0.8. *NASH* values for the EANN-CCA model estimation of the mean values of the four considered water quality variables are higher than those for the EANN model estimation. This indicates that the EANN model in the CCA space is better than the EANN.

(Table 3)

RMSE and *RMSEr* indices provide an assessment of the prediction accuracy on an absolute and relative scale, respectively. The EANN-CCA model performed better than the EANN model according to these two indices. The *BIAS* and *BIASr* indices provide an indication about whether a model tends to overestimate or underestimate. The analysis based on the *BIAS* index suggests that the EANN model performed slightly better than the EANN-CCA model for the estimation of the BOD and TVS mean values. However, the *BIASr* indicates that the EANN-CCA model performed better than the EANN model for the estimation of the mean values of the four considered variables. This indicates that the errors obtained when using the EANN model are more symmetric around zero but show more dispersion than those obtained when using the EANN-CCA model. This result is illustrated by the jackknife plots for the BOD and TVS in Figure 3.

Overall, the EANN-CCA model yielded a significantly better performance than the EANN model according to the *NASH*, *RMSE*, *RMSEr* and *BIASr* indices. These results indicate that applying ANN models in the CCA attributes space can greatly improve the performance of ANN models in the original attribute space. ANNs are nonparametric approaches, which are limited in their ability to extrapolate beyond the range of the observed data used in the training process. The EANN-CCA approach combines both parametric and nonparametric methods, which seems to help the performance of the ANNs. Transformation of inputs through CCA before using EANN seems also to filter some of the noise in the data. The EANN model focuses then on reproducing the true signal rather than the noise. In this sense, combining CCA with EANN helps avoid overfitting. Shu and Ouarda (2007) concluded that the CCA technique is better able to characterize the physiographic space for conducting flood quantile estimation. In agreement with their conclusions, the research results of this paper show that CCA is capable of characterizing the SU attribute space for the estimation of water quality mean values.

Using the jackknife validation procedure, estimates for the selected water quality variables (BOD, TVS, TN and TDS) using the EANN and EANN-CCA models are shown in Figure 3. From Figure 3, it can be observed that the EANN-CCA model tends to provide a better estimation than the EANN model for the estimation of the mean values of the four selected water quality variables. Plots corresponding to the EANN model show more dispersion around the 45 degree line than in plots corresponding to the EANN-CCA model. Both models overestimate the mean values at sites with values lower than 34 mg/l for BOD, 10 mg/l for TVS and 12 mg/l for TN and underestimate the mean values at sites with values higher than 60 mg/l for BOD, 20 mg/l for TVS, 20 mg/l for TN and 4000 mg/l for TDS. However, using the EANN-CCA model, the error and bias for those sites are lower than when using the EANN model. Overestimation of low water quality mean values and underestimation of high mean values is mainly due to the limited number of SUs represent these values. These SUs are underrepresented in this case study. For instance, there are only four SUs in the database that show BOD mean values greater than 60 mg/l and only five SUs that show BOD mean values less than 34 mg/l. Thus less training data is available in the variable space occupied by these small or high values.

(Figure 3)

5 Conclusions

Two water quality models for the estimation of water quality mean values at ungauged sites are presented in this paper: the ANN ensemble (EANN) and ANN ensemble in the canonical space (EANN-CCA) models. The EANN model establishes the functional relationship between water quality mean values and basin attributes in the original space. For the CCA-based EANN model, the CCA is used to project the basin attributes into the canonical attribute space. EANN models are then used to approximate the functional relationship between the water quality mean values and the projected attribute variables. The two models are developed and applied to the data of the case study.

Jackknife validation is used to assess the performance of the two models. Results showed that both models performed acceptably. Although the EANN-CCA model has a better prediction accuracy than the EANN model, the EANN model leads to a less biased estimation of the BOD and TVS mean values. Results indicate that applying EANN models in the CCA attribute space can greatly improve the performance of EANN models in the original attribute space. Thus, a CCA-based EANN model can be used to provide an estimate for water quality mean values at ungauged sites.

Models developed in this paper are designed to estimate the water quality mean values for four selected variables at ungauged sites in the Nile Delta, Egypt. Modification of the methods presented in this paper to account for non-stationary variables and the development of methods which allow the estimation of detailed water quality time series represent important directions of future research.

Acknowledgment: The authors wish to thank Prof. Shaden Abdel-Gawad, president of the National Water Research Center of Egypt, for providing the data used in this study. The financial support provided by Helwan University, Cairo, Egypt, the Canada Research Chair Program and the Natural Sciences and Engineering Research Council of Canada (NSERC) is acknowledged.

6 References

[1]. Abdel-Gawad, S.T., H.M. Kandil and T.M. Sadek (2004). Water scarcity prospects in Egypt 2000-2050, in: Marquina (ed.) Environmental Challenges in the Mediterranean 2000-2050, Dordrecht: Kluwer Academic Publishers, 187 - 203.

[2]. Abu-salama, M.S.M. (2007). Spatial and temporal consolidation of drainage water quality monitoring networks. Ph.D. dissertation, Universität Lüneburg, Fakultät III, Umwelt und Technik, Germany.

[3]. Agrafiotis, D.K., W. Cedeno and V.S. Lobanov (2002). On the use of neural network ensembles in QSAR and QSPR. J. Chem. Inf. Compu. Sci., 42, 903-911.

[4]. Beable, M.E. and A.I. McKerchar (1982). Regional flood estimation in New Zealand, Water Soil Tech. Publ. 20, 139 pp. Ministry of Works and Development, Wellington, New Zealand.

[5]. Breiman, L. (1996). Bagging predictors, Mach. Learn., 24 (2), 123-140.

[6]. Burn, D.H. (1990). An appraisal of the “region of influence” approach to flood frequency analysis, Hydrol. Sci. J., 35, 149-165.

[7]. Cavadias C.S. (1990). The canonical correlation approach to regional flood estimation, IAHS Publ., 191, 171-178.

[8]. Chapman, D. (1996). Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring. Chapman & Hall, London.

[9]. Chokmani, K. and T.B.M.J. Ouarda (2004). Physiographic space-based kriging for regional flood frequency estimation at ungauged sites, Water Resour. Res., 40, W12514, doi:10.1029/2003WR002983.

[10]. Cunnane, C. (1988). Methods and merits of regional flood frequency analysis, Journal of Hydrology, 100, 269-290.

[11]. Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, Mach. Learn., 40 (2), 139-157, doi:10.1923/A:1007607513941.

[12]. Dijkman, J.P.M. (1993). Environmental Action Plan of Egypt, A Working Paper on Water Resources. Directorate of General International Cooperation, Ministry of Foreign Affairs, the Netherlands, 116 - 127.

[13]. DRI (Drainage Research Institute) - MADWQ (1998). Monitoring and analysis of drainage water quality in Egypt, Interim Report, Cairo.

[14]. Frenken, K., (2005). Irrigation in Africa in Figures, Aquastat Survey (2005), Food & Agriculture Org, Rome, Italy, 88 p.

[15]. Groupe de Recherche en Hydrologie Statistique (GREHYS) (1996). Intercomparison of regional flood frequency procedures for Canadian rivers, Journal of Hydrology, 186, 85-103.

[16]. Hansen, L., and P. Salamon (1990). Neural network ensembles, IEEE Trans. Pattern Anal. Mach. Intell., 12, 993-1001, doi:10.1109/34.58871.

[17]. Harmancioglu, N.B., M.N. Alpaslan and V.P. Singh (1992). Design of water quality monitoring networks, in R.N. Chowdhury (ed.), Geomechanics and Water Engineering in Environmental Management, ch. 8, pp. 267-296.

[18]. Harmancioglu, N.B., O. Fistikoglu, S.D. Ozkul, V.P. Singh and M.N. Alpaslan (1999). Water Quality Monitoring Network Design. Kluwer Academic Publishers, Dordrecht, the Netherlands, 290 p.

[19]. Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24 (6), 417 - 441.

[20]. Guillemette, N., A. St-Hilaire, T.B.M.J. Ouarda, N. Bergeron, E. Robichaud, L. Bilodeau. (2009). Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in multivariate space. Journal of hydrology, 364,1-12.

[21]. Jobson, J.D. (1992). Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods. New York, Springer-Verlag, 768 p.

[22]. Khalil, B. and T.B.M.J. Ouarda (2009). Statistical approaches used to assess and redesign surface water quality monitoring networks, Journal of Environmental Monitoring, doi: 10.1039/b909521g., 11, 1915 - 1929.

[23]. Khalil, B., T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana (2010). A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks. Journal of Hydrology, 386, 173-185.

[24]. Matalas, N.C., J.R. Slack and J.R. Wallis (1975). Regional skew in search of a parent, Water Resources Research, 11, 815-826.

[25]. Merz, C. J. (1998). Classification and regression by combining models, Ph.D. thesis, Dep. of Inf. and Comput. Sci., Univ. of Calif, Irvine.

[26]. Muirhead, R.J. (1982). Aspect of Multivariate Statistical Theory, John Wiley, Hoboken, N.J.

[27]. MWRI, Ministry of Water Resources and Irrigation (1997). Review of Egypt's Water Policies, Strengthening the Planning Sector Project, Ministry OF Water Resources and Irrigation, Cairo, Egypt.

[28]. NAWQAM, National Water Quality and Availability Management Project (2001). Evaluation and Design of Egypt National Water Quality Monitoring Network. Report no.: WQ-TE-0110-005-DR, NAWQAM, NWRC, Cairo, Egypt.

[29]. Opitz, D., and R. Maclin (1999). Popular ensemble methods: An empirical study, J. Artif. Intell. Res., 11, 169-198.

[30]. Ouarda, T.B.M.J., C. Girard, G.S. Cavadias and B. Bobée (2001). Regional flood frequency estimation with canonical correlation analysis, J. Hydrol. 254, 157-173.

[31]. Ouarda, T.B.M.J. and C. Shu (2009). Regional low-flow frequency analysis using single and ensemble artificial neural networks, Water Resour. Res., 45, W11428, doi:10.1029/2008WR007196.

[32]. Ouarda, T.B.M.J., M. Haché, P. Bruneau and B. Bobée (2000). Regional flood peak and volume estimation in a northern Canadian basin, Journal of Cold Region Engineering, 14, 176-191.

[33]. Rumelhart, D.E. and J.L. McClelland (Eds.) (1986). Parallel Distributed Processing: Explorations in the Microstructure of Congnition, vol. 1, Foundations, MIT Press, Cambride, Mass.

[34]. Schapire, R.E. (1990). The strength of weak learnability, Mach. Learn., 5 (2), 197-227.

[35]. Sharkey, A.J.C. (Ed.) (1999). Combining Artificial Neural Nets: Ensemble and Modular Multi-net Systems, Springer, London.

[36]. Shu, C. and D.H. Burn (2004). Artificial neural network ensembles and their application in pooled flood frequency analysis, Water Resources Research, 40, W09301, doi: 10.1029/2003WR002816.

[37]. Shu, C. and T.B.M.J Ouarda (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43, W07438, doi:10.1029/2006WR005142.

[38]. Tabachnick, B.G. and L.S. Fidell (1996). Using Multivariate Statistics. Allyn and Bacon, Boston, London, 879 p.

[39]. Whitfield, P. (1988). Goals and data collection designs for water quality monitoring, Water Resources Bulletin, AWRA, 24 (4), 775-780.

[40]. Wiltshire, S.E. (1986). Regional flood frequency analysis I: Homogeneity stations, Hydrol. Sci. J., 31, 321-333.

[41]. Wolpert, D.H. (1992). Stacked generalization, Neural Networks, 5, 241-259, doi: 10.1016/S0893-6080(05)80023-1.

[42]. Zaier, I., C. Shu, T.B.M.J. Ouarda, O. Seidou and F. Chebana (2010). Estimation of ice thickness on lakes using artificial neural networks ensembles. Journal of Hydrology, 380, 330-340.

Table 1. Descriptive Statistics of SU attributes and selected water quality variables

SU Attributes and WQ variables	Minimum	Mean	Maximum	STDV
Cultivated area (fedd)	912.51	63323.11	294852.06	50902.16
Soil Salinity (ppm)	1280.00	2027.57	2560.00	281.77
Soil hydraulic conductivity (m/day)	0.08	0.19	1.00	0.15
Average Rainfall (mm/year)	37.50	87.43	150.00	35.07
Drains total length (km)	18.12	173.25	540.97	115.11
Total Industrial effluent (m^3/day)	0.00	4420.42	123561.00	18314.81
Total WWTP capacity (m^3/day)	0.00	37078.78	1415000.00	200098.72
Livestock (count)	1957.00	62470.27	206552.00	49447.63
Total annual fertilizers applied (tones/year)	444.64	11730.73	31047.36	7887.32
Biochemical Oxygen Demand (mg/l)	26.36	45.80	94.79	13.05
Total Volatile Solids (mg/l)	5.33	13.60	30.82	5.11
Total Nitrogen (mg/l)	6.78	14.45	25.29	3.65
Total Dissolved Solids (mg/l)	698.60	1685.62	5768.84	1010.07

Table 2. PCA loading matrix for the Nile Delta water quality variables

Water Quality Variables	**Principal Components**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Discharge (Q)	0.427	-0.103	0.156	0.105	-0.116	0.094	0.602
Total Coliform (TCol)	-0.180	**0.931**	0.164	0.057	-0.076	0.138	-0.001
Fecal Coliform (FCol)	-0.182	**0.926**	0.160	0.058	-0.080	0.165	0.002
Biochemical Oxygen Demand (BOD)	-0.031	**0.961**	-0.032	0.014	0.111	-0.106	-0.057
Chemical Oxygen Demand (COD)	-0.021	**0.943**	-0.022	0.033	0.111	-0.113	-0.063
Total Suspended Solids (TSS)	0.125	0.012	**0.972**	0.001	0.049	0.003	-0.053
Total Volatile Solids (TVS)	0.115	0.006	**0.974**	0.009	0.047	-0.008	-0.067
Nitrate (NO_3)	0.210	-0.005	-0.015	**0.947**	-0.102	-0.054	0.030
Ammonium (NH_4)	0.134	0.536	-0.042	**0.781**	-0.013	-0.098	-0.034
Total Phosphorus (TP)	-0.162	**0.860**	0.209	0.130	-0.088	0.113	-0.109
Total Nitrogen (TN)	0.241	-0.078	-0.012	**0.954**	-0.050	0.023	0.023
Cadmium (Cd)	-0.183	-0.068	-0.083	-0.028	0.110	-0.129	0.693
Copper (Cu)	0.087	0.114	-0.038	-0.090	-0.641	0.179	-0.036
Iron (Fe)	-0.088	-0.060	0.070	-0.020	0.208	**0.888**	-0.140
Manganese (Mn)	0.136	0.301	0.385	-0.058	-0.170	0.616	0.071
Zinc (Zn)	0.164	-0.085	-0.130	-0.250	0.634	0.354	0.061
Lead (Pb)	0.165	0.226	0.037	0.029	0.569	0.110	0.285
Nickel (Ni)	0.130	0.036	0.007	-0.164	**0.769**	0.051	-0.206
Boron (B)	0.548	-0.042	-0.253	0.118	0.354	-0.040	-0.314
pH	0.331	**-0.700**	0.168	0.367	0.097	-0.124	-0.162
Electric Conductivity (EC)	**0.981**	-0.110	0.093	0.087	0.034	0.009	0.023
Total Dissolved Solids (TDS)	**0.982**	-0.116	0.076	0.085	0.035	0.004	0.016
Calcium (Ca)	**0.940**	-0.144	-0.130	0.130	0.085	-0.051	-0.029
Magnesium (Mg)	**0.969**	-0.116	0.134	0.090	0.001	0.036	0.023
Sodium (Na)	**0.978**	-0.107	0.121	0.067	0.027	0.017	0.033
Potassium (K)	**0.937**	-0.014	0.009	0.049	0.164	-0.013	-0.039
Bi-Carbonate (HCO_3)	**0.913**	-0.009	0.299	0.092	-0.115	0.049	-0.014
Sulphate (SO_4)	**0.935**	-0.176	-0.150	0.145	0.141	-0.061	-0.054
Chloride (Cl)	**0.973**	-0.094	0.147	0.053	0.006	0.029	0.052
Water temperature (Temp.)	0.018	**0.732**	-0.102	0.129	0.213	-0.427	-0.167
Dissolved Oxygen (DO)	0.343	-0.682	-0.197	0.231	0.256	-0.231	-0.288
Turbidity (Turbid.)	0.117	0.024	**0.951**	0.028	-0.014	0.090	0.081
Visibility (Visib.)	0.030	-0.351	**-0.793**	0.103	0.135	-0.272	-0.099
% Variance exp.	31.83	19.14	12.17	8.35	5.59	4.39	3.61

Bold numbers indicate variables with an absolute factor correlation coefficient greater than or equal to 0.7. Shaded cells indicate the factor correlation coefficients for the selected variables.

Table 3. Jackknife validation results for the performance of the two models

Metrics	Variables	EANN	EANN-CCA
NASH	BOD	0.62	0.84
	TVS	0.68	0.78
	TN	0.61	0.72
	TDS	0.80	0.82
RMSE (mg/l)	BOD	7.95	5.15
	TVS	2.85	2.37
	TN	2.26	1.92
	TDS	451.71	424.41
RMSEr (%)	BOD	15.09	8.60
	TVS	31.42	23.42
	TN	21.24	12.87
	TDS	28.35	18.02
BIAS (mg/l)	BOD	-0.02	0.08
	TVS	0.35	0.41
	TN	0.47	0.15
	TDS	127.74	84.36
BIASr (%)	BOD	2.46	1.80
	TVS	8.78	7.44
	TN	6.95	2.72
	TDS	12.31	10.19

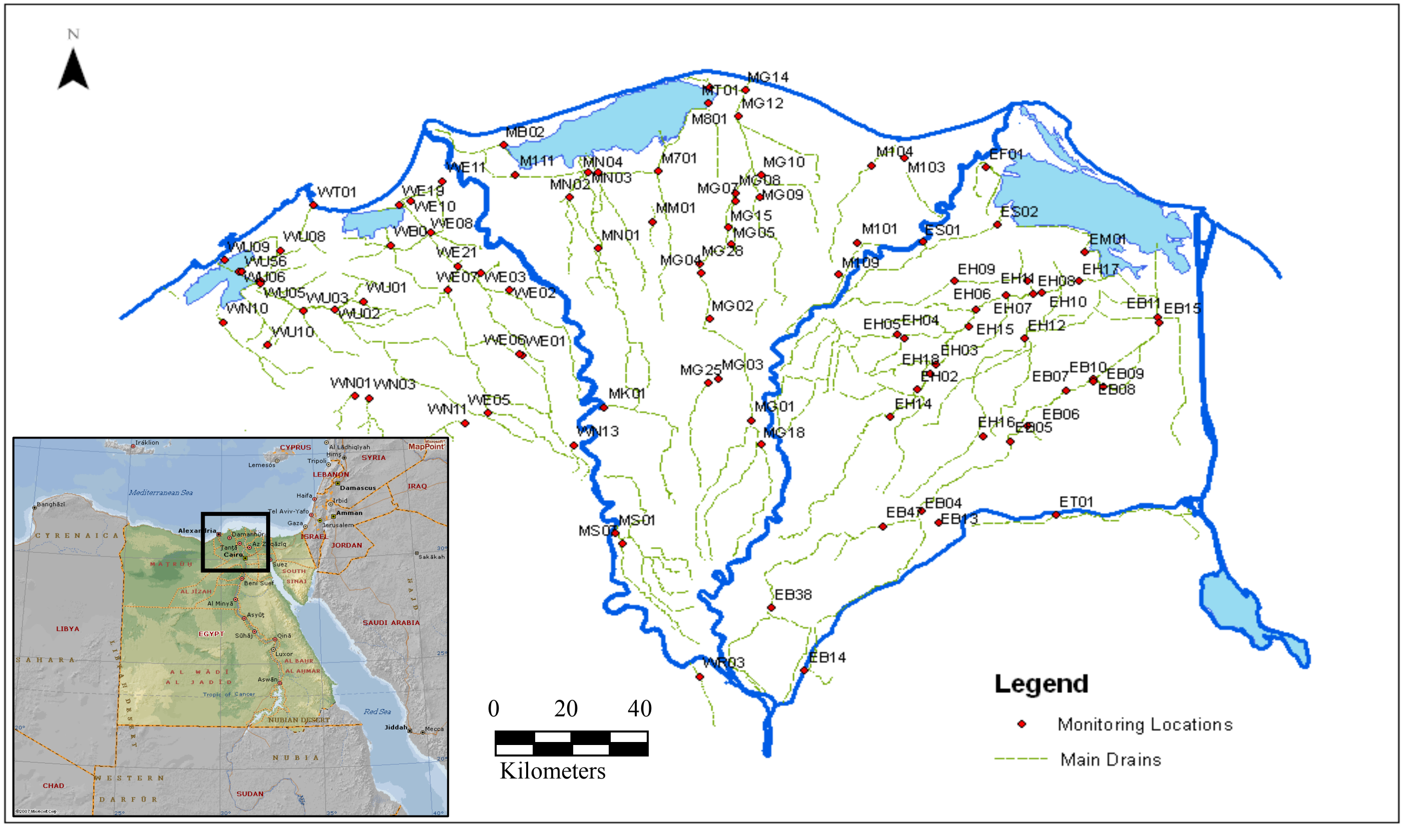


Figure 1. The Nile Delta surface WQM sites (source: NAWQAM, 2001)

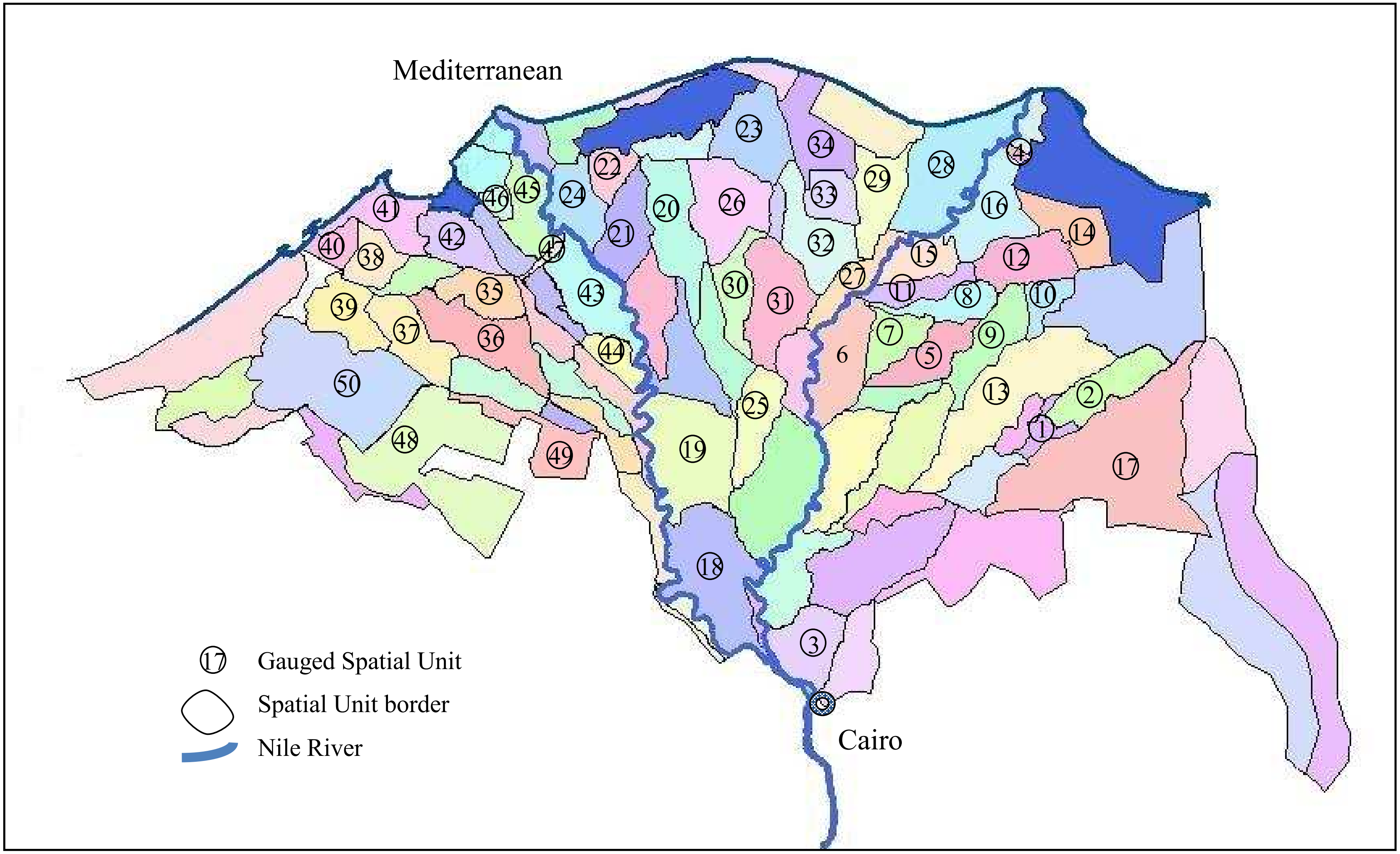


Figure 2. The Nile Delta Spatial Units

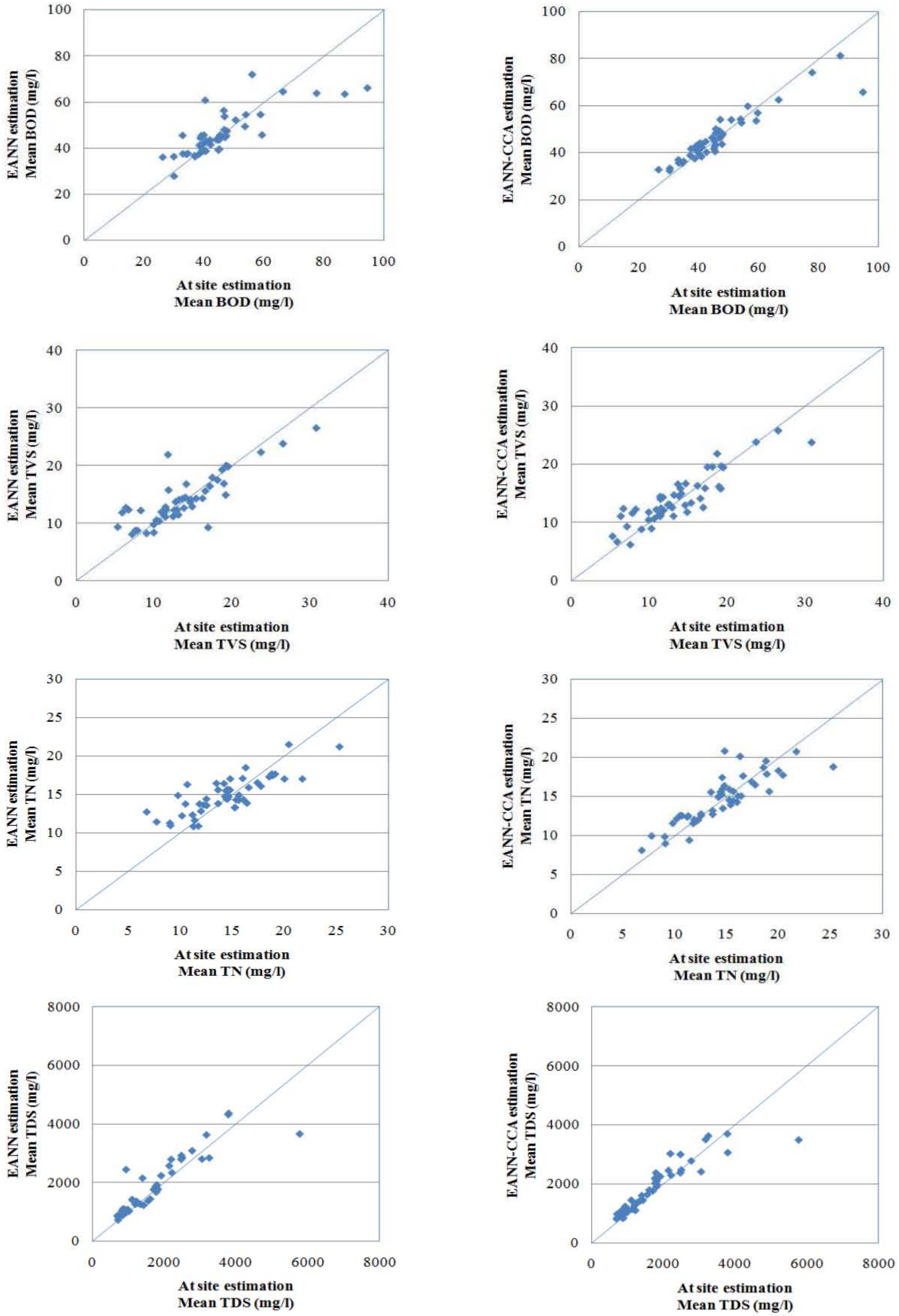
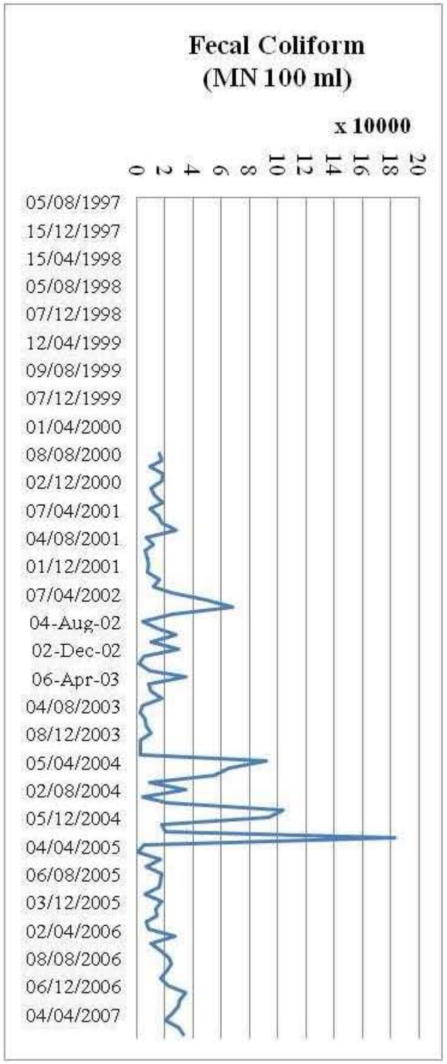
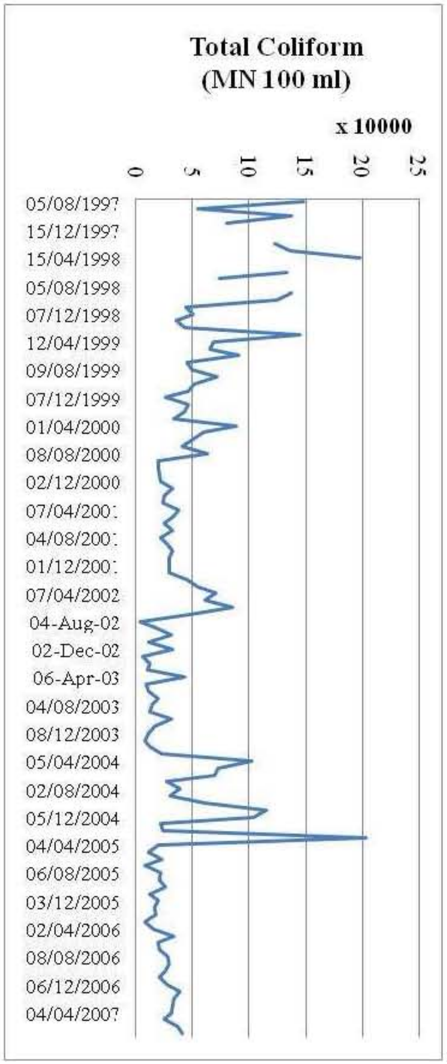


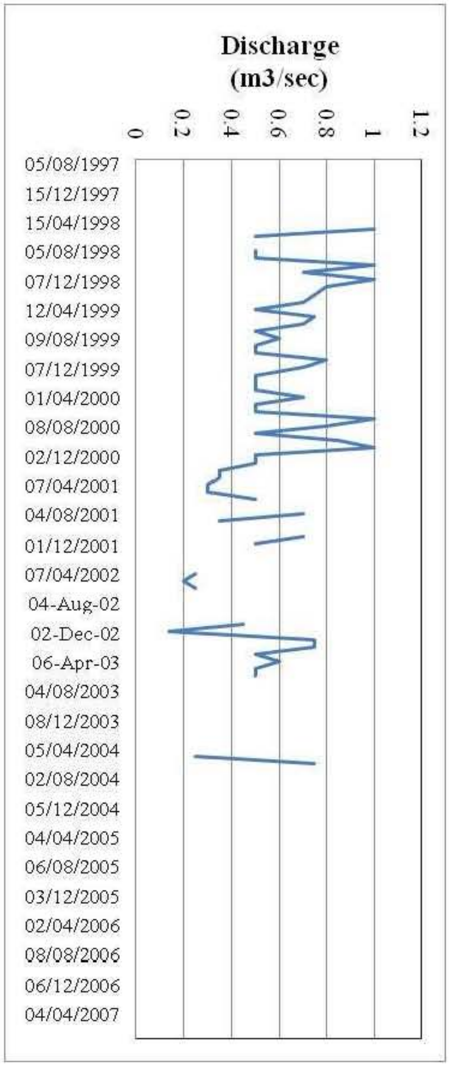
Figure 3. Jackknife estimation using the EANN and EANN-CCA models

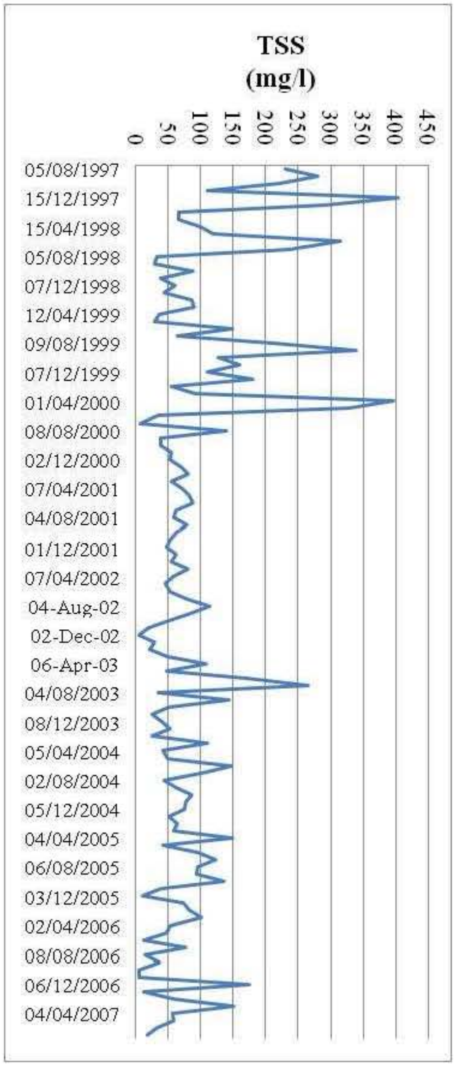
Appendix

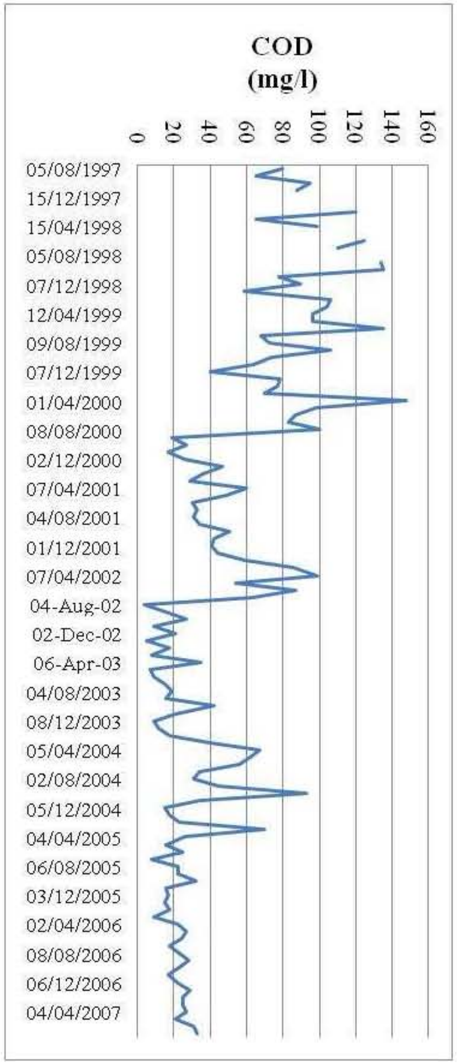
Preliminary analysis
Bani-Ebeid drain EH06, Bahr-Hadus drainage system, Western Delta

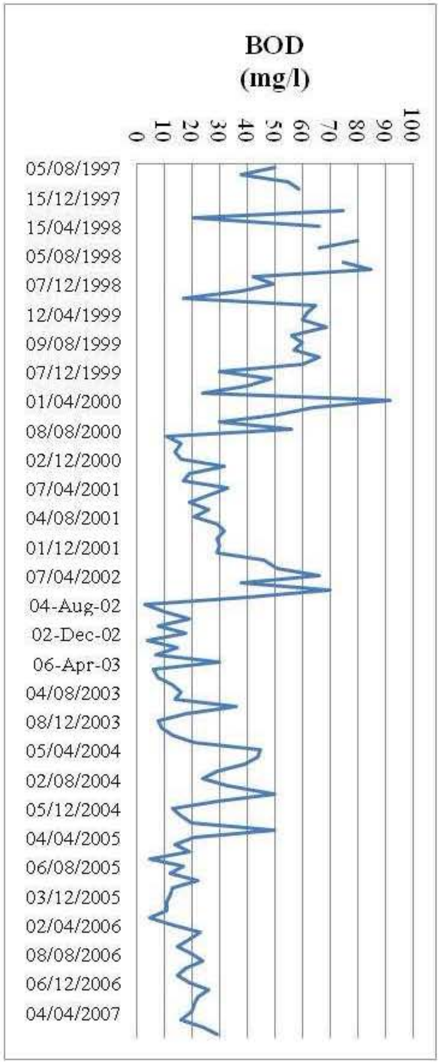

Fecal Coliform
(MN 100 ml)
x 10000
0
2
4
6
8
10
12
14
16
18
20
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

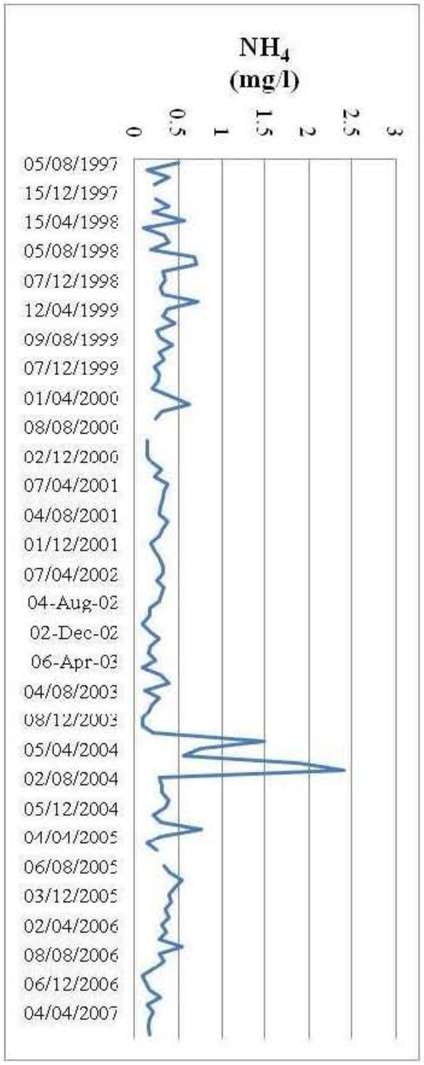

Total Coliform
(MN 100 ml)
x 10000
0
5
10
15
20
25
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

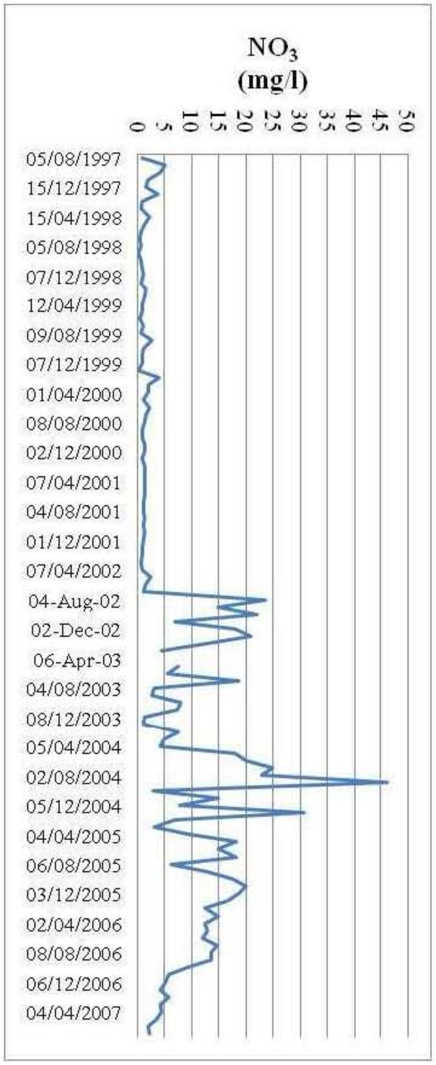

Discharge
(m3/sec)
0
0.2
0.4
0.6
0.8
1
1.2
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

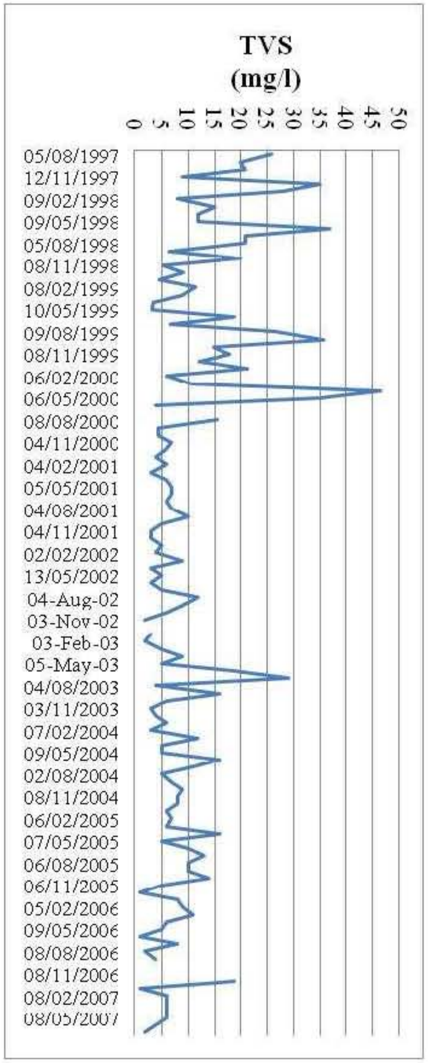

TSS
(mg/l)
0
50
100
150
200
250
300
350
400
450
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

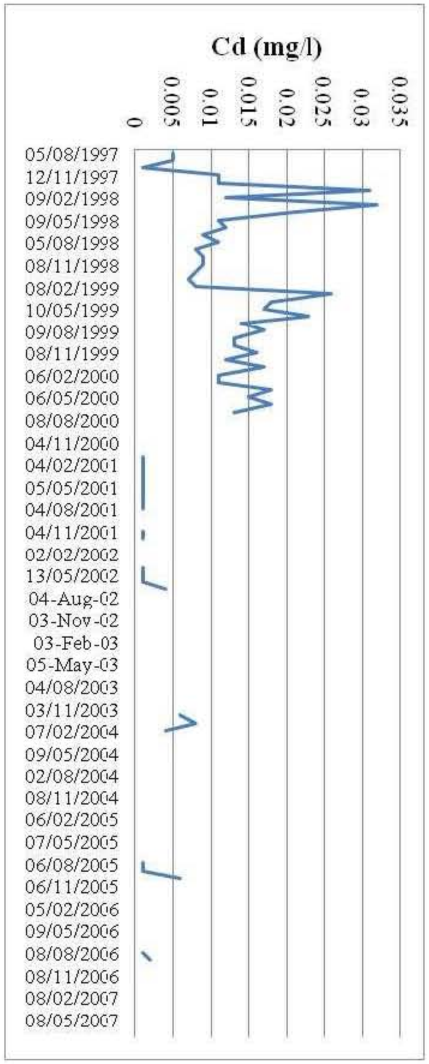

COD
(mg/l)
0
20
40
60
80
100
120
140
160
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

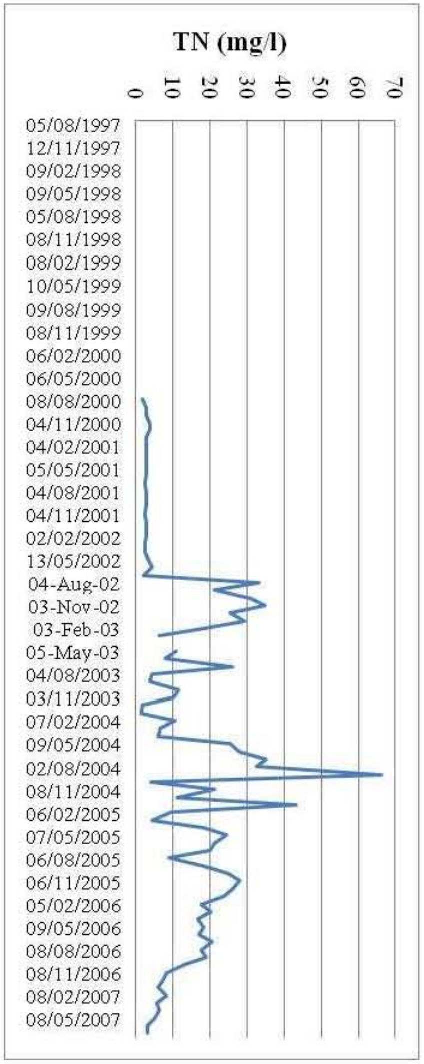

BOD
(mg/l)
0
10
20
30
40
50
60
70
80
90
100
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

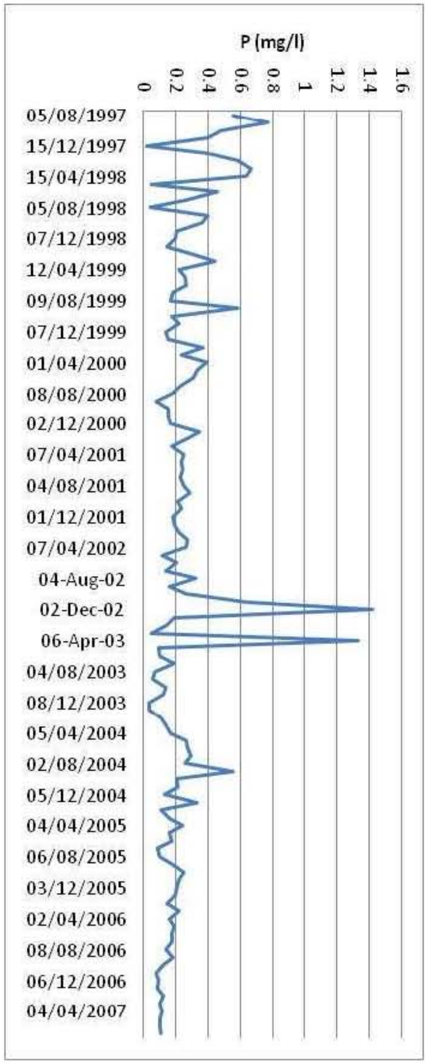

NH$_4$
(mg/l)
0
0.5
1
1.5
2
2.5
3
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

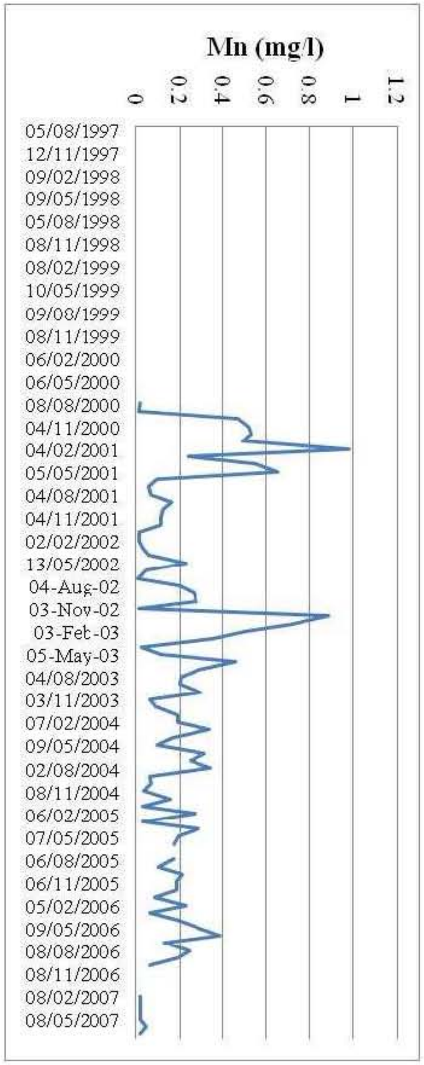

NO$_3$
(mg/l)
0
5
10
15
20
25
30
35
40
45
50
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

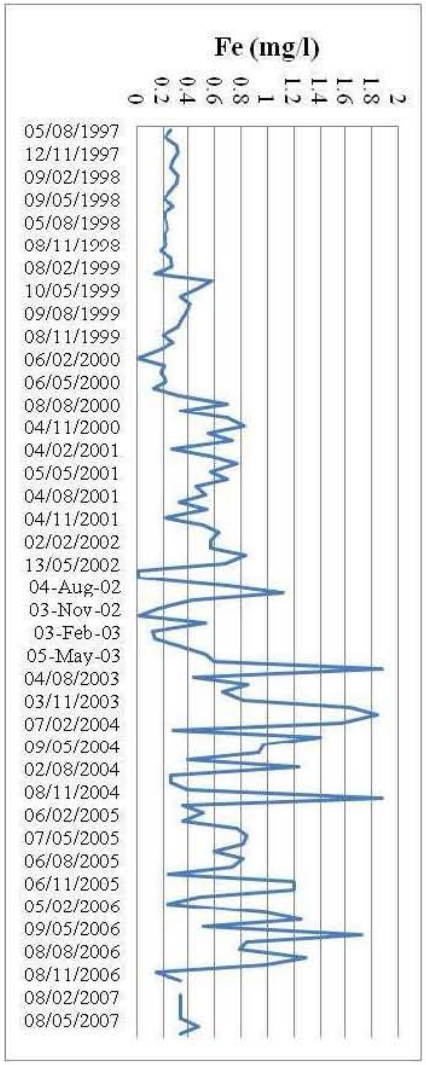

TVS
(mg/l)
0
5
10
15
20
25
30
35
40
45
50
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

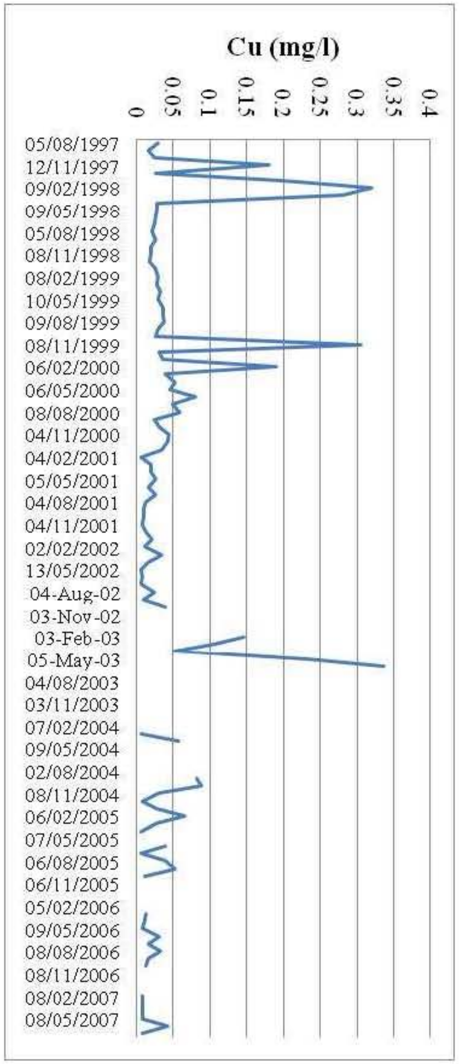

Cd (mg/l)
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

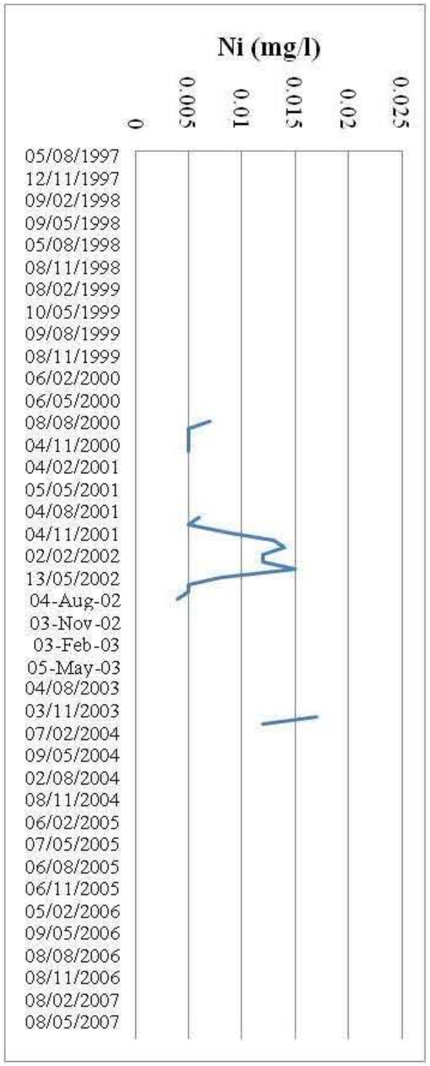

TN (mg/l)
0
10
20
30
40
50
60
70
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

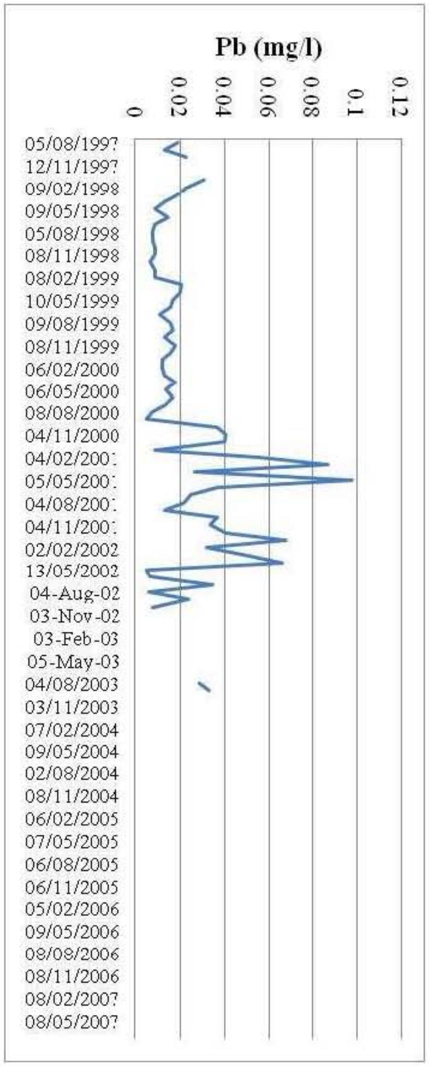

P (mg/l)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
05/08/1997
15/12/1997
15/04/1998
05/08/1998
07/12/1998
12/04/1999
09/08/1999
07/12/1999
01/04/2000
08/08/2000
02/12/2000
07/04/2001
04/08/2001
01/12/2001
07/04/2002
04-Aug-02
02-Dec-02
06-Apr-03
04/08/2003
08/12/2003
05/04/2004
02/08/2004
05/12/2004
04/04/2005
06/08/2005
03/12/2005
02/04/2006
08/08/2006
06/12/2006
04/04/2007

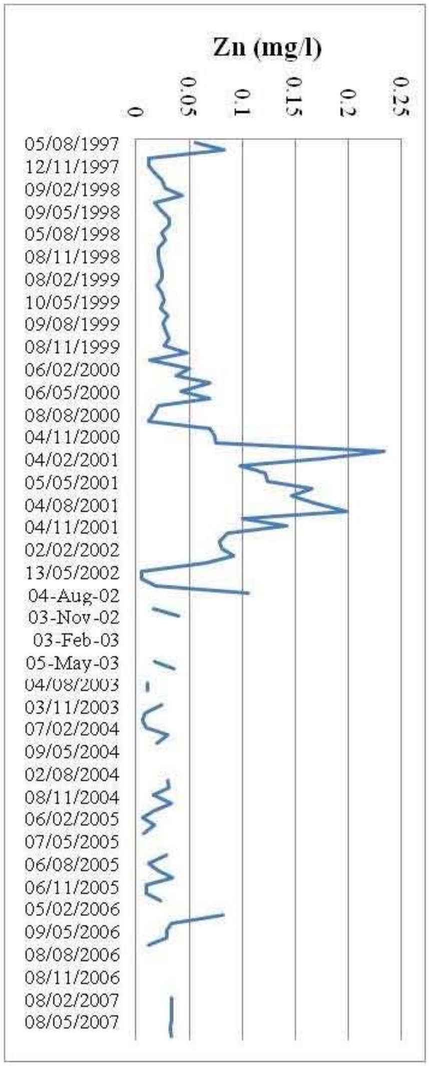

Mn (mg/l)
0
0.2
0.4
0.6
0.8
1
1.2
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

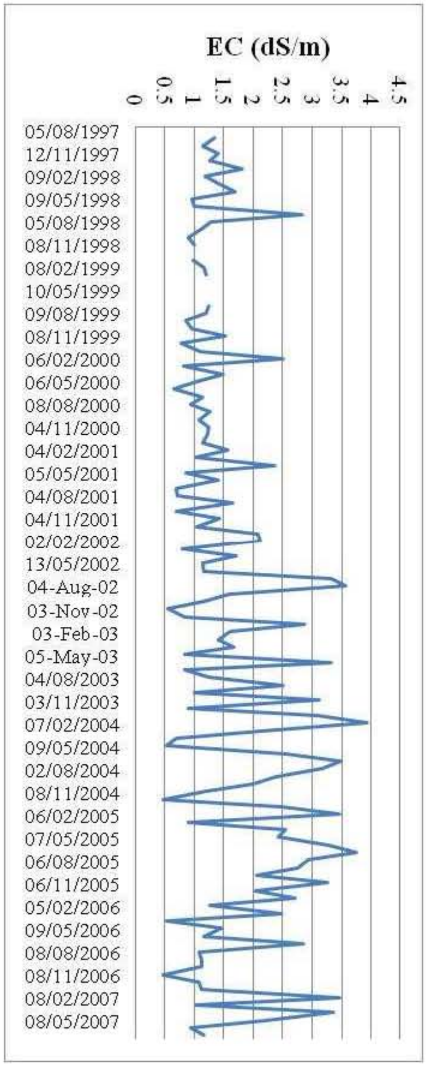

Fe (mg/l)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

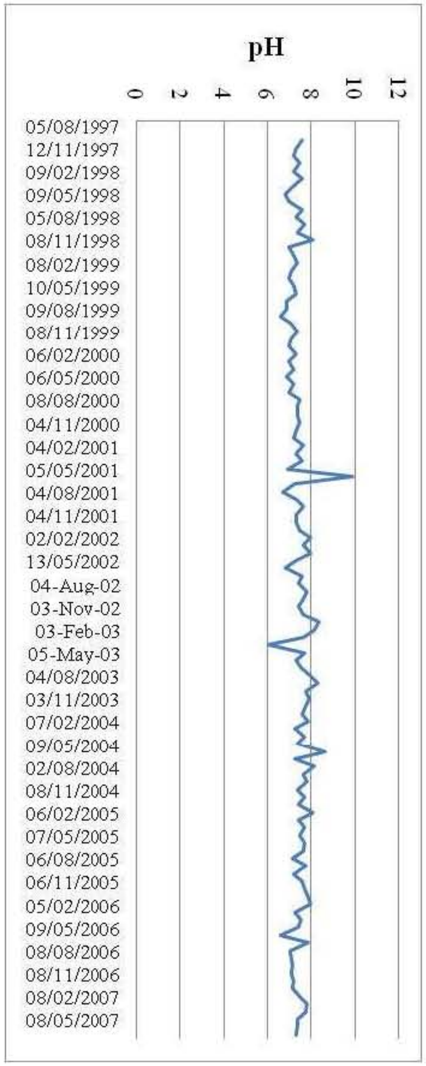

Cu (mg/l)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

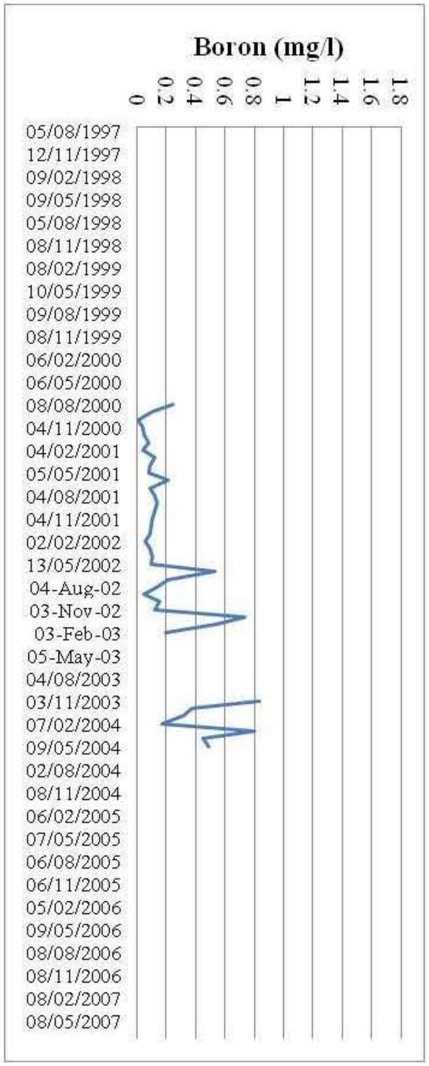

Ni (mg/l)
0
0.005
0.01
0.015
0.02
0.025
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

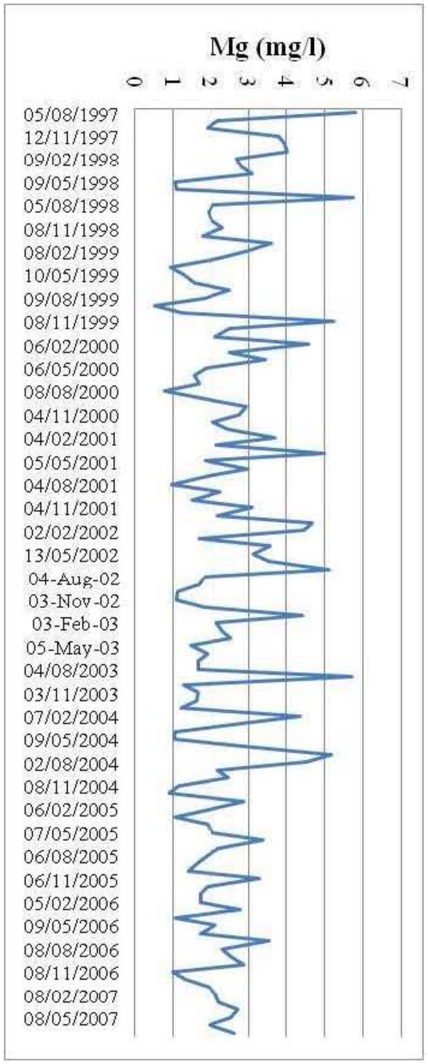

Pb (mg/l)
0
0.02
0.04
0.06
0.08
0.1
0.12
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

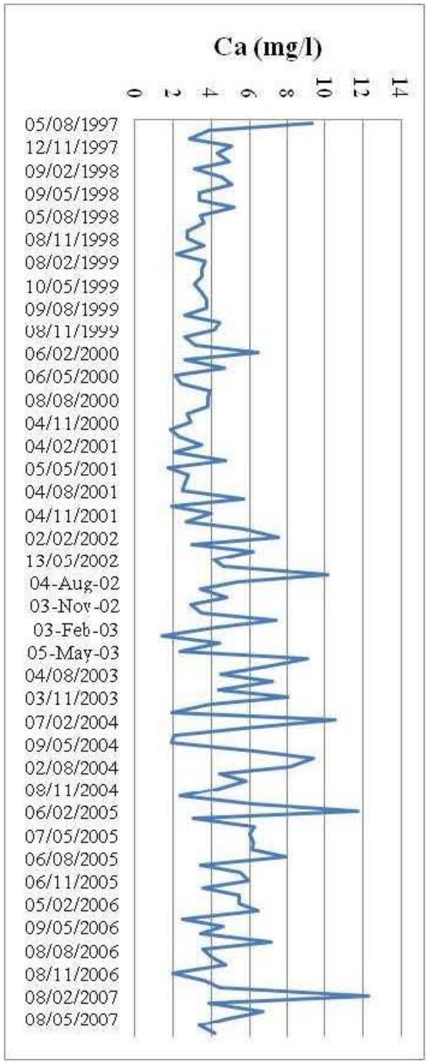

Zn (mg/l)
0
0.05
0.1
0.15
0.2
0.25
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

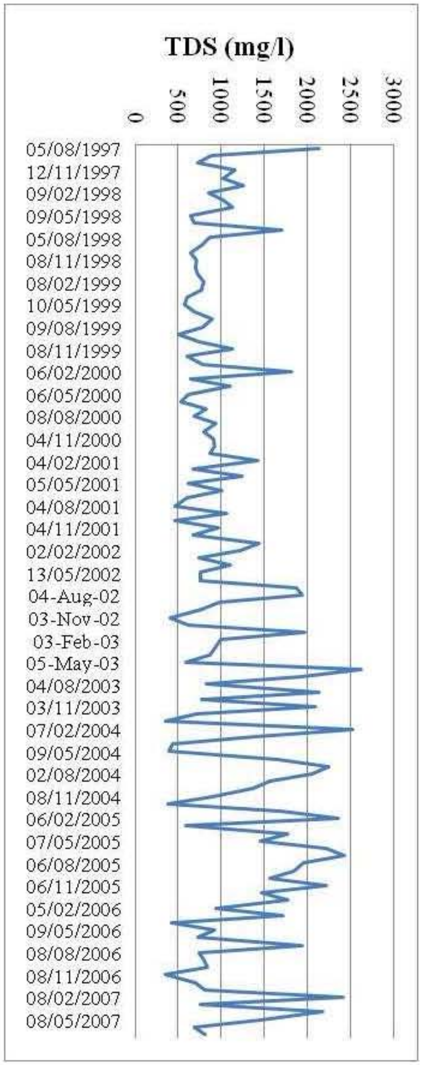

EC (dS/m)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

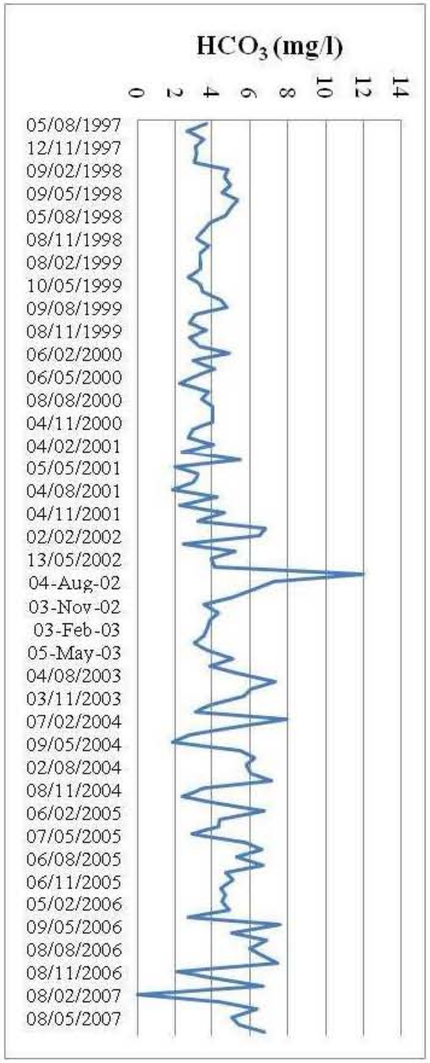

pH
0
2
4
6
8
10
12
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

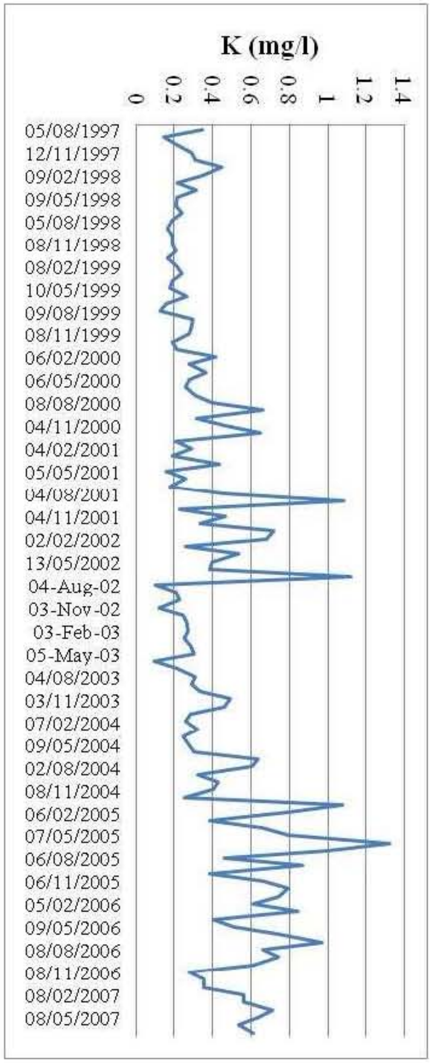

Boron (mg/l)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

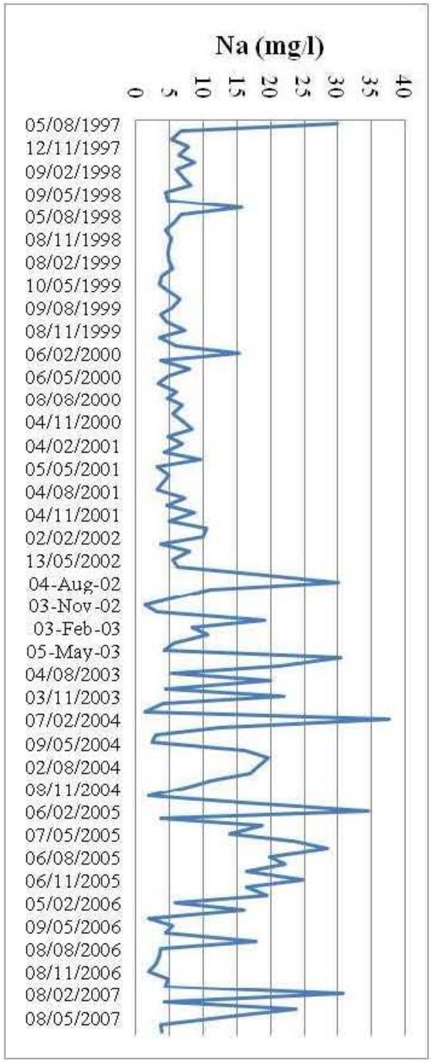

Mg (mg/l)
0
1
2
3
4
5
6
7
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

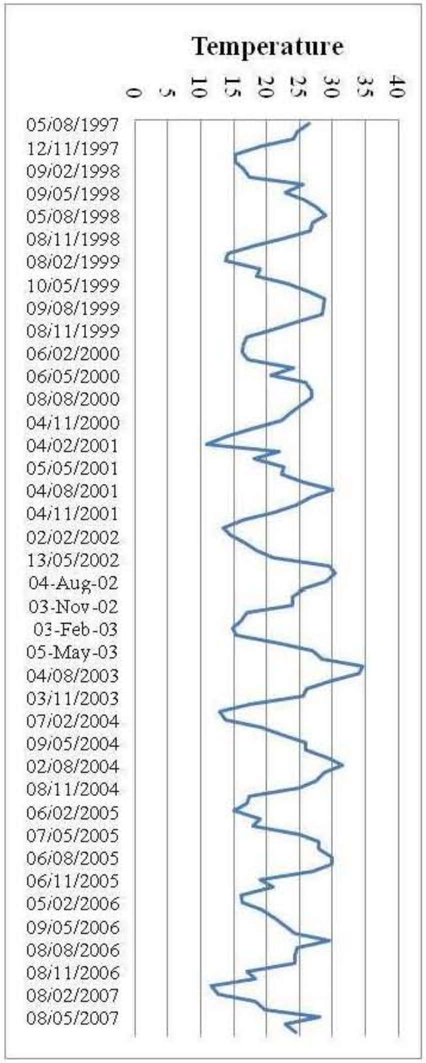

Ca (mg/l)
0
2
4
6
8
10
12
14
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

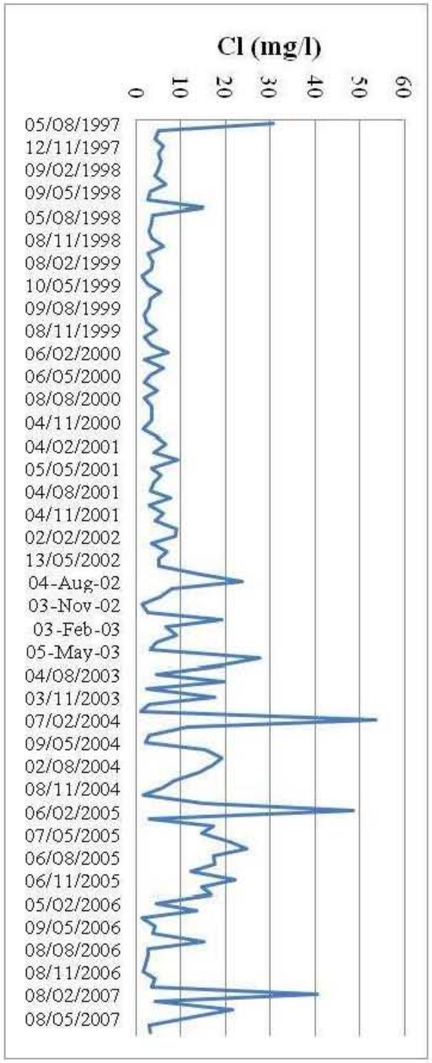

TDS (mg/l)
0
500
1000
1500
2000
2500
3000
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

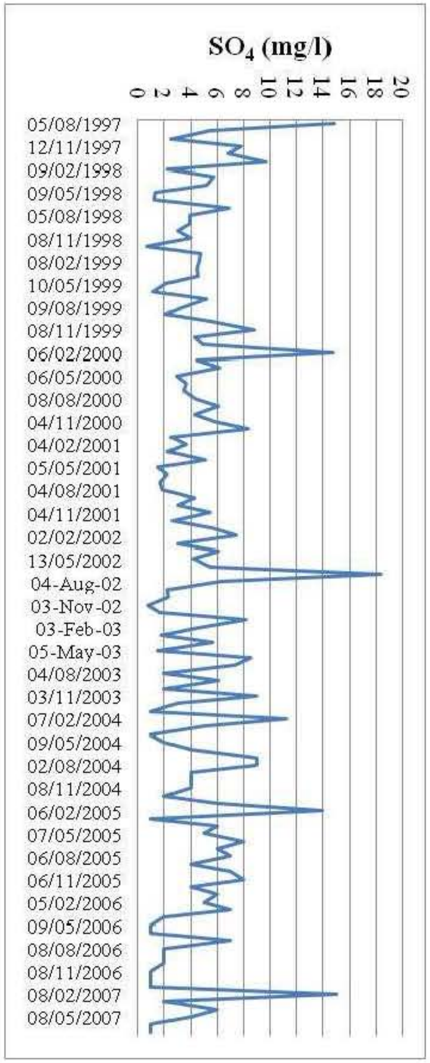

HCO$_3$ (mg/l)
0
2
4
6
8
10
12
14
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

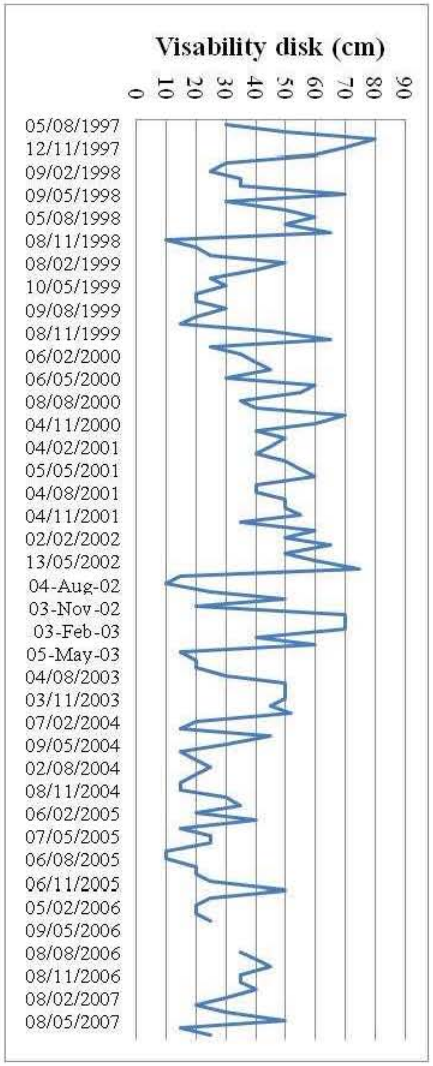

K (mg/l)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

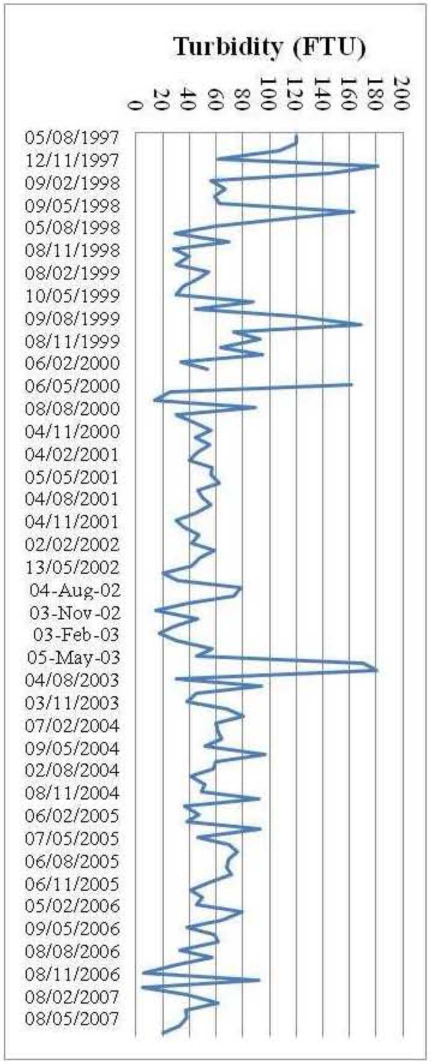

Na (mg/l)
0
5
10
15
20
25
30
35
40
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

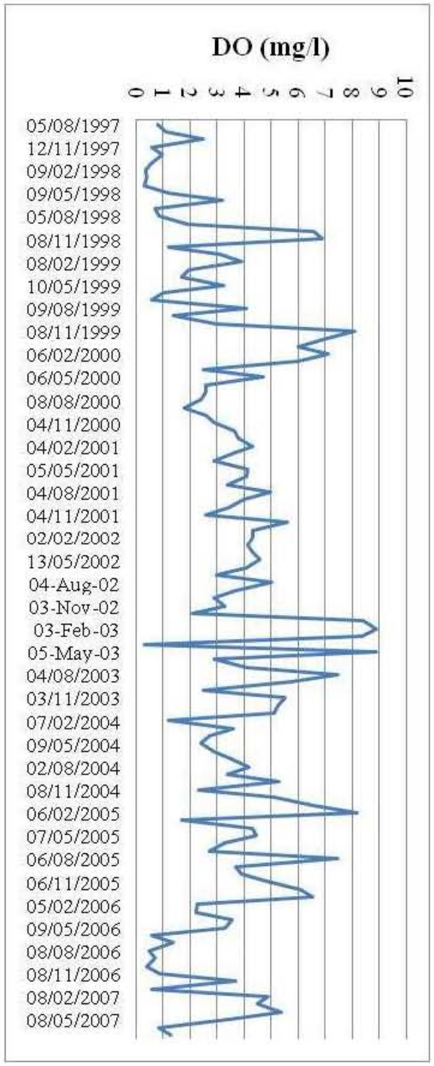

Temperature
0
5
10
15
20
25
30
35
40
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007


Cl (mg/l)
0
10
20
30
40
50
60
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007


SO_4 (mg/l)
0
2
4
6
8
10
12
14
16
18
20
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007


Visability disk (cm)
0
10
20
30
40
50
60
70
80
90
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007


Turbidity (FTU)
0
20
40
60
80
100
120
140
160
180
200
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007


DO (mg/l)
0
1
2
3
4
5
6
7
8
9
10
05/08/1997
12/11/1997
09/02/1998
09/05/1998
05/08/1998
08/11/1998
08/02/1999
10/05/1999
09/08/1999
08/11/1999
06/02/2000
06/05/2000
08/08/2000
04/11/2000
04/02/2001
05/05/2001
04/08/2001
04/11/2001
02/02/2002
13/05/2002
04-Aug-02
03-Nov-02
03-Feb-03
05-May-03
04/08/2003
03/11/2003
07/02/2004
09/05/2004
02/08/2004
08/11/2004
06/02/2005
07/05/2005
06/08/2005
06/11/2005
05/02/2006
09/05/2006
08/08/2006
08/11/2006
08/02/2007
08/05/2007

Descriptive statistics for water quality variables at Arin drain (EH06), western Delta

Variables	No. samples	Average	Median	Standard deviation	Variance	Skewness	Kurtosis
Discharge	58	0.57	0.50	0.21	0.05	0.25	-0.35
Total Colif.	116	46076.9	32952	39470.3	1557901010.3	1.90	3.65
Fecal Colif.	84	22259.3	16567	26534.6	704083475.9	3.64	17.11
BOD	116	31.64	24.00	20.78	431.71	0.86	-0.18
COD	116	48.53	34.00	35.19	1238.44	0.89	-0.18
TSS	120	91.57	63.50	80.18	6429.35	2.00	4.10
TVS	116	10.07	7.00	8.56	73.21	1.91	3.89
NO_3	119	6.59	2.60	7.95	63.26	1.89	4.65
NH_4	116	0.35	0.30	0.30	0.09	4.54	25.47
P	120	0.25	0.20	0.20	0.04	3.26	14.58
TN	83	12.99	8.18	12.07	145.72	1.53	3.43
Cd	61	0.01	0.01	0.01	0.00	0.97	0.67
Cu	102	0.05	0.03	0.07	0.00	2.79	7.26
Fe	119	0.55	0.40	0.41	0.17	1.49	2.09
Mn	81	0.21	0.18	0.20	0.04	1.70	3.34
Zn	102	0.04	0.03	0.05	0.00	2.10	4.41
Pb	67	0.02	0.02	0.02	0.00	1.87	3.92
Ni	25	0.01	0.01	0.00	0.00	1.09	0.75
Boron	39	0.25	0.12	0.31	0.10	2.68	8.82
pH	119	7.46	7.40	0.45	0.20	1.30	7.41
EC	116	1.64	1.25	0.89	0.79	0.93	-0.32
TDS	120	1110.96	878.63	577.16	333111.48	0.94	-0.25
Ca	120	4.52	3.98	2.14	4.57	1.34	2.02
Mg	120	2.43	2.14	1.17	1.37	1.06	0.73
Na	120	9.79	6.13	8.04	64.60	1.49	1.55
K	120	0.41	0.32	0.24	0.06	1.34	1.59
HCO_3	120	4.43	4.24	1.65	2.73	0.93	2.81
SO_4	120	4.71	4.23	3.27	10.71	1.56	3.40
Cl	120	8.55	4.50	9.15	83.68	2.47	7.65
Temprature	120	22.30	22.80	5.36	28.75	-0.07	-0.90
DO	120	3.51	3.40	2.13	4.54	0.55	-0.15
Turbidity	119	60.83	54.61	35.46	1257.10	1.65	2.99
Visability Disk	117	37.75	35.00	17.35	301.10	0.36	-0.81