ASSESSMENT OF THE LAKE WINNIPEG
WATER QUALITY MONITORING NETWORK

ASSESSMENT OF THE LAKE WINNIPEG WATER QUALITY MONITORING NETWORK

RESEARCH REPORT

R-1197

Par

Bahaa Khalil
Dan Beveridge
André St-Hilaire
Taha B.M.J. Ouarda

Institut national de la recherche scientifique
Centre Eau, Terre et Environnement
(INRS-ETE)
490 De la Couronne, Québec, G1K 9A9

March 2010

Suggested citation:

Khalil, B., D. Beveridge, A. St-Hilaire, T.B.M.J. Ouarda. 2010. Assessment of the Lake Winnipeg water quality monitoring network. Research report R-1197, INRS-ETE. vi+182 pages.

© INRS-ETE, 2010

ISBN : 978-2-89146-738-4

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

---

1 INTRODUCTION

Water quality monitoring programs aid in the understanding of various water quality processes and provide water managers with the necessary information for water resources management, in general, and water quality management, in particular (Khalil et al., 2010a).

Water quality monitoring 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 (Khalil and Ouarda, 2009).

The design of monitoring networks is the translation of the monitoring objectives to specify the sampling sites, sampling frequency and variables to be measured (Khalil et al., 2010a). The quality of a water body is usually described by a set of physical, chemical and biological variables that are mutually interrelated. Water quality can be defined in terms of one variable or hundreds of compounds and for multiple usages. This is a very complex issue because there are numerous variables that can represent surface water quality (Sanders et al., 1983; Harmancioglu et al., 1999).

Selection of the initial water quality variables to measure, or the addition of new variables, is primarily based on the monitoring objectives, facilities and available budget. However, statistical approaches are used if the objective is to reduce the number of water quality variables measured.

Another important aspect of water quality monitoring network design is the sampling frequency. Sampling frequency affects not only data utility but also operational cost. With too-frequent sampling, the obtained information is redundant and costly, 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. For instance, the effective sampling method proposed by Lettenmaier (1976) is used to assess the sampling frequency when the monitoring objective is to detect temporal trends.

In large continuous water bodies such as lakes, sampling location and spatial density of sampling sites must be selected such that the spatial variability of water quality is captured without spending resources in acquiring redundant information.

The readers are referred to Khalil and Ouarda (2009) for more detailed information regarding the statistical approaches used to assess and redesign water quality monitoring networks.

The main goal of this study was to provide a first statistical assessment of the the current applied monitoring activities in Lake Winnipeg, Manitoba, Canada. As such, this firs report does not provide an assessment of the complete water quality data base gathered on Lake Winnipeg. The specific objectives of this report are the assessment of:

- the number of isotope stations and their spatial configuration;
- the number of water quality variables to measure; and
- the sampling frequency.

2 DATA

Two water quality datasets were used in this study. The first comprised the $\delta^{18}O$ and $\delta^{2}H$ isotope data collected between 14/09/2009 and 04/10/2009. The samples were collected once at each of 240 locations throughout the lake. As such, the isotope database is perhaps de most dense sampling network on Lake Winnipeg.

The second dataset used in this study were provided by the Water Quality Management Section, Manitoba Water Stewardship, Winnipeg. The database consisted of 168 sampling sites, at which 13 water quality variables were measured on an irregular basis (Table 1). Of 168 sampling locations, 55 were sampled only once between 1992 and 2008. Only seven sampling sites were sampled more than 30 times during this period. These seven sampling sites were selected in this study for assessment of the number of water quality variables to measure and the sampling frequencies of these variables. The selected sampling site codes are as follows: MB05RAS078; MB05RBS003; MB05SAS004; MB05SBS126; MB05SCS005; MB05SHS004; and MB05SHS014. The seven selected sites represent the three main regions of Lake Winnipeg. The sampling sites MB05RAS078 and MB05RBS003 are located in the narrows. Sites MB05SAS004, MB05SBS126 and MB05SCS005 represent the south basin, and sites MB05SHS004 and MB05SHS014 represent the north basin.

Table 1. Water quality variables

Water Quality Variable	**Symbol**	**units**
Alkalinity Total $CaCO_3$	$CaCO_3$	mg/L
Alkalinity Total HCO_3	HCO_3	mg/L
Chloride Dissolved	Cl	mg/L
Colour	Colour	CU
Conductivity (at $25C^o$)	EC	US/cm
Hardness Total $CaCO_3$	H. $CaCO_3$	mg/L
pH	pH	pH units
Sodium Total	Na	mg/L
Sulphate Dissolved	SO_4	mg/L
Water Temperature (Field)	Temp.	Deg C
Total Dissolved Solids (at 180 C^o)	TDS	mg/L at180C
Total Suspended Solids	TSS	mg/L
Turbidity	Turbid.	NTU

3 METHODOLOGY

The methodology consists of several steps. The first step included preliminary analyses, followed by statistical approaches to assess the number of sampling sites, water quality variables and sampling frequency. The objectives of the preliminary analyses were to understand the characteristics of the available water quality data and verify the statistical assumptions required to apply the designed statistical approaches to the assessment and redesign of the monitoring network. The following subsections present the preliminary analyses, sampling sites assessment, water quality variables assessment and the sampling frequency assessment.

3.1 Preliminary analyses

The preliminary analyses included descriptive statistics as well as tests to check normality, autocorrelation and existence of trend. The descriptive statistics included measures of central tendency, variability, skewness and kurtosis. The mean and median were computed to characterise the central tendency. The minimum, maximum, range, standard deviation and variance were computed to characterise the variability. The skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. The kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution; i.e., data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

The box-and-whisker plot (boxplot) is a common five-number summary of the distribution of a variable in graphical form. The box extends from the 25th to the 75th percentile and is crossed by a bar at the median. The H-spread is a term given to the differences between the 25th and the 75th percentile. A step is 1.5 times the H-spread, and the whiskers extend one step. Values that are more extreme, but within two steps, are considered outliers and indicated with an “o”. Values outside are considered extreme outliers and indicated with an asterisk “*”. Boxplots may also be used for spatial or temporal comparison. Comparing the boxplots of the same variable at different locations, or at different consequent years, provides a visual inspection for possible trends, outliers, and data distribution.

The Mann-Kendall nonparametric trend test was employed to test for the existence of a trend. Mann (1945) first suggested using the significance test for Kendall's tau, where the x variable is time, as a test for trend (monotonic change). No assumption of normality is required, but there must be no serial correlation for the resulting p-values to be correct.

The autocorrelation functions (ACFs) for the available water quality variables were also computed. The plot of the autocorrelation function (ACF) against the lag time is called a correlogram. It should be mentioned that trends affect the correlogram function. Therefore, the time series were adjusted (de-trended) before creating the correlogram. To check whether the water quality data followed a normal distribution, the Kolmogorov-Smirnov test for normality was performed.

3.2 Isotope station number and configuration

3.2.1 Kriging

We used kriging to identify water quality stations that contribute the most and least information to prediction maps. Kriging is an interpolation method that predicts the value of a variable at an unmeasured location as a function of observed data from nearby locations. The spatial dependence between points is expressed using a variogram. The variogram captures how the differences in values of points change as the points become further away from each other. Kriging produces an optimal prediction, in the sense that errors and bias are minimized (Journel and Huijbregts 1978).

Before kriging, we will explore the data using techniques such as normal QQ plots, trend analysis, and semivariogram clouds to determine whether the water quality data require transformation or trend removal.

Once the data are prepared for kriging, the variogram model is fit, i.e. we will determine the optimum values of a few variogram parameters. The *sill* is the value at which the variance no longer increases and should be close to the sample variance. The *range* is the distance at which the semivariogram model reaches the sill; beyond this distance, observations appear independent. Anisotropy is the presence of directional dependence, i.e. there is a difference in the correlation between points along different axes. If the data are anisotropic, the direction of strongest correlation defines the direction of the *major range.* The *minor range* is perpendicular to the major range and its length relative to the major range is expressed as the *anisotropy ratio.* Different mathematical functions can be used to model the semivariance, e.g. linear, spherical, gaussian or exponential. The best type of function will be iteratively determined and based on that which yields the smallest nugget effect and Root Mean Square Error (RMSE). The nugget effect represents measurement error or variation at a spatial scale finer than the data.

Once the variogram model is fit, the water quality data will be predicted using ordinary kriging. We will use ordinary kriging which assumes that the mean is constant within the neighborhoud of the point to be estimated, but allows this mean to vary as an unknown constant trend in the whole region.

Kriging is briefly summarized in Appendix 1. It is well-described by Journel and Huijbregts (1978).

3.2.2 Cross-validation

In cross-validation, data are divided into a calibration set, which is used for parameter estimation, and a validation set, with which the results of the calibration set are validated. It is typically used to assess how the results of an analysis based on one dataset will generalize to a different data set. We used it to identify water quality stations that are particularly useful or not useful in predicting water quality.

We used the leave-one-out cross-validation procedure in which one station at a time will be left out as validation data while a kriging map is created from the remaining, i.e. *n*-1 data, where *n*=the number of water quality stations, e.g. *n*=237 in the case of the isotope data. This process will be repeated *n* times.

Each time the training process is repeated, the residual and the kriging variance will be calculated. The *residual* is the difference between the value that the kriging model predicts at the location of the removed station and the observed value. A large residual likely indicates that the missing station is important to prediction. Conversely, a small residual likely indicates that there is significant redundancy between the missing station and its neighbours and that that location can be predicted with less cost. The kriging variance, also known as the kriging error, is by definition minimized in a kriging model, so when a removing a station produced a kriging map with high variance, that station is likely important to prediction. Conversely, when removing a station yields a kriging map with relatively low variance, it is likely redundant with its neighbours.

3.2.3 Improving prediction with additional predictors

The kriging method used above only used the relative position and orientation of water quality stations to predict the values of water quality parameters. Other factors influence water quality and are potentially measured directly or indirectly from available data. Two such factors are water depth, i.e. the elevation of the lake bed, and the proximity of a water quality station to the mouth of a contributing river.

Lake bathymetry data are available. Lake bed elevation will be kriged to predict the elevation at water quality stations.

Lake and river hydrography is available. The distance of a water quality station to the mouth of major contributing rivers will be calculated.

These additional variables can potentially improve the quality of prediction. Unfortunately, kriging can only use two predictive variables at a time. Co-kriging allows information of an auxiliary variable (the co-variable) to be used in the prediction of the variable of interest. The co-variable does not need to be measured at the same locations as the target variable but it does need to demonstrate a spatial structure and correlation and covariance over space with the target variable.

To incorporate more than three predictive variables at a time, we have to use Principal Component Analysis (PCA). PCA can compress the information contained in many variables into the small number of axes permitted in kriging. PCA transforms a large number of intra-correlated variables into uncorrelated variables called principal components. In the process, much of the variability in the original dataset is explained in a few e.g. two or three, principal components. PCA is briefly described in Appendix 2.

We conducted PCA on a dataset containing geographic coordinates, elevation, and distances to the mouths of contributing rivers of water quality stations. The first principal components was used to define a new coordinate system onto which the water quality stations were be projected. Once the stations are located in this new space, we can use the leave-one-out cross-validation procedure outlined above to identify water quality stations that are more or less important when kriging prediction maps.

A variety of measures were be used to evaluate the performance of the kriging model. The *Mean Error* is the mean difference between observed and predicted values and is ideally 0. The *Median Square Prediction Error* (MSPE) is calculated by squaring all the errors and taking their median; it is ideally a small number. The *Mean Square Normalized Error* (MSNE) was calculated as in Equation 1.

$$MSNE = \frac{\left(\frac{residual}{standard\ error}\right)^2}{n} \tag{1}$$

where *n* is the number of water quality stations. MSNE is ideally close to 1. The correlation between observed and predicted values, ideally 1, and the correlation between predicted and residual, ideally 0, will also be used.

These were used in combination to identify the best kriging model.

3.3 Water quality variables

A review of the literature reveals that the most commonly used statistical approach to reduce the number of variables being measured is the correlation-regression (CR) approach. The CR approach is based on three steps. First, correlation analysis is used to assess the level of association among the variables being measured. If high correlation exists among the variables, some of the information produced may be redundant. Second, the water quality variables to be continuously measured or discontinued are selected. 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, and the cost of laboratory analysis. Third, the reconstitution of information about the discontinued variables using auxiliary variables from the continuously measured variables is examined.

The main advantage of the CR approach is that it allows the reconstitution of information about the discontinued variables using regression analysis. However, Khalil et al. (2010a) identified three main deficiencies in the CR approach as commonly practiced for water quality variable reduction. The first deficiency involves 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 based on subjective preference. Thus, studies using the same variables but performed by different investigators may lead to different results. The second deficiency is the absence of a criterion to identify the combination of variables that should be continuously measured or 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).

Khalil et al. (2010a) modified the CR approach to overcome these three deficiencies by using criteria from record-augmentation procedures to identify a correlation coefficient threshold. To identify optimal combinations of variables to be continuously measured and variables to discontinue, Khalil et al. (2010a) proposed an information performance index to evaluate different combinations of variables. For reconstitution of information about the discontinued variables, Khalil et al. (2010a) recommended the use of the maintenance of variance extension technique type 3 (MOVE3), whereas the regression technique was recommended for the reconstitution of missing values.

In this study, the approach proposed by Khalil et al. (2010a) was applied to identify optimal combinations of variables to be continuously measured and variables to discontinue. The following subsection provides background on the record augmentation and extension approaches. The second subsection presents the proposed approach.

3.3.1 Background on record augmentation and extension approaches

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

3.3.1.1 Record augmentation

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

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

In the case of water quality variables reduction, year n_1 can be considered as the year where assessment and selection took place. After n_1 years, assessment of the water quality variables reveals that measurement of the variable y can be stopped, whereas measurement of variable x can be continued. Assume that after n_2 years our interest is to estimate the mean μ_y and 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 the following:

$$\hat{\mu}_y = \bar{y}_1 + \frac{n_2}{n_1 + n_2}\hat{\beta}(\bar{x}_2 - \bar{x}_1) \tag{2}$$

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 the following:

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

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). 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 the following:

$$\rho^2 > 1/(n_1 - 2) \tag{4}$$

Therefore, estimation of the mean from the extended series is profitable only if the correlation coefficient between the two variables exceeds $\left(|n_1 - 2|\right)^{-1/2}$ (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 $\hat{\sigma}_y^2$:

$$\hat{\sigma}_y^2 = \hat{\beta}^2 s_x^2 + \left[1 - \frac{n_1 + n_2 - 3}{(n_1 - 3)(n_1 + n_2 - 1)}\right]\frac{n_1 - 1}{n_1 - 2}(s_{y1}^2 - \hat{\beta}\, s_{x1}^2) \qquad (5)$$

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 the following:

$$Var\{\hat{\sigma}_y^2\} = \frac{2\sigma_y^4}{n_1 - 1} + \frac{n_2 \sigma_y^4}{(n_1 + n_2 - 1)^2 (n_1 - 3)}(A\rho^2 + B\rho + C) \qquad (6)$$

where

$$A = \frac{(n_2 + 2)(n_1 - 6)(n_1 - 8)}{(n_1 - 5)} + (n_1 - 4)\left(\frac{n_1 n_2 (n_1 - 4)}{(n_1 - 3)(n_1 - 2)} - \frac{2n_2(n_1 - 4)}{(n_1 - 3)} - 4\right) \qquad (7)$$

$$B = \frac{6(n_2 + 2)(n_1 - 6)}{(n_1 - 5)} + 2\left(n_1^2 - n_1 - 14\right) + (n_1 - 4)\left(\frac{2n_2(n_1 - 5)}{(n_1 - 3)} - 2(n_1 + 3) - \frac{2n_1 n_2 (n_1 - 4)}{(n_1 - 3)(n_1 - 2)}\right) \qquad (8)$$

$$C = 2(n_1+1) + \frac{3(n_2+2)}{(n_1-5)} - \frac{(n_1+1)(2n_1+n_2-2)(n_1-3)}{(n_1-1)} + (n_1-4)\left(\frac{2n_2}{(n_1-3)} + 2(n_1+1) + \frac{n_1 n_2 (n_1-4)}{(n_1-3)(n_1-2)}\right) \quad (9)$$

The first term on the right-hand side in Equation 6 is equal to the variance of the variance estimator based on the n_1 years of the y-series; therefore, estimation is profitable when:

$$\rho^2 > (-B \pm \sqrt{B^2 - 4AC}) / 2A \quad (10)$$

Khalil et al. (2010a) used Equations 4 and 10 as criteria for assessment of the required correlation coefficient threshold and identification of highly associated pairs of variables. By using such criteria, Khalil et al. (2010a) overcame the first disadvantage of the CR approach. In this study, the higher correlation coefficient obtained from Equations 4 and 10 is used as a criterion to evaluate the correlation coefficients among water quality variables measured at the seven selected locations.

3.3.1.2 Record extension

To estimate records of the discontinued variable y for the period n_1+1 through n_2 years, the simple linear regression of y on x can be used.

$$\hat{y}_i = a + bx_i \quad (11)$$

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 minimise the sum of the squared differences 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 11 is the following:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) \qquad (12)$$

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 obtained if the following equation is used:

$$\hat{y}_i = \bar{y}_1 + r\,(s_{y1}/s_{x1})\,(x_i - \bar{x}_1) + \alpha\left(1 - r^2\right)^{1/2} s_{y1}\, e_i \qquad (13)$$

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 13 is that studies of the same sequence of x and y by different investigators will almost certainly 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 11 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 and an empirical analysis. Both the Monte Carlo study and the empirical analysis showed that regression cannot be expected to provide records with the appropriate variability.

In practice, Equation 11 is used to generate the $\hat{y}_i$ only for $i = n_1 + 1, ..., n_1 + n_2$. This suggests that Hirsch (1982) 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 11 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; Equations 2 and 5). Estimates of a and b for the MOVE3 method are obtained by rewriting Equation 11 as follows:

$$\hat{y}_i = a + b(x_i - \bar{x}_2) \tag{14}$$

where estimates of a and b are obtained as follows:

$$a = \left[(n_1 + n_2)\hat{\mu}_y - n_1 \bar{y}_1\right] / n_2 \tag{15}$$

$$b^2 = \left[(n_1 + n_2 - 1)\hat{\sigma}_y^2 - (n_1 - 1)s_{y1}^2 - n_1(\bar{y}_1 - \hat{\mu}_y)^2 - n_2(a - \hat{\mu}_y)^2\right]\left[(n_2 - 1)s_{x2}^2\right]^{-1} \tag{26}$$

The Kendall-Theil robust line (KTRL) method has been proposed as an analogue to regression, with the advantage of being robust in the presence of extreme values. The KTRL method is based on 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 that is 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 regression (Helsel and Hirsch, 2002).

A robust estimate of the slope for this nonparametric fitted line was first described by Theil (1950). The Theil slope is estimated 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):

$$b_K = median\,\frac{y_j - y_i}{x_j - x_i} \qquad (37)$$
$$\forall\, i < j \qquad i = 1,2,\ldots\ldots n-1 \qquad j = 2,3\ldots\ldots\ldots,n$$

The intercept (a_K) is defined as follows:

$$a_K = median\,(y) - b_K * median\,(x) \qquad (48)$$

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

Khalil et al. (2010b) generated a modified KTRL method, referred to as KTRL2, that utilises MOVE techniques to reduce the bias in the estimation of the variance while incorporating the robustness of the KTRL method in the presence of extreme values. The KTRL2 method proposed by Khalil et al. (2010b) follows the KTRL method, 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 approximations of the CDFs of the actual records. The developmental goal of KTRL2 is to find the values of a_q and b_q in Equation 11 that minimise the error when estimating the y percentiles; a_q and b_q are defined as follows:

$$b_q = median \frac{q(y)_j - q(y)_i}{q(x)_j - q(x)_i} \qquad (59)$$

$$\forall\, i < j \qquad i = 5^{th}, 10^{th}, \ldots\ldots 90^{th} \qquad j = 10^{th}, 15^{th} \ldots\ldots 95^{th}$$

The intercept is defined as follows:

$$a_q = median(y) - b_q * median(x) \qquad (206)$$

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… 95th percentile. Thus, a set of 19 (x, y) pairs of percentiles will result in 171 [*n (n-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. Consequently, the objective in developing the KTRL2 method was to minimise the error in estimating the y percentiles rather than to minimise the error in estimating the y records.

3.3.2 *Proposed approach*

The approach proposed by Khalil et al. (2010a) consists of four main steps. The first step is to assess the level of association among the variables being measured and to define the 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 the 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 from which the information about discontinued variables can be reconstituted from the continuously measured variables.

3.3.2.1 Association Assessment

To identify highly associated water quality variables, cluster analysis was employed with criteria developed from the record augmentation procedures. Hierarchical clustering was performed 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. Khalil et al. (2010a) used Equations 4 and 10 to identify the dissimilar, highly correlated variables that may be grouped in clusters as follows:

$$d_m < 1 - \frac{1}{n_1 - 2} \tag{71}$$

$$d_v < 1 - \left(- B \pm \sqrt{B^2 - 4AC} \right) / 2A \tag{82}$$

Assessment of the correlation between water quality variables was applied for each of the seven sampling sites separately. It should be noted that some of the clusters contain only one variable each (single-variable cluster). The final, rationalised 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.

3.3.2.2 Best auxiliary variables

After identifying clusters of highly associated variables, the following step was 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. In this study, Equation 3 was used to identify the best auxiliary variable that minimises the variance of the estimated mean of each discontinued variable.

Using Equation 3, the choice of the best auxiliary variable is based on the number of concurrent years of measurement, the correlation coefficient and 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 3) 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 three years; therefore, reconstitution of information about discontinued variables would occur three years after the assessment and reselection took place.

3.3.2.3 Selection of discontinued variables

Khalil et al. (2010a) proposed an information index to define the optimum combination of variables that could be continuously measured or discontinued. For instance, consider the case where budget cuts require k variables to be discontinued. The question becomes 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, the information index is computed and the combinations are ranked based on their information index values. Such a procedure provides the decision maker with the rank of the best combinations to discontinue. The aggregate information index (I_a) is defined as follows:

$$I_a = \sum_{\text{var}\,iables\ X} \sqrt{Var\{\hat{\mu}\{X\}\}} \qquad (93)$$

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 3). This summation was performed over all variables in the study. For the discontinued variables, the variance of the mean value estimator after n_2 years was estimated using Equation 3. The population parameters in Equation 3 were 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 was 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 was applied to the standardised variables to remove the dimensionality and the scale effects from the variables. Using such an aggregated performance index, the second disadvantage in the conventional CR approach can be overcome. Figure 1 illustrates the flow of the analyses as described above.

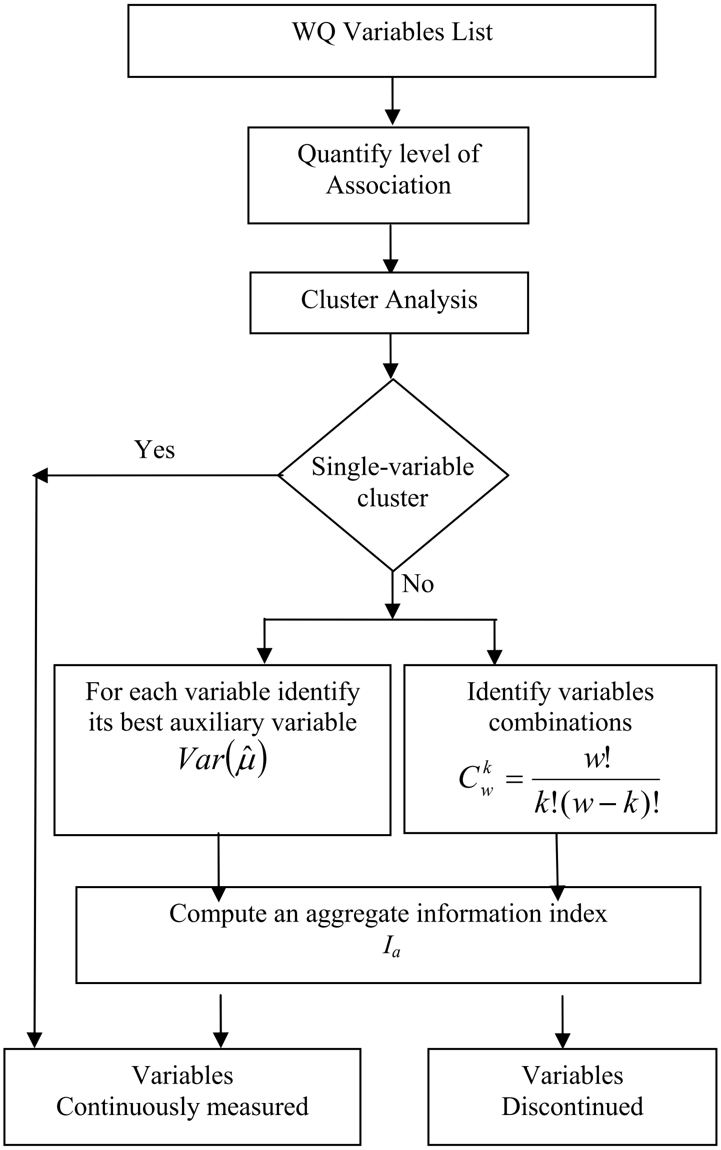


Figure 1. Flow chart of the proposed rationalization approach (Khalil et al., 2010a)

3.3.2.4 Information transfer

The objective in this step is to estimate daily, weekly or monthly records of the discontinued variable for the n_2 years $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$, while maintaining the 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 values 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). In this study, the linear regression and KTRL2 record-extension techniques were applied to reconstitute information about discontinued variables.

An empirical experiment was designed to examine the utility of the simple linear regression and the KTRL2 techniques for preserving the statistical characteristics of the discontinued water quality variables measured at Lake Winnipeg. To evaluate the performance of the two record-extension techniques, a cross-validation (jack-knife) was conducted. In the cross-validation, three years of records were removed, in turn, from the available nine years of data. All possible combinations of successive or non-successive three-year periods were considered. Thus, from the available nine years of records, $C(9,3) = 84$ possible combinations were considered. The values for these three years of observations were then estimated using the two record-extension techniques calibrated with the remaining six years.

The experimental design was as follows: for each pair of water quality variables identified in the previous step as the discontinued variable and its best auxiliary, the two record-extension techniques were applied. Thus, different realisations of extended water quality variable records (i.e., the possible pairs × 84 in the 7 selected locations) were generated for cross-validation. Important characteristics of the observed and generated records during the extension period were computed for the two record-extension techniques. The evaluation of records generated by the extension techniques involved determining the ability of the techniques to reproduce the various statistical properties of the observed records.

The extended records $\{\hat{y}_{n_1+1},....,\hat{y}_{n_2}\}$ were compared to the observed records $\{y_{n_1+1},....,y_{n_2}\}$ based on estimation of the mean, standard deviation and the full range of percentiles (from the 5^{th} to the 95^{th} percentile). Different performance measures were applied. First, the ratio U was calculated based on each statistic computed from the extended series over those computed from the observed records. The ratio *U* is used to assess the performance of the record-extension techniques for preserving the discontinued water quality variable characteristics. If the ratio *U* for a given statistical parameter was larger than one, the applied technique overestimated this parameter. If it was less than one, the technique underestimated the parameter.

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

$$BIASr = \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \frac{\hat{y}_i - y_i}{y_i} \tag{104}$$

$$RMSEr = \sqrt{\frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \left[\frac{\hat{y}_i - y_i}{y_i} \right]^2} \tag{115}$$

where $\hat{y}_i$ and y_i are, respectively, the estimated and the measured values of the dependent variable for $i = n_1 + 1, \ldots n_2$. The two performance measures of Equations 24 and 25 were applied to compare the estimated records with the observed records.

3.4 Sampling frequency

Two statistical approaches were employed to assess the sampling frequency. The semi-variogram approach and confidence interval width around the mean were used to identify the optimum sampling frequency for water quality variables measured at each of the selected monitoring locations.

3.4.1 *Semi-variogram approach*

The semi-variogram method is based on the fundamental work of Krige (1951) and Matheron (1963), and it is the first stage in a geostatistical analysis. Khalil et al. (2004) introduced the semi-variogram as a way to determine the optimal sampling intervals in water quality monitoring. As stated previously, the semi-variogram 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 semi-variogram is a plot of one distance, or "lag", separating pairs of points (x-axis) (Figure 2). The general equation for the semi-variogram is as follows:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)-1} \left[Z(x_{i+h}) - Z(x_i) \right]^2 \qquad (126)$$

where $\gamma(h)$ is the semi-variance; $n(h)$ 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 semi-variogram 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 (Figure 2). The curve fitted to the data usually consists of two distinct parts: an initial rising segment and a horizontal flat segment (Figure 2).

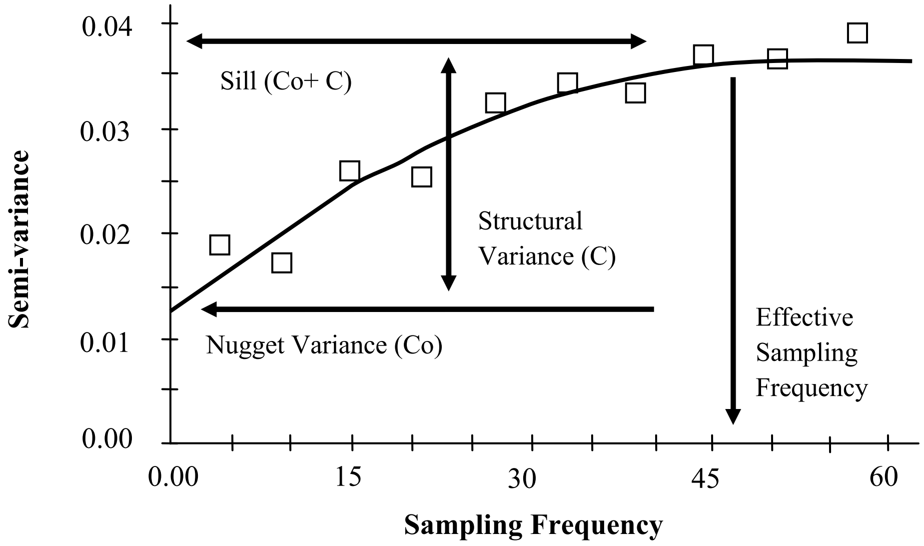


Figure 2. Semi-variogram parameters

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 semi-variance 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 temporally uncorrelated. A sample can be considered representative of the time defined by the effective sampling frequency, which is the range of serial correlation.

3.4.2 Confidence Interval width

Sanders and Adrean (1978) developed a methodology to determine the required sampling frequencies in time if the information sought is the true mean value of a water quality variable at a specified level of statistical confidence. The method is based on the assumption that the monitoring objectives are the determination of ambient water quality conditions and an assessment of annual trends. The question is how many samples should be taken to determine the true mean ($\bar{x}$) with a certain level of confidence. Assuming that the population variance is known and that the samples are independent, the variance of the sample mean ($\mathrm{var}(\bar{x})$) may be computed from the following:

$$\mathrm{var}(\bar{x}) = \frac{\sigma^2}{n} \qquad (137)$$

where σ is the population variance and n is the number of samples. The number of samples required to obtain a given degree of confidence can be derived from the following:

$$\left[\bar{x} - z_{\alpha/2}\,\mathrm{var}(\bar{x})^{\frac{1}{2}} \le \mu \le \bar{x} + z_{\alpha/2}\,\mathrm{var}(\bar{x})^{\frac{1}{2}}\right] \qquad (148)$$

Replacing *var* $(\bar{x})$ by σ^2/n, the n required to obtain a future estimate of the mean can be computed with a known level of confidence by the following (Sanders et al., 1983):

$$n \geq \left[\frac{z_{\alpha/2}\,\sigma}{\mu - \bar{x}}\right]^2 \qquad (159)$$

Substituting the sample variance (*s*) instead of the population variance (σ) in Equation 29 requires Student's "*t*" statistic instead of *z*. The difference between the true population mean and the sample mean (μ-$\bar{x}$) is sometimes referred to as the error (*E*). This substitution results in the following:

$$n \geq \left[\frac{t_{\alpha/2}\ s}{E} \right]^2 \qquad (3016)$$

Using Equation 30, an acceptable confidence level (1-α) and the error (E) must be specified to determine n. Because the value of the t is dependent on the number of samples (degree of freedom), which is what we are looking for, the computation of n becomes an iteration problem. A value for n is assumed, the degrees of freedom are determined, and a first approximation of n is computed using Equation 30. A new n is selected, and a second approximation is then computed. The procedure is repeated until successive approximations are nearly equal. To illustrate the effect of confidence levels and the error expected on the number of samples, three confidence levels (95%, 90% and 80%) and several expected errors (10% to 60% from the sample mean) were considered in this study.

In practice, autocorrelation may be present, meaning that part of the information contained in one measurement is also contained in subsequent measurements. In this case, the variance of $\bar{x}$ is as follows (Loftis and Ward, 1979):

$$Var(\bar{x}) = \frac{\sigma^2}{n} \left[1 + \frac{2}{n} \sum_{k=1}^{n-1} (n-k)\rho_k \right] \qquad (171)$$

where ρ_k is the autocorrelation coefficient for lag k. When $Var(\bar{x})$ given by Equation (31) is substituted into Equation (28), a quadratic equation is obtained, which can be solved for n. Gilbert (1987) obtained an approximate expression by ignoring the term in the quadratic equation with n^2 in the denominator:

$$n \geq D \left[1 + 2 \sum_{k=1}^{n-1} \rho_k \right] \qquad \text{where D} = \left[\frac{(t_{\alpha/2})\ s}{E} \right]^2. \qquad (182)$$

The relationship between the confidence interval and the number of samples, as shown in Equation (30), becomes theoretically valid if the variance of the stationary component computed from different sampling intervals stabilises after a certain sampling interval (Harmencioglu et al., 1999). 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.

4 RESULTS

The results are presented in four subsections: the first subsection presents the results of preliminary analyses; the second subsection presents the isotope station number and configuration assessment; the third subsection presents the water quality variable assessments; and the last subsection presents the sampling frequency assessment.

4.1 Preliminary analyses

Results of preliminary analyses for the monitoring location MB05SBS126 are presented here. A complete set of results for the seven sampling sites is presented in Appendix 4. Table 2 shows the descriptive statistics for the 13 water quality variables measured at MB05SBS126.

The maximum number of records available at MB05SBS126 was 35, while the minimum number of samples was 21, for temperature. The mean and median for each of the measured variables were similar, indicating that the data distribution was almost symmetric. The skewness and kurtosis measures indicated that the variables under analysis were symmetric around their mean values, with normal tails.

Table 2. Location MB05SBS126, descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Colour	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	**mg/l**	**mg/l**	**mg/l**	**CU**	**US/cm**	**mg/l**	**pH units**	**mg/l**	**mg/l**	**deg C**	**mg/l**	**mg/l**	**NTU**
n	33	33	33	33	35	31	35	31	33	21	33	29	33
Mean	111.13	134.14	11.07	40.70	347.89	141.14	7.95	14.37	50.68	11.45	235.97	19.97	23.35
Median	112.00	137.00	11.50	35.00	334.00	135.00	8.02	13.90	45.00	12.60	210.00	11.00	17.10
Std. Deviation	28.37	33.51	4.35	24.40	113.44	45.62	0.36	6.49	26.61	8.14	78.81	18.56	19.88
Variance	805.04	1123.09	18.91	595.28	12869.75	2080.76	0.13	42.07	708.35	66.21	6211.03	344.53	395.32
Skewness	0.22	0.29	-0.21	3.85	0.11	0.22	-1.03	0.11	0.22	-0.12	0.97	1.77	1.69
Std. Error of Skewness	0.41	0.41	0.41	0.41	0.40	0.42	0.40	0.42	0.41	0.50	0.41	0.43	0.41
Kurtosis	-0.39	-0.10	-1.01	18.44	-1.02	-0.79	1.25	-0.60	-1.20	-1.06	1.06	2.59	3.07
Std. Error of Kurtosis	0.80	0.80	0.80	0.80	0.78	0.82	0.78	0.82	0.80	0.97	0.80	0.85	0.80
Range	108.90	133.40	14.52	142.00	385.00	164.00	1.58	26.40	87.80	24.00	344.00	72.00	87.10
Minimum	61.10	74.60	3.48	18.00	159.00	68.00	6.90	1.60	9.30	0.00	122.00	2.00	3.90
Maximum	170.00	208.00	18.00	160.00	544.00	232.00	8.48	28.00	97.10	24.00	466.00	74.00	91.00

Table 3. Location MB05SBS126, Kolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Colour	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	**Absolute**	0.074	0.072	0.080	0.230	0.114	0.074	0.168	0.074	0.137	0.157	0.156	0.226	0.209
	Positive	0.074	0.072	0.080	0.230	0.078	0.072	0.075	0.074	0.132	0.157	0.156	0.226	0.209
	Negative	-0.057	-0.055	-0.076	-0.176	-0.114	-0.074	-0.168	-0.065	-0.137	-0.118	-0.074	-0.176	-0.164
Kolmogorov-Smirnov Z		0.426	0.414	0.462	1.323	0.676	0.411	0.991	0.409	0.788	0.720	0.898	1.218	1.202
Asymp. Sig. (2-tailed)		0.993	0.995	0.983	0.060	0.750	0.996	0.279	0.996	0.565	0.677	0.395	0.103	0.111

To check normality, the Kolmogrov-Smirnov goodness of fit test was employed. The test results indicated that the null hypothesis that the data is normally distributed could not be rejected with a 5% level of significance (Table 3).

The coefficient of variation is the ratio of the standard deviation to the mean. The coefficient of variation indicated that pH had low variation, i.e., only 4.5% of its mean value, while for alkalinity measures ($CaCO_3$ and HCO_3), the variation was around 25%. The variation reached 85% for turbidity, and 93% for TSS.

Boxplots for the 13 water quality variables (Figure 3) showed that the records were almost symmetric around their median values. Boxplots for Colour, TDS, TSS and Turb indicate that some of the records may be considered outliers. Table 4 shows the results of the Mann-Kendal nonparametric trend test. The Z statistic and the probability of accepting the null hypothesis (p-value) suggests there is no trend in the data. Results show that the null hypothesis cannot be rejected for any of the 13 variables measured at MB05SBS126. The ACFs were computed for the considered water quality variables (Figure 4). Figure 4 shows the autocorrelation coefficients for the first 12 lags (monthly basis). Due to the irregularly applied sampling frequency and the limited number of available records, some of the lags did not have enough cases to be computed.

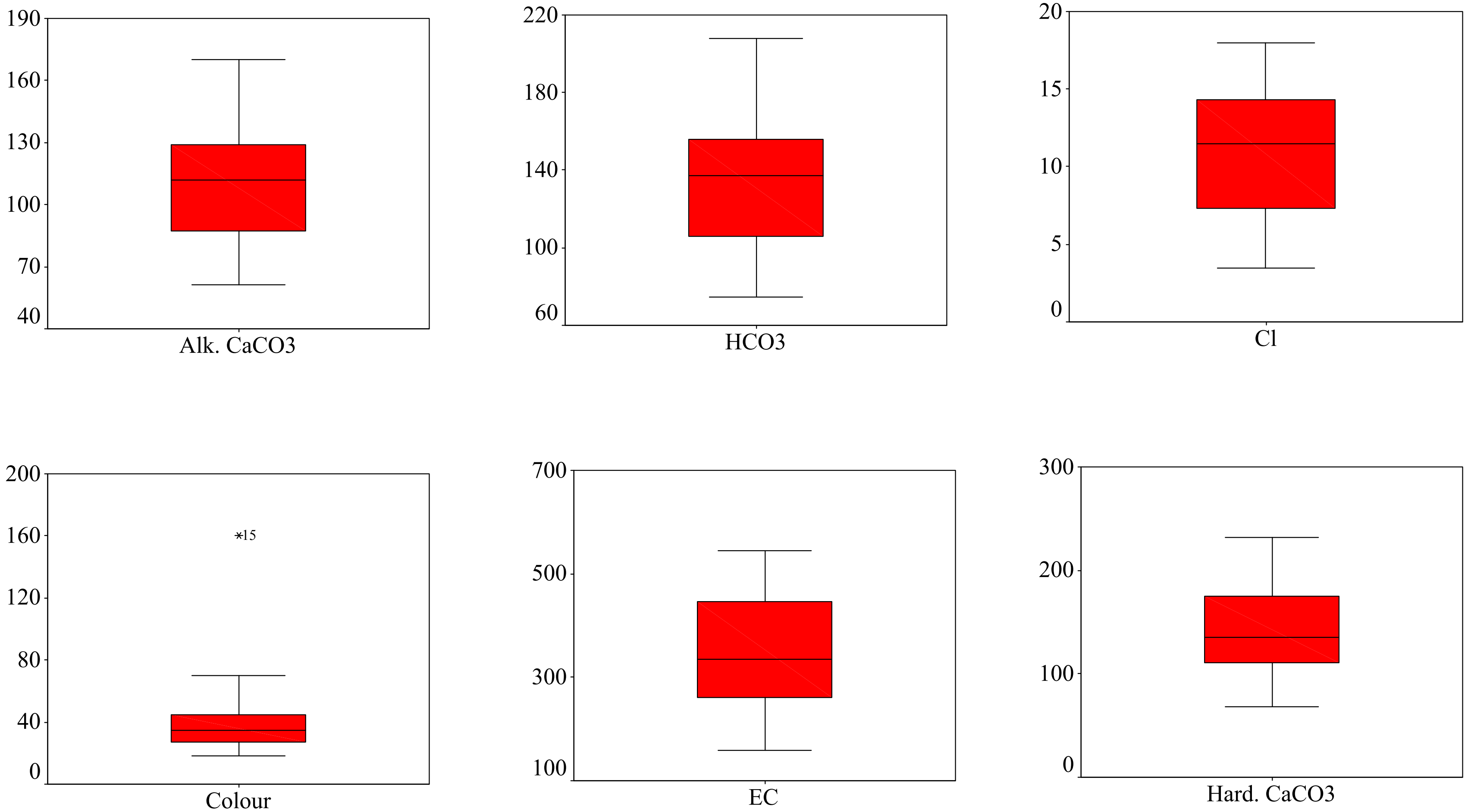


Figure 3. Location MB05SBS126, WQ variables Box-plots

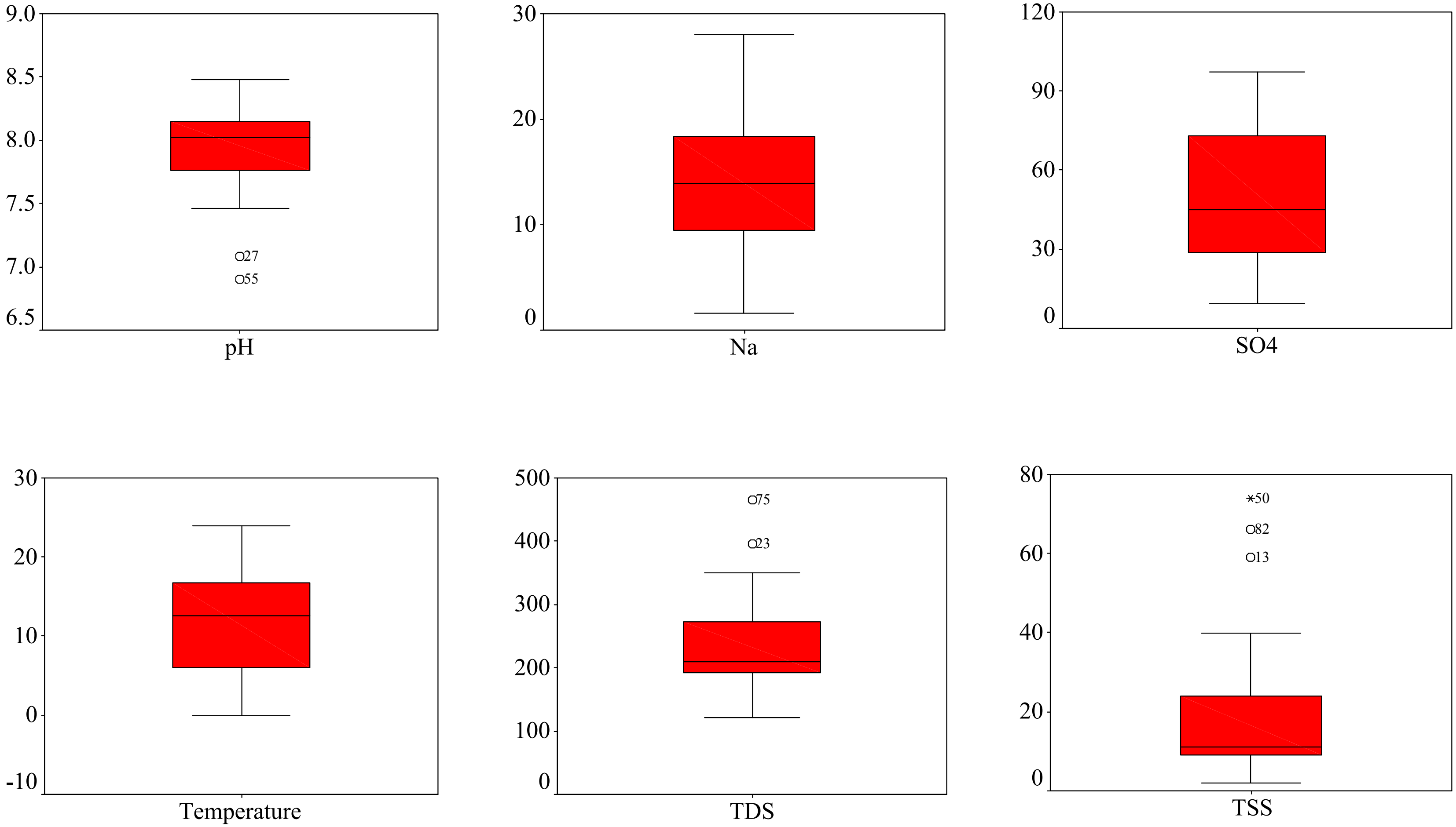


Figure 3 (cont.). Location MB05SBS126, WQ variables Box-plots

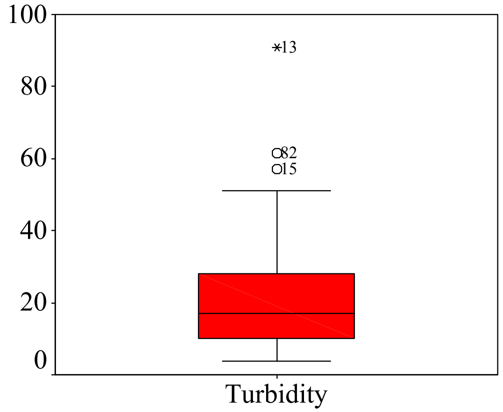


Figure 3 (cont.). Location MB05SBS126, WQ variables Box-plots

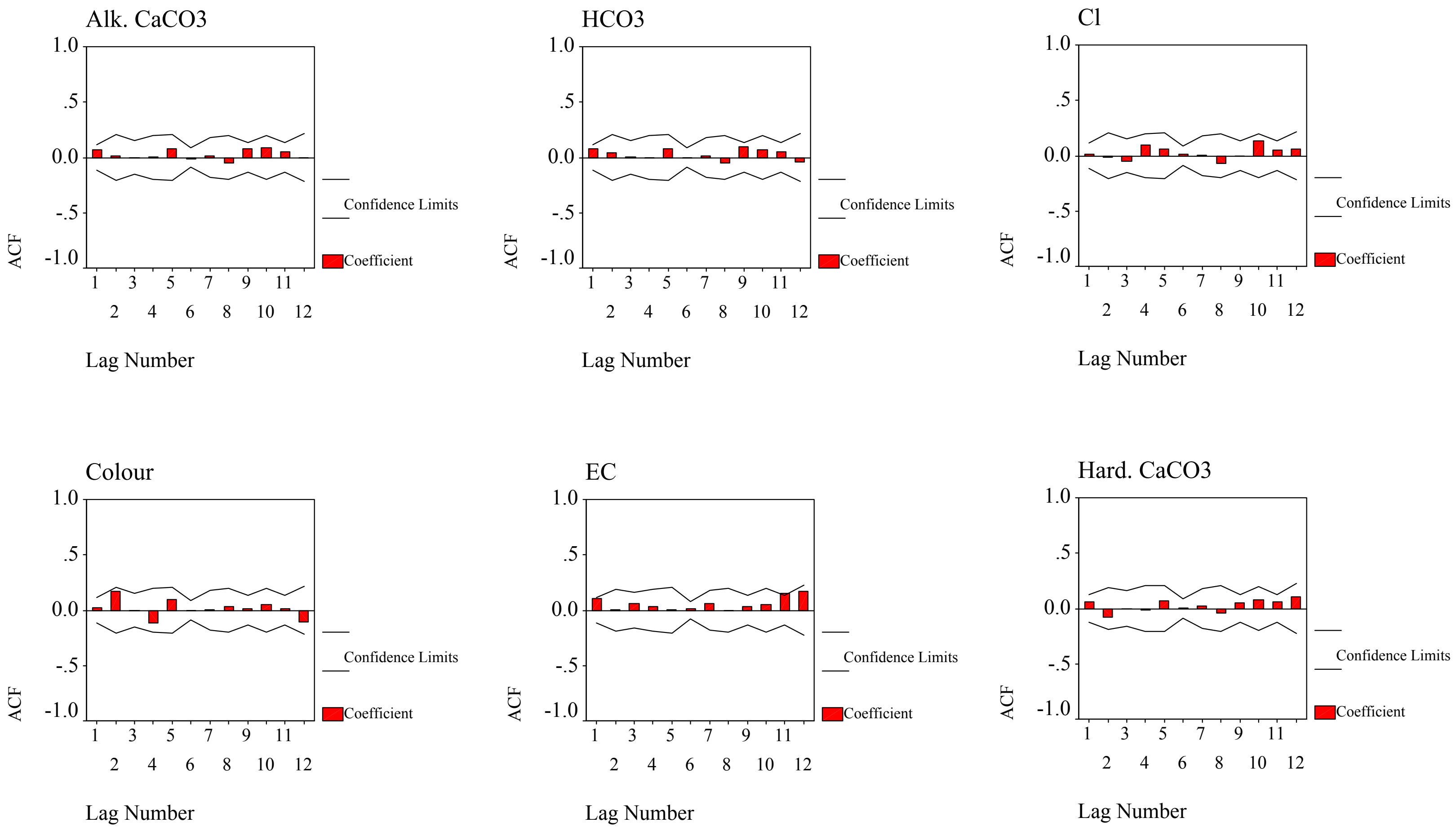


Figure 4. Location MB05SBS126, WQ variables Corellogram

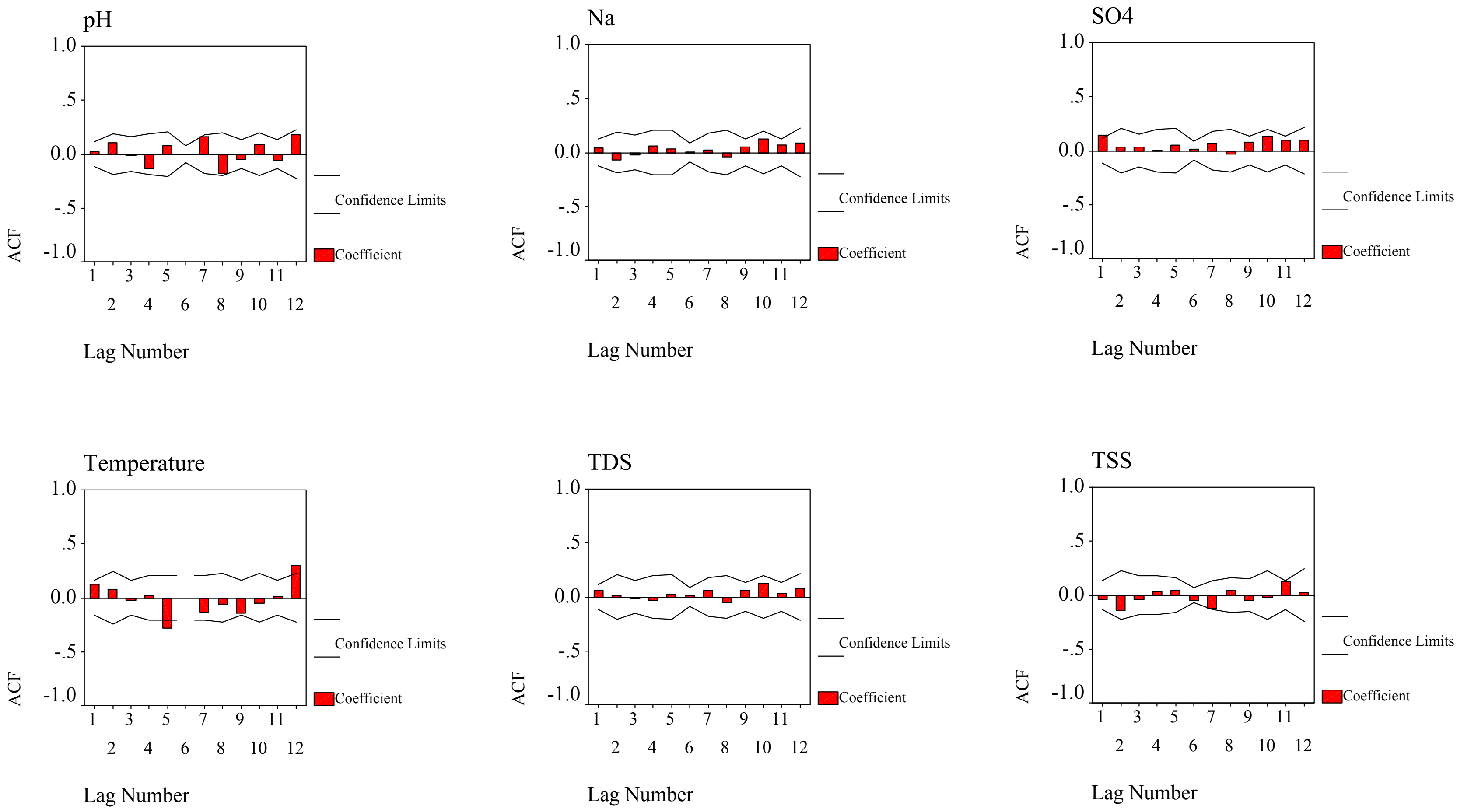


Figure 4 (cont.). Location MB05SBS126, WQ variables Corellogram

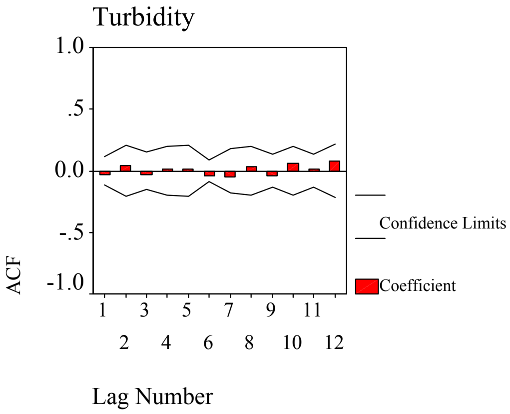


Figure 4 (cont.). Location MB05SBS126, WQ variables Corellogram

Table 4. Mann-Kendall nonparametric trend test results

Location	**MB05RAS078**		**MB05RBS003**		**MB05SAS004**		**MB05SBS126**		**MB05SCS005**		**MB05SHS004**		**MB05SHS014**	
WQ variable	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**
$CaCO_3$	0.41	0.68	-0.12	0.91	-1.57	0.12	-0.19	0.85	0.49	0.62	-2.29	**0.02**	0.38	0.71
HCO_3	0.36	0.72	-0.08	0.93	-1.39	0.16	0.00	1.00	0.56	0.57	-1.23	0.22	1.14	0.25
Cl	-0.56	0.57	-1.36	0.17	-1.87	0.06	-1.16	0.25	-1.34	0.18	-2.75	**0.01**	-1.12	0.26
Colour	-1.16	0.25	0.74	0.46	-0.32	0.75	-1.04	0.30	-1.53	0.13	1.22	0.22	2.27	**0.02**
EC	0.25	0.80	-0.10	0.92	-1.65	0.10	0.41	0.68	0.61	0.54	-1.88	0.06	0.75	0.45
H. $CaCO_3$	0.24	0.81	-0.09	0.93	-0.75	0.45	0.27	0.79	0.25	0.80	-2.94	**0.003**	0.02	0.99
Ph	1.47	0.14	-0.76	0.44	0.94	0.35	0.31	0.75	0.83	0.40	-1.58	0.11	0.56	0.57
Na	0.15	0.88	-0.51	0.61	-1.86	0.06	0.32	0.75	0.46	0.65	-1.38	0.17	-0.98	0.33
SO_4	-0.13	0.90	-0.31	0.76	-2.49	**0.01**	0.22	0.83	-0.03	0.98	-1.54	0.12	-1.27	0.21
Temp.	-0.31	0.76	0.27	0.78	0.53	0.60	1.51	0.13	1.44	0.15	0.25	0.81	0.20	0.85
TDS	0.17	0.86	0.58	0.56	-1.64	0.10	0.05	0.96	0.50	0.62	-2.27	**0.02**	-0.32	0.75
TSS	-1.08	0.28	-2.13	**0.03**	0.04	0.97	-1.22	0.22	-1.28	0.20	0.18	0.86	0.33	0.74
Turb.	-1.07	0.28	-1.24	0.22	-0.18	0.86	-1.69	0.09	-1.31	0.19	1.17	0.24	1.32	0.19

In general, for the seven selected sampling sites, the Kolmogrov-Smirnov goodness of fit test showed that the null hypothesis could be rejected for most of the 13 water quality variables measured. For variables that do not follow normal distributions, logarithmic transformations were applied. Table 5 summarises the results for the Kolmogrov-Smirnov goodness of fit test applied to the 13 water quality variables at the 7 selected sampling sites. Detailed results for each water quality variable at each sampling site are presented in Appendix 4.

Table 5. Kolmogrov-Smirnov test results summary

Sampling site	MB05 RAS078	MB05 RBS003	MB05 SAS004	MB05 SBS126	MB0 5SCS005	MB05 SHS004	MB05 SHS014
$CaCO_3$	N*	N	N	N	N	N	N
HCO_3	N	N	N	N	N	N	N
Cl	N	N	LN**	N	N	SQ***	N
Color	N	N	N	N	N	N	N
EC	N	N	LN	N	N	N	N
H. $CaCO_3$	N	N	N	N	N	N	N
pH	N	N	N	N	N	N	N
Na	N	N	LN	N	N	N	N
SO_4	N	N	LN	N	N	N	N
Temp.	N	N	N	N	N	N	N
TDS	N	N	N	N	N	N	N
TSS	N	N	LN	N	N	N	N
Turb.	N	N	N	N	N	LN	LN

*N stands for normal distribution; **LN stands for log-normal distribution; ***SQ indicates that the Cl is transformed as (Cl^2)

Boxplots for the 13 variables at the 7 selected sampling sites are also presented in Appendix 4. The boxplots show that some of the records could be considered outliers. For the seven sites, few variables had significant temporal trends (Table 4). The variables that did show significant temporal trends were TSS at MBO5RBS003, SO_4 at MBO5SAS004, $CaCO_3$, Cl, H.$CaCO_3$ and TDS at MB05SHS004, and Colour measured at MB05SHS014.

For some of the water quality variables, autocorrelation for some lags could not be computed due to a limited number (or absence) of cases representing these lags. The correlograms for each of the water quality variables measured at the seven sampling sites are presented in Appendix 4.

4.2 Isotope station number and configuration

After having performed PCA on the predictor variables, the stations were plotted along the first two principal components. As shown in Figure 5, the stations are arranged in the shape of an arch.

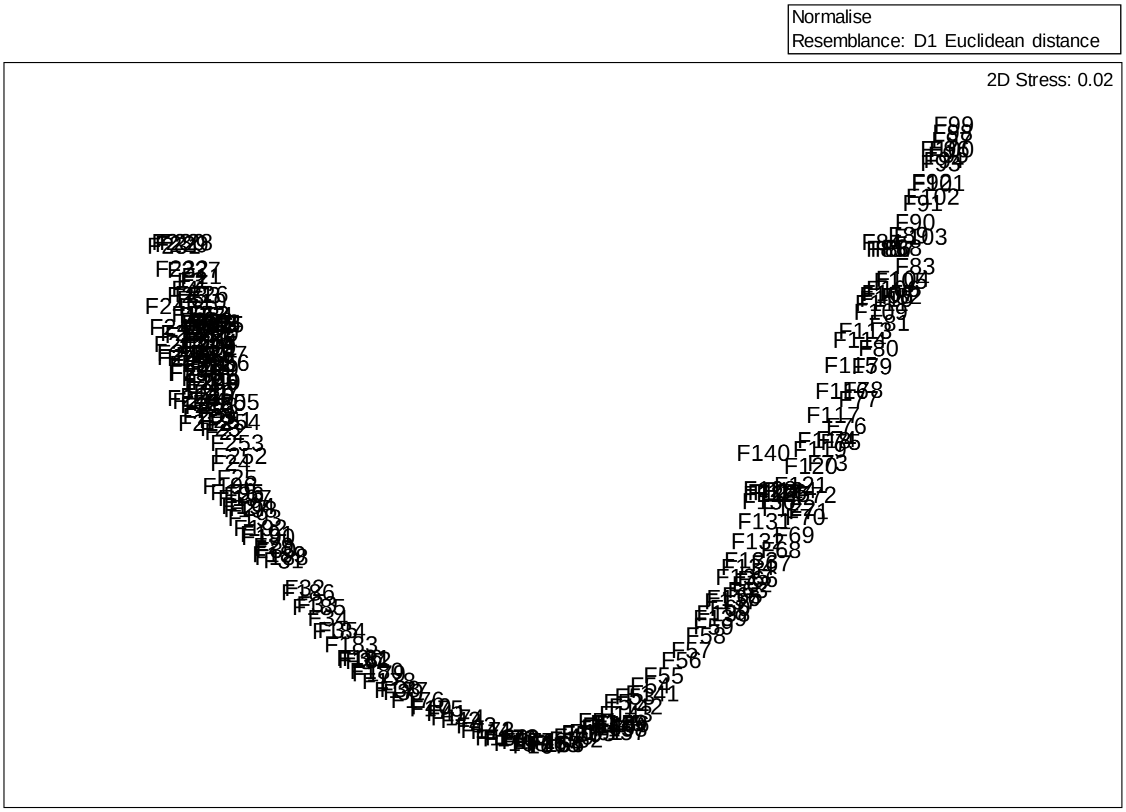


Figure 5. Isotope stations plotted along the first two principal components of a PCA on the predictor variables.
Stations are arranged in the form of an arch.

This effect is typically observed when rendering non-linear phenomena in a two-dimensional Euclidean space (Legendre and Legendre 1998). Presence of an arch does not mean that the plot is an invalid interpretation. However, it does indicate that the new PCA space is slightly distorted and stations at the ends of the horseshoes are closer together than they might ideally be.

It is very difficult to eliminate the effects of this distortion. Techniques to straighten the plane can themselves introduce undesirable artefacts (Legendre and Legendre 1998). The distortion can be lessened however. In this case, two steps were taken. The first was the transformation towards multivariate normality using power transformations. The optimal multivariate power transformations were estimated by a maximum likelihood-like method (Weisberg 2005). The second step was to use non-metric multidimensional scaling, a non-parametric approach similar to PCA. These two techniques diminished the distortion as shown in Figure 6.

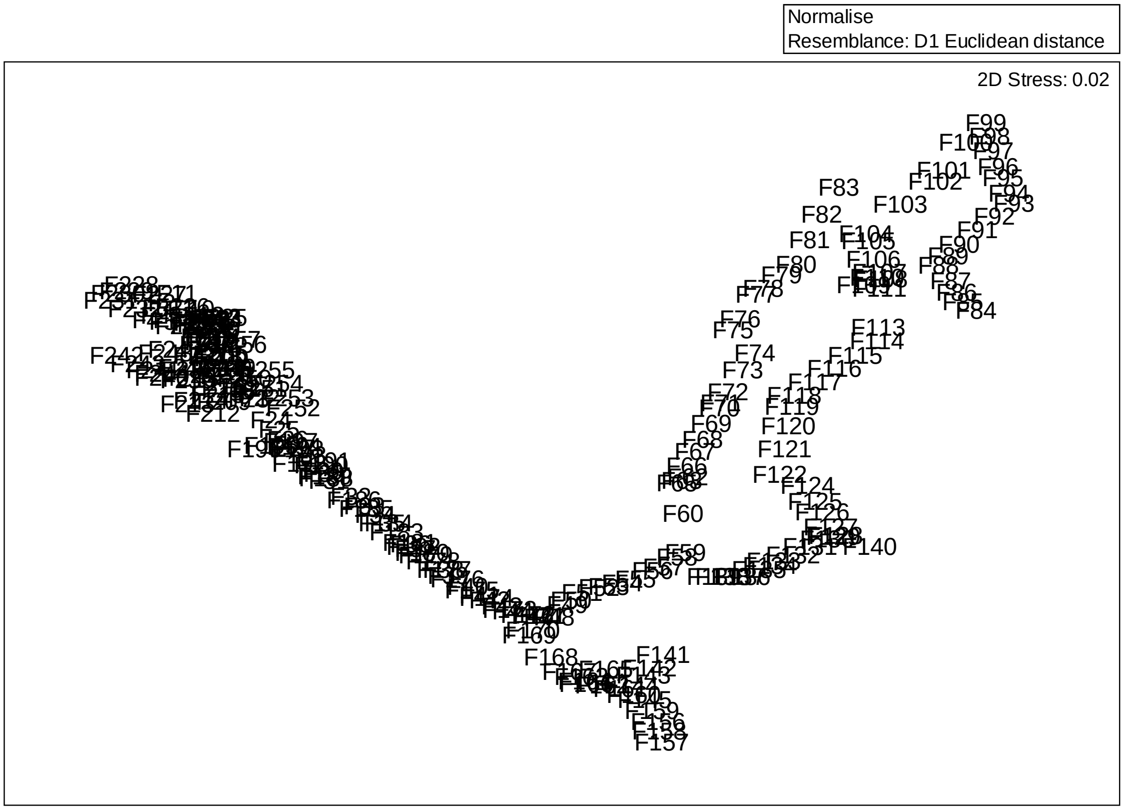


Figure 6. Isotope stations plotted in two-dimensional NMDS space.

There is less arching in NMDS space than in PCA space. It is difficult, if not undesirable, to completely remove arching.

Various combinations of predictor variables and ordinations, i.e. PCA and NMDS, were used to identify the best kriging model. The best kriging model was developed by using two dimensions from a non-metric dimensional scaling analysis on the planar coordinates, i.e. the X and Y coordinates of each station in UTM zone 14N, along with the distance from each station to the mouth of the Red River. The NMDS dimensions formed the basis for the new coordinate system used in the kriging model. Elevation was used as the co-variable.

Table 6 shows the values of the various measures that were used to identify the best kriging models. The best models for both δ^{18}O and δ^{2}H performed quite well with close to ideal values for mean error and MSNE, and almost no correlation between predicted and residual values. δ^{2}H had significantly higher MSPE although both δ^{18}O and δ^{2}H were relatively low. Both δ^{2}H and δ^{18}O models had moderate to moderately high correlations between observed and predicted values. That the correlations were not higher is partly due to the high absolute values and the small range, e.g. δ^{2}H had a mean of 84.09 and a standard deviation of only 5.98; the Pearson correlation coefficient is best suited to quantifying linear c-relationships. In summary, good kriging models could be developed for both δ^{2}H and δ^{18}O.

Table 6. Measures of δ^{18}O and δ^{2}H kriging model performance

	δ^{2}H	δ^{18}O
Mean Error	0.02	0
Median Square Prediction Error	4.03	0.11
Mean Square Normalized Error	1.06	0.97
Correlation between observed and predicted	0.77	0.56
Correlation between predicted and residual	0.00	0.02

Once kriging models had been developed for each isotope, the same measures that were used to evaluate kriging model performance were used to evaluate individual stations. Figure 7 shows the squared residuals for δ^{2}H from the leave-one-out cross-validation, i.e. when predicting δ^{2}H at the location of each station when that location is left out of the kriging data. Most stations can be predicted with little error. Some stations, particularly in the south basin, cannot be predicted with low error. These stations should not be removed from the water quality monitoring network. Figure 8 shows the kriging variance at every iteration of the cross-validation of δ^{2}H. Many stations that contain unique information are found in the north basin. Also highlighted is the cluster of redundant stations in the centre of the south basin.

The spatial distribution of δ^{18}O was easier to model. A few stations in the south basin contributed unique information as shown by the high squared errors in Figure 9. Most stations had some redundancy with their neighbours as highlighted in Figure 10, which plots kriging variance. The kriging variance values over all cross-validations were relatively low and had relatively limited range.

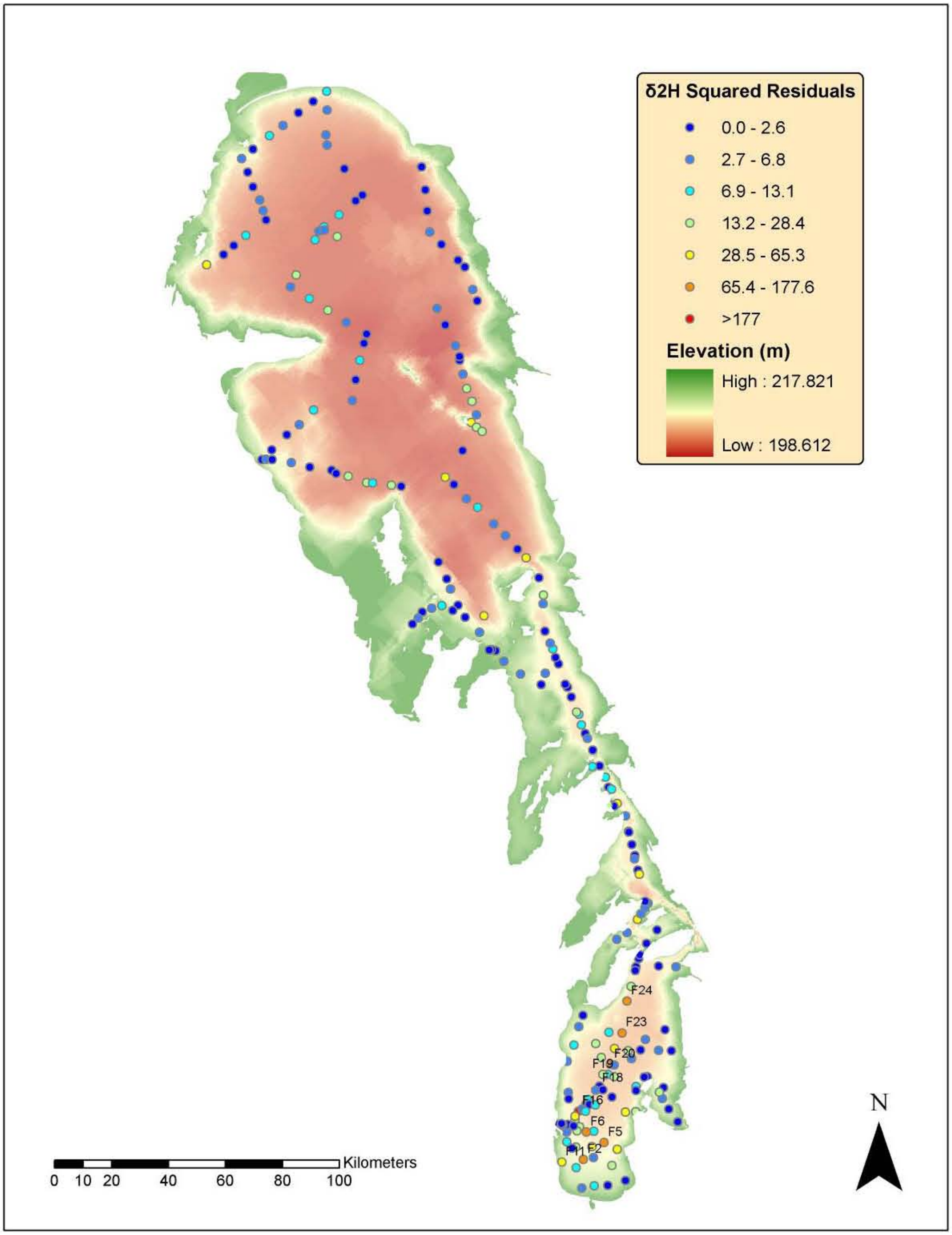


Figure 7. **Squared residual from leave-one-out cross-validation of δ^2H.**

Stations with low values were removed from the kriging model with little effect on prediction error. Station values that could not be predicted with little error when those stations were removed from the kriging model are indicated in red. Most stations important for model calibration are in the South basin.

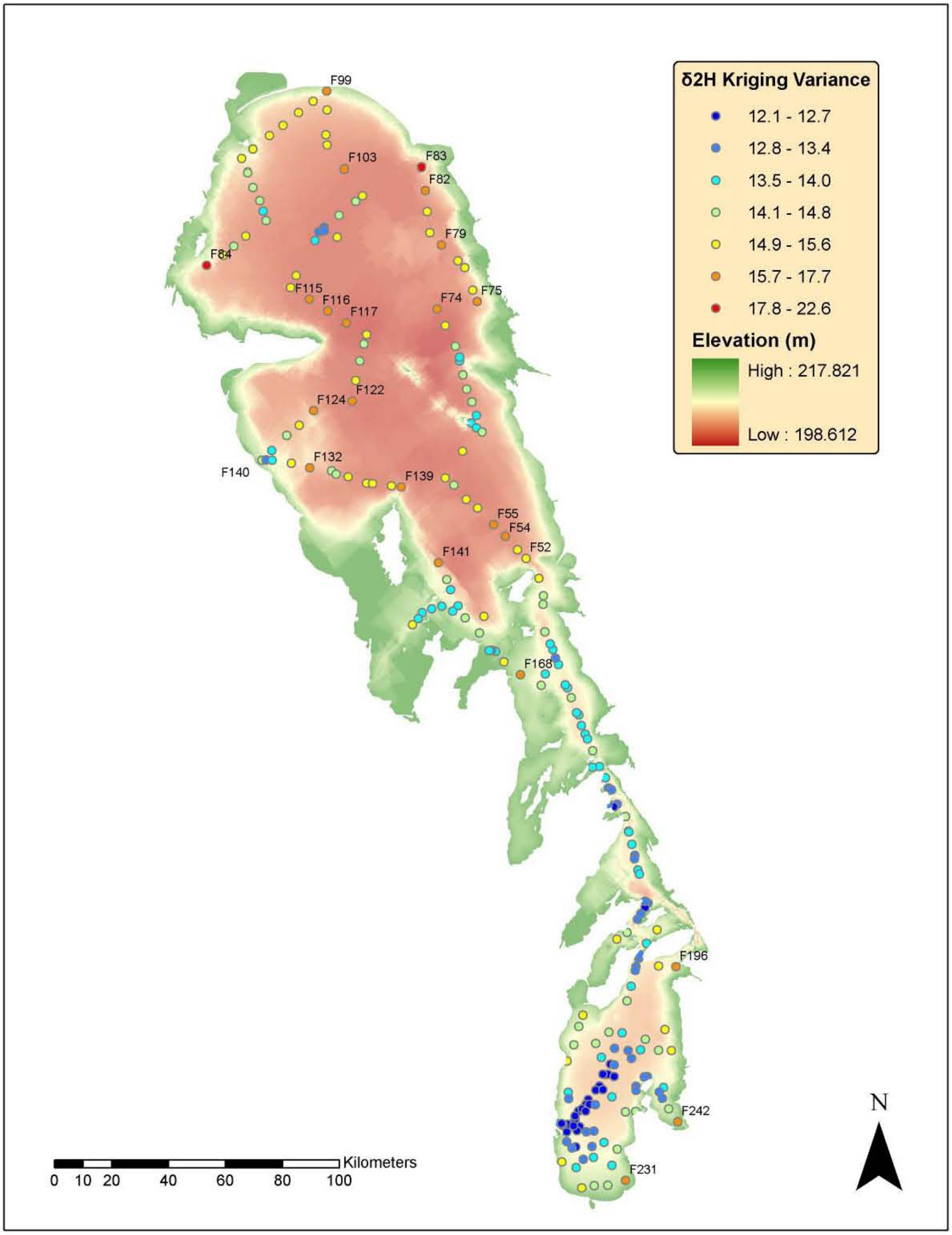


Figure 8. δ^2H kriging variance from leave-one-out cross-validation.

The kriging variance at every iteration of the cross-validation is indicated by the colour of the station that was left out. Lower values in cool colours indicate that those stations when removed yielded kriging maps with lower variance.

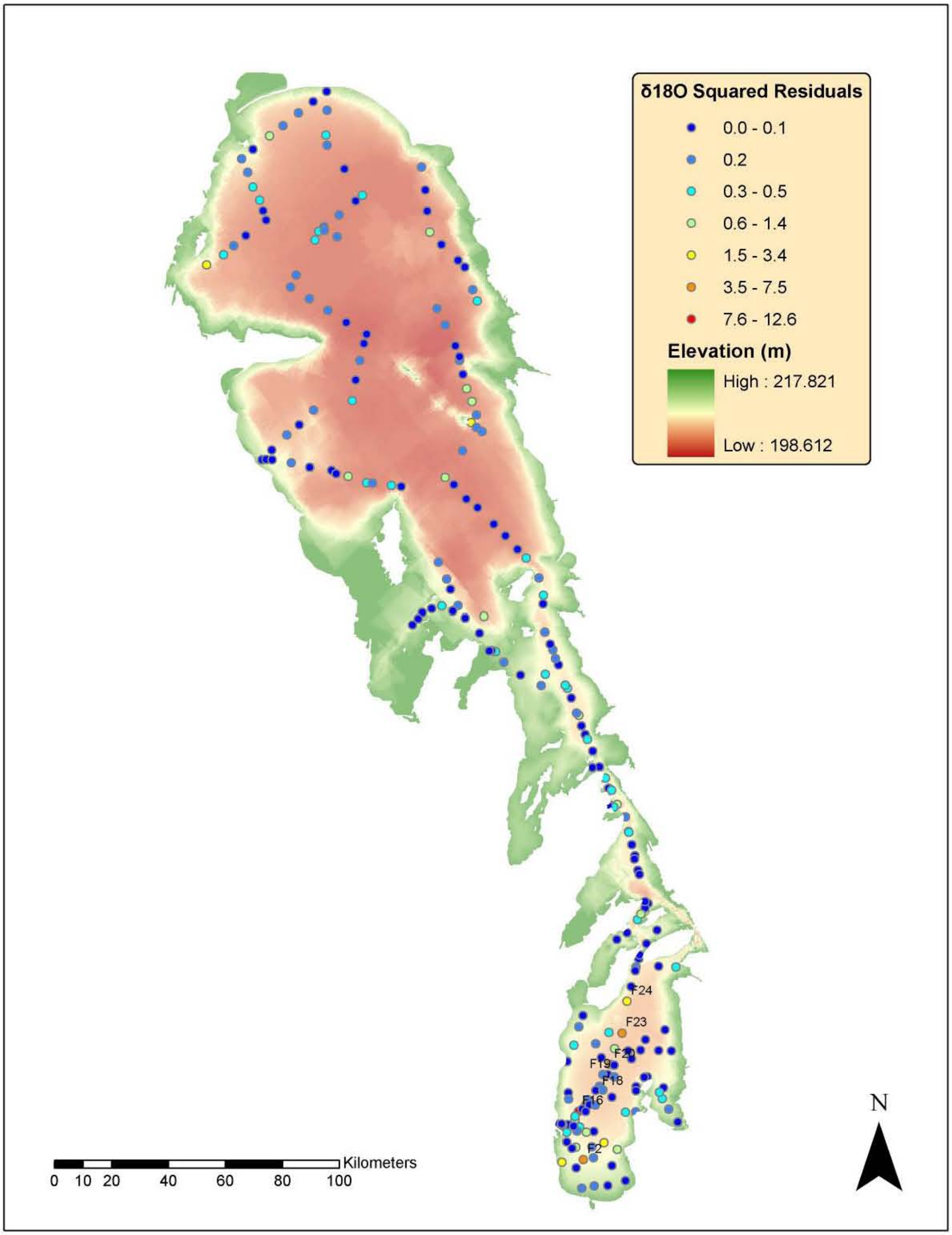


Figure 9. **Squared residual from leave-one-out cross-validation of δ^{18}O.**

Stations with low values in blue were removed from the kriging model with little effect on prediction error. Station values that could not be predicted with little error when removed from the kriging model are indicated in red. As with δ^2H, most stations important for model calibration, i.e. not substitutable, are in the South basin.

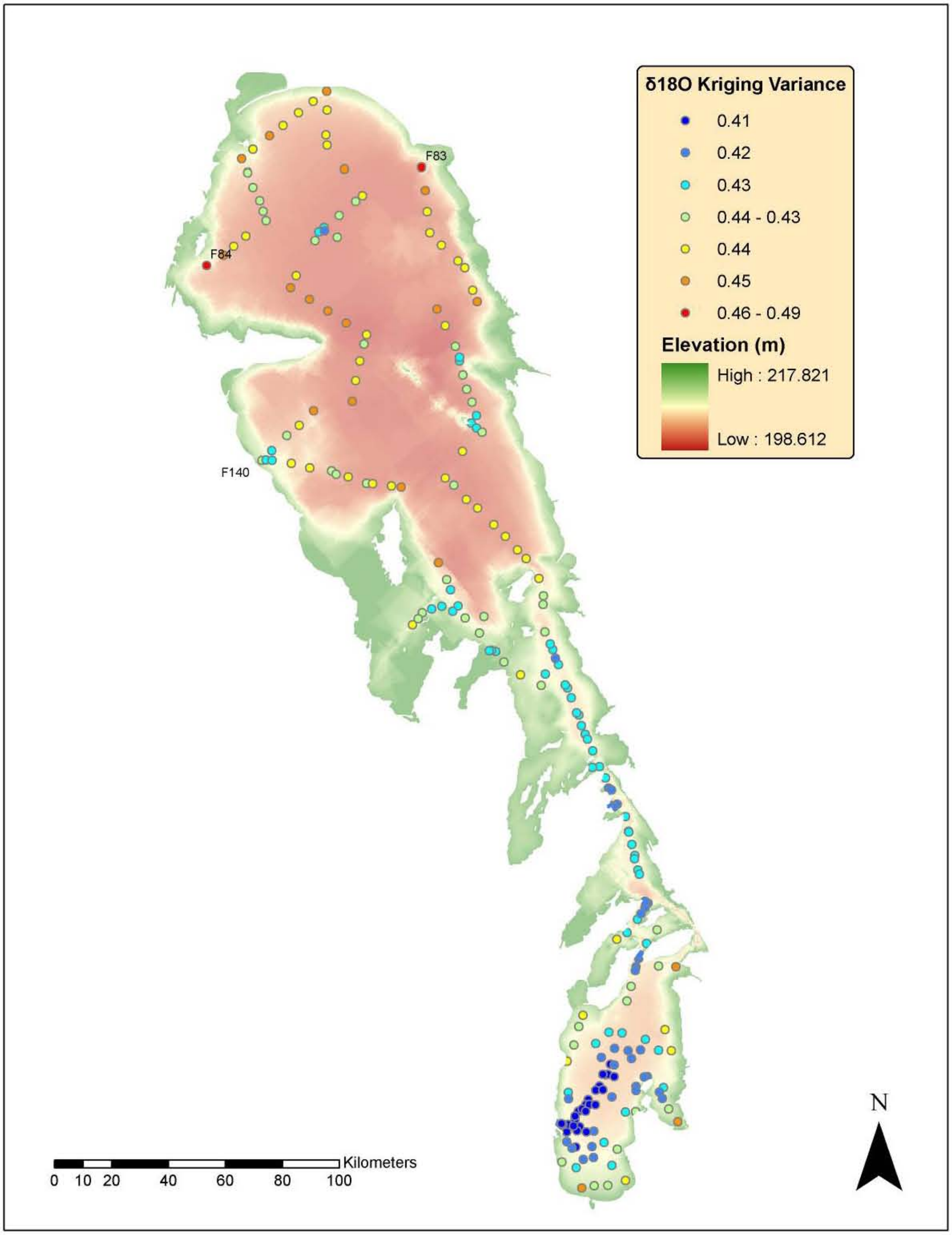


Figure 10. δ^{18}O kriging variance from leave-one-out cross-validation.

The kriging variance at every iteration of the cross-validation is indicated by the colour of the station that was left out. Stations with lower values in blue indicate that kriging maps with lower variance are produced when those stations are removed. All iterations had relatively low variance.

Some individual stations that were highlighted for potential inclusion or removal are shown in Table 7.

Table 7. Stations highlighted by kriging analysis for potential removal or retention.

Potential removal	Potential retention
F121	F11
F125	F140
F127	F16
F128	F165
F13	F18
F134	F19
F142	F2
F15	F20
F157	F21
F158	F220
F159	F23
F175	F24
F177	F247
F189	F5
F190	F52
F193	F6
F195	F62
F207	F84
F213	
F218	
F222	
F226	
F239	
F240	
F246	
F252	
F31	
F54	
F57	
F60	
F78	
F81	

F88
F89
F99

4.3 Water quality variables

This section discusses the results obtained from the application of the proposed approach to rationalize water quality variables measured at Lake Winnipeg. The proposed approach is applied to each of the seven selected sampling sites. Results are presented in two subsections. The first subsection presents the results obtained from the application of the rationalization approach to sampling site MB05SBS126. The second subsection presents the results obtained from the empirical experiment to identify the best technique to use in reconstituting information about discontinued variables.

4.3.1 Rationalisation results

Using cluster analysis (CA) and the criteria developed for assessment of the correlation coefficient (d_m and d_v), groups of highly correlated water quality variables were identified. Figure 6 shows the cluster tree for MB05SBS126, 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 20 and 21) was applied. Results show that $d_m = 0.75$ and $d_v = 0.34$. Thus, the d_v criterion was applied as indicated by the horizontal line at a cut-off distance of 0.34 (Figure 11). Using the higher value (d_m instead of d_v), it cannot be guaranteed 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 were divided into five clusters, three of which were single-variable clusters. These three variables should be continuously monitored because the information they provide cannot be estimated from other variables. Figures representing the CA for the seven selected sampling sites are presented in Appendix 5.

Two other clusters were considered for further analysis. $CaCO_3$, HCO_3, Cl, EC, H.$CaCO_3$, Na, SO_4 and TDS form the first cluster. The second cluster consisted of the TSS and Turb. At least one variable from each of these two clusters should be continuously measured. Given that the two clusters contained ten water quality variables, the maximum number of variables to discontinue would be eight.

Within each multiple-variable cluster, each variable was assumed to be discontinued, and its best auxiliary variable was identified based on Equation 3. The second cluster consisted of only two variables. In this case, each variable was an auxiliary for the other variable, and only one variable could be discontinued from this cluster, i.e., either Turb or TSS. The first cluster consisted of eight variables. Therefore, each variable was considered discontinued, and Equation 3 was applied to identify the best auxiliary from the other seven variables in the same cluster.

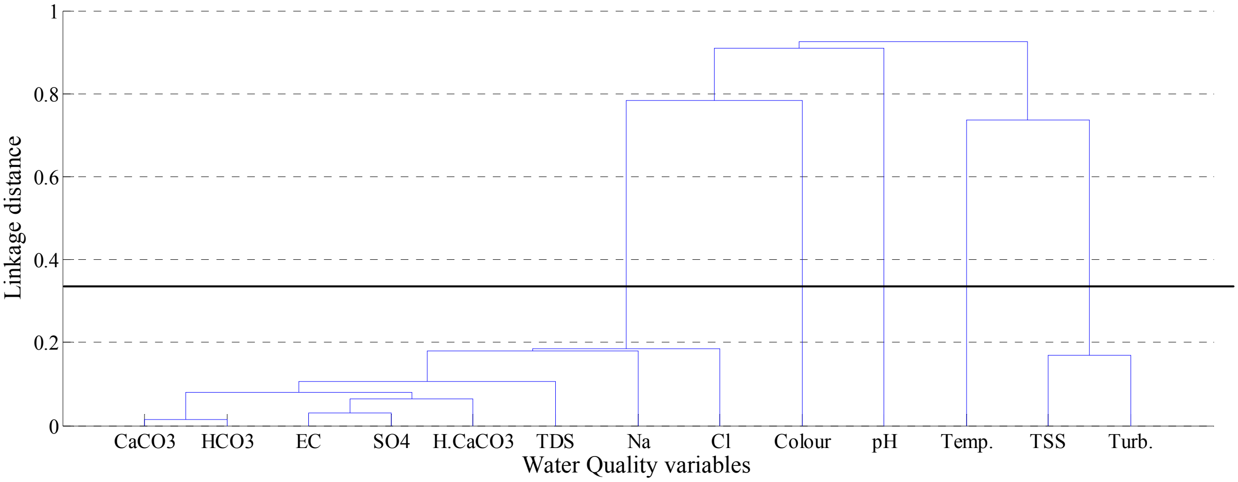


Figure 11. Cluster tree for water quality variables at MB05SBS126

Using Equation 3, Cl was found to be the best auxiliary for HCO_3, TDS, $CaCO_3$, SO_4, EC and H.$CaCO_3$. HCO_3 was the best auxiliary for Cl, and TDS was the best auxiliary for Na. Table 8 shows all of the alternatives; assuming each variable was discontinued (first column), the best auxiliary variable was identified (second column). The correlation between the discontinued variable and its best auxiliary, the number of available observations, the estimated variance after n_2 years obtained from the application of

Equation 3, and the performance index are presented in the remaining columns, respectively.

If one variable is to be discontinued from the list of variables being measured at this site, the aggregated performance index is applied (Equation 23). The variables listed in Table 8 are ranked based on the performance index (sixth column). If only one variable is to be discontinued, from a statistical point of view it can be HCO_3, Cl or TDS. Discontinuation of any of these three variables leads to an equal performance index. Table 8 provides the rank of the variables to be discontinued from a statistical point of view. Thus, it can be used along with other criteria to choose which variable to discontinue. Other criteria may include stakeholders' preference, the significance of the variable in specific studies or laboratory cost analysis. Tables represent variables to be discontinued and its best auxiliary variables for the seven selected sampling sites are presented in Appendix 5.

Table 8. Variable to discontinue and its best auxiliary variable

Discontinued variable	**Best auxiliary variable**	***r***	**Number of samples**	$Var(\hat{\mu})$	I_a
HCO_3	Cl	0.88	32	0.09	1.93
Cl	HCO_3	0.88	32	0.09	1.93
TDS	Cl	0.88	32	0.09	1.94
$CaCO_3$	Cl	0.90	32	0.10	1.95
SO_4	Cl	0.92	32	0.11	1.98
Na	TDS	0.87	30	0.13	1.99
EC	Cl	0.93	32	0.12	1.99
$H.CaCO_3$	Cl	0.91	30	0.14	2.01
TSS	Turb.	0.91	27	0.18	2.05
Turb.	TSS	0.91	27	0.18	2.06

When more than one variable are to be discontinued, the information index was applied to rank the different combinations. For example, if two water quality variables must be discontinued, $C(13,2)=78$ different combinations may be considered. Table 9 shows the first 15 combinations ranked based on the information index for the cases of two variables and three variables to be discontinued. Similarly, the proposed approach can provide monitoring network designers with a variety of scenarios (combinations) for any number of variables to be discontinued. Tables representing different combinations of two, three, four and five variables to be discontinued from each of the seven selected sampling sites are presented in Appendix 5. For the seven selected sampling sites, cluster analysis indicated that the levels of correlation between water quality variables varied from site to site. Consequently, groups of highly associated variables also varied.

Table 9. Combinations of variables to discontinue

Two variables		I_a%	Three variables			I_a%
HCO_3	TDS	1.32	HCO_3	TDS	$CaCO_3$	1.10
HCO_3	$CaCO_3$	1.33	HCO_3	TDS	SO_4	1.12
TDS	$CaCO_3$	1.33	HCO_3	$CaCO_3$	SO_4	1.13
HCO_3	SO_4	1.34	TDS	$CaCO_3$	SO_4	1.13
TDS	SO_4	1.35	HCO_3	TDS	EC	1.13
Cl	Na	1.35	HCO_3	$CaCO_3$	Na	1.13
HCO_3	Na	1.35	HCO_3	$CaCO_3$	EC	1.14
$CaCO_3$	SO_4	1.35	HCO_3	TDS	$H.CaCO_3$	1.14
HCO_3	EC	1.36	TDS	$CaCO_3$	EC	1.14
TDS	EC	1.36	HCO_3	SO4	Na	1.15
$CaCO_3$	Na	1.36	HCO_3	$CaCO_3$	$H.CaCO_3$	1.15
HCO_3	$H.CaCO_3$	1.37	TDS	$CaCO_3$	$H.CaCO_3$	1.15
$CaCO_3$	EC	1.37	HCO_3	SO_4	EC	1.16
TDS	$H.CaCO_3$	1.37	TDS	SO_4	EC	1.16
SO_4	Na	1.38	$CaCO_3$	SO_4	Na	1.16

For the case when only one variable should be discontinued from the seven sampling sites, the discontinued variable should be $CaCO_3$, HCO_3, Na or $H.CaCO_3$. These variables appear for possible discontinuation in all of the sites. However, a variable to discontinue can be chosen from some of the sampling sites, while another variable can be discontinued from the remaining sites. For instance, SO_4 could be discontinued from the five sites representing the narrows and the south basin, and Na could be discontinued from the two sites representing the north basin. Another alternative is to discontinue TDS from the two sites representing the narrows and discontinue $CaCO_3$ or HCO_3 from the five sites representing the north and south basins. Similarly, for discontinuation of two or more variables, the same combination can be chosen from all the sampling sites or different combinations can be chosen from different groups of sites.

It should be emphasised that the proposed approach provides the decision maker with the optimal combinations of variables to discontinue from a statistical point of view. However, various criteria could be integrated when deciding which variables to discontinue and which variables should be continuously measured. Such criteria may include the presence of the variable in water quality standards, preference, significance for particular studies, or laboratory cost analysis.

4.3.2 Information transfer

Figures 12 and 13 summarise the results obtained for the ratio *U*. Figure 12 shows the *U* distribution for the estimation of the statistical moments, and Figure 13 shows the *U* distribution for the estimation of different non-exceedance percentiles. From the previous step, 55 pairs of variables were considered for all of the sampling sites. Boxplots in Figures 12 and 13 were constructed from 4,620 records. The accuracy of each approach can be judged by the degree of dispersion in the boxplots, by the closeness of the median to a value of one, and by the symmetry of the boxplot around the value of one (Hirsch, 1982; Vogel and Stedinger, 1985).

In Figure 12, boxplots represent the distribution of *U* for the estimation of the mean and the standard deviation. For the estimation of the mean, the regression and KTRL2 techniques led to median values of *U* equal to one. Boxes were symmetric around one and had nearly the same dispersion. Figure 12 shows that the regression technique tended to underestimate the standard deviations. The cross-validation showed that approximately 75% of the regression standard deviations were lower than those estimated from the observed values, with a median value of 0.86. With the KTRL2 technique, the median value was one. The standard deviation boxplots had similar dispersions, but the KTRL2 standard deviation boxplot was more symmetric around one.

Regression

KTRL2

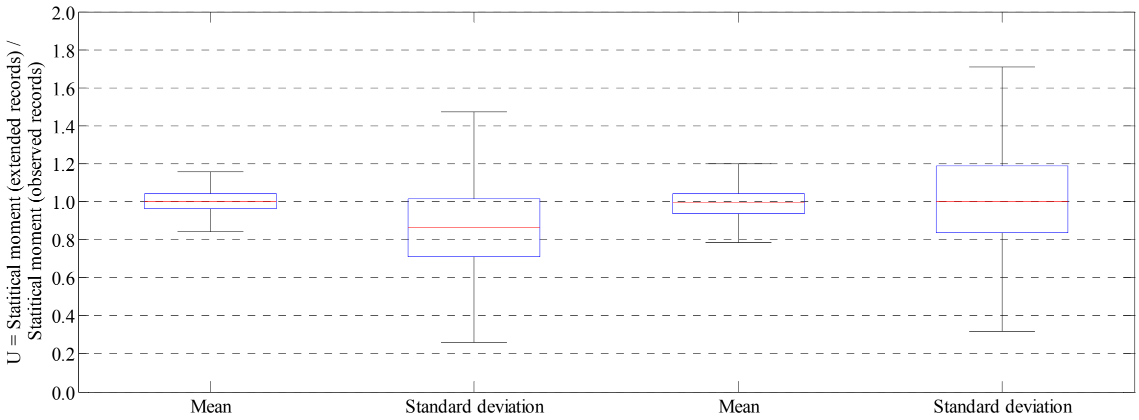


Figure 12. Box plots of the ratio *U* for the mean and standard deviation

Figure 13 shows that when using the regression technique, the median values of *U* for low percentiles were higher than one, while those corresponding to high percentiles were lower than one. However, the median values of *U* for the KTRL2 percentiles were very close to one. The *U* median values ranged between 0.97 and 1.06 for the regression, and between 0.98 and 1 for the KTRL2 technique. In general, the KTRL2 boxplots were symmetric around one and showed slightly more dispersion than those corresponding to the regression. These results suggest that the regression tended to overestimate low percentiles and underestimate high percentiles. The KTRL2 technique reduced the bias exhibited by the regression.

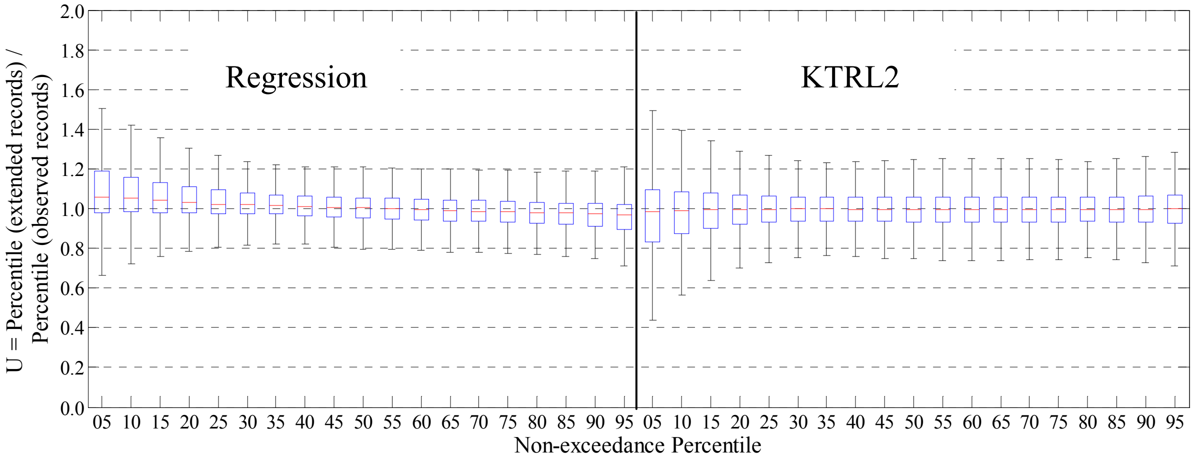


Figure 13. Box plots of the ratio *U* for different percentiles

Figure 14 illustrates the *BIASr* exhibited by the two extension techniques for estimating the water quality concentration percentiles. Results indicate that the KTRL2 *BIASr* values were closer to zero when compared to the regression *BIASr* values. The curves corresponding to the KTRL2 and the regression method intersect at the median, as the regression approach overestimated low percentiles and underestimated high percentiles.

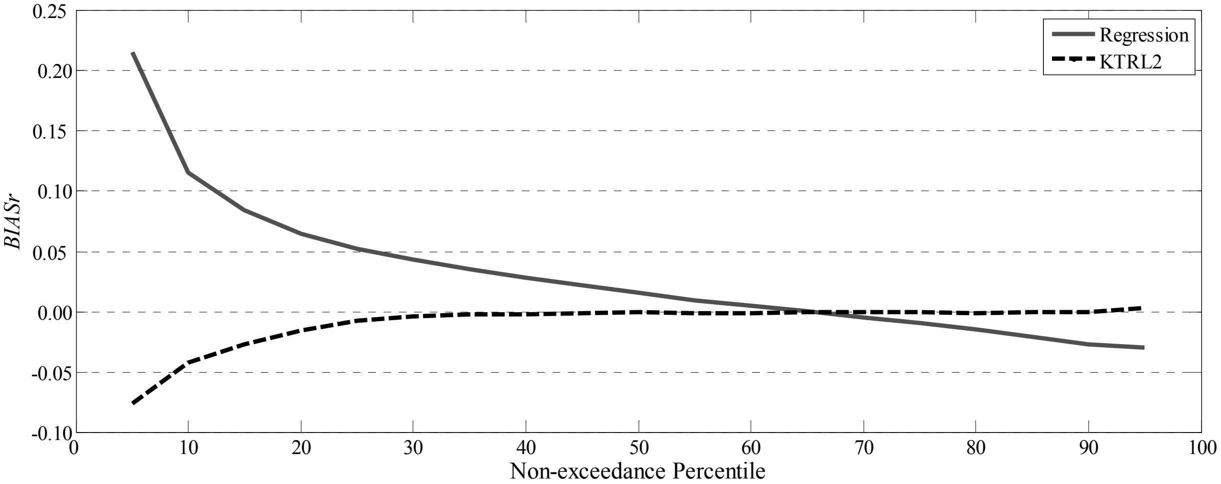


Figure 14. *BIASr* of the tested extension techniques in estimating various percentiles

Figure 15 illustrates the *RMSEr* exhibited by the two extension techniques for estimating the water quality concentration percentiles. There is no significant difference between the *RMSEr* values corresponding to the regression and those corresponding to the KTRL2 technique. Figures 13 and 14 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.

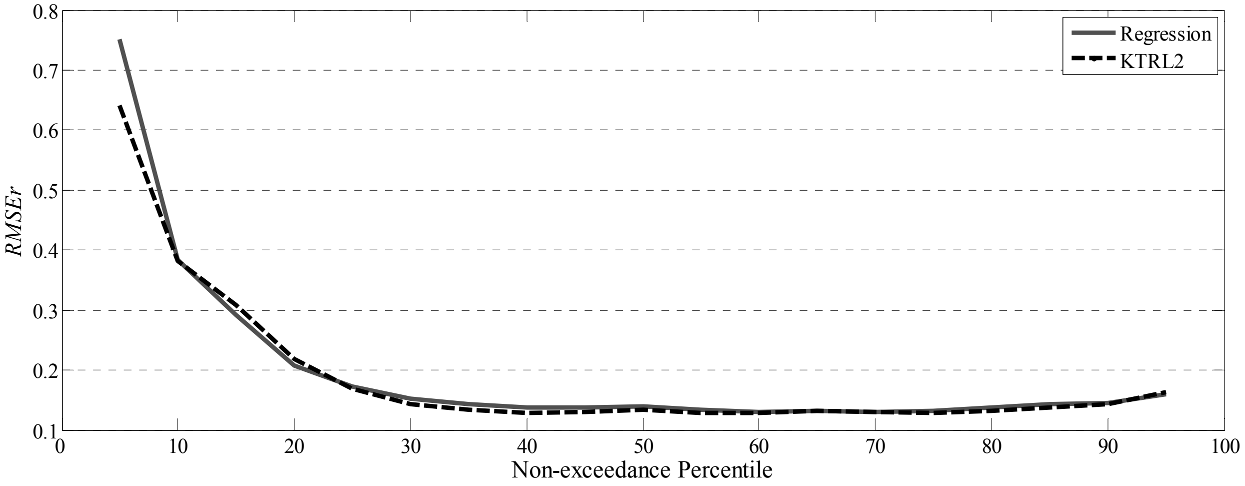


Figure 15. *RMSEr* of the tested extension techniques in estimating various percentiles

The *BIASr* and the *RMSEr* were also computed for the estimation of the mean value and standard deviation as well as for individual water quality records, where each error represents the difference between the extended and actual water quality records. Table 10 shows that for the estimation of the mean values, regression and KTRL2 had similar *BIASr* and *RMSEr* values. The *RMSEr* values determined using both record-extension techniques were similar for the estimation of the standard deviation and individual records. However, the *BIASr* value derived from the KTRL2 method was lower than that derived from the regression for the estimation of the standard deviation and for the estimation of individual values.

The results in this subsection are summarised as follows. The two extension techniques produced extended records that were unbiased in the mean. Furthermore, regression substantially reduced the variability, and KTRL2 tended to preserve variability. The regression technique underestimated high concentration values and overestimated low values, while the KTRL2 technique tended to reduce the bias in the estimation of both high and low concentration values. The KTRL2 technique produced extended records that preserved both the high and low percentiles relatively well.

Table 10. Average error measures for record-extension techniques

Statistic	**Mean value**		**Standard deviation**		**Individual records**	
Technique	**Regression**	**KTRL2**	**Regression**	**KTRL2**	**Regression**	**KTRL2**
BIASr	0.006	-0.008	-0.104	0.041	0.046	-0.003
RMSRr	0.114	0.113	0.326	0.341	0.209	0.207

4.4 Sampling frequency

Two approaches were employed to assess the sampling frequency for water quality variables at the seven sampling sites. These were the semi-variogram and confidence interval approaches. Results obtained from the application of the semi-variogram approach are presented in the following subsection. Results obtained from the application of the confidence interval approach are presented in the second subsection.

4.4.1 Semi-variogram approach

The variogram analysis was applied for water quality variables measured at the seven selected sampling sites. However, the irregular sampling frequency and the limited number of records available did not allow a good fitting of experimental variogram models. For example, only 32 observations are available for $CaCO_3$ measured at MB05RAS078 in the period from August 1999 to October 2008. Thus, in duration of 111 months, sampling took place in 32 months only.

Two parameters should be identified to carry out variogram analysis. These are the active lag distance and the lag class interval. The active lag distance specifies the range over which autocorrelation will be calculated. The minimum active lag distance is the minimum distance between adjacent records in the data set (one month in our case), while the maximum distance is the maximum distance between any two points (111 months in our case). The lag class distance interval defines how pairs of points will be grouped into lag classes. Each point in a variogram graph represents the average autocorrelation for a single lag class, which is a group of pairs separated by a certain lag class distance interval. This interval is uniformly distributed across the active lag distance.

Figure 16 shows the experimental semi-variogram for $CaCO_3$ measured at MB05RAS078. In this analysis, the active lag distance is 100 and the lag interval is equal to 1.

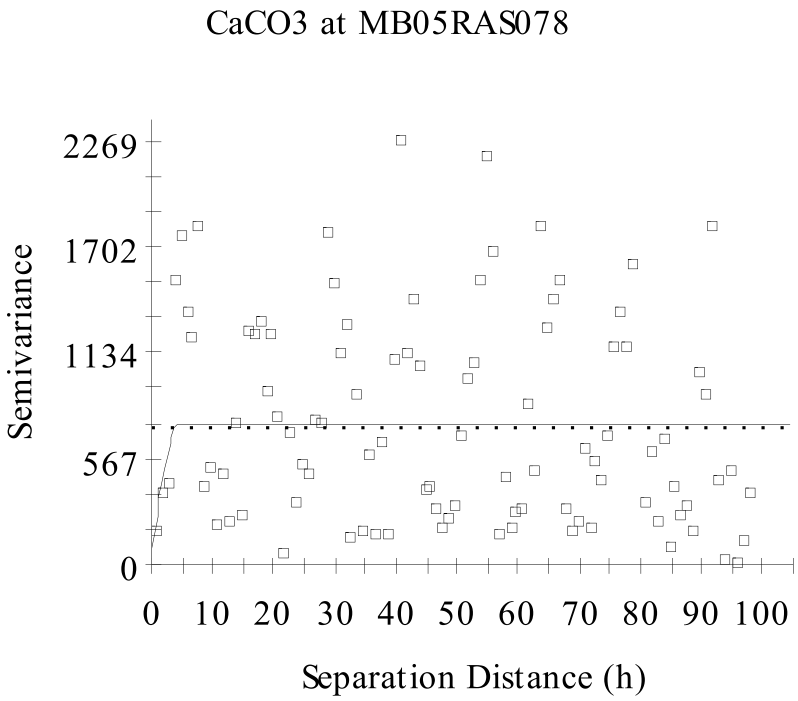


Figure 16. Variogram for CaCO3 measured at MB05RAS078 (lag class interval =1)

In Figure 16, small boxes indicate the variance calculated at each lag, the doted horizontal line indicate the variance calculated from the observed 32 records and the solid line represents the experimental semi-variogram model. Although, the experimental model indicates that the variance is stable after separation distance equal to 5 months, but the small boxes represent the variance at each lag distance indicate a high variation in the variances computed at each lag distance. This is mainly due to the limited number of available observations and the irregular sampling frequency applied. For some lag distances the number of available cases represents this lag may be very limited to represent the variance at this lag distance.

Consider the lag class interval is equal to 2, 3 or 4, this increases the number of cases available at each lag class interval. Figure 17 shows the variogram analysis for CaCO3 measured at MB05RAS078 when considering the active lag distance is equal to 40 and lag class interval is equal to 4.

From Figure 17, the experimental variogram indicates that the variance is stabilizing after lag 5. Comparing Figures 16 and 17, the variation in the variances computed for each lag class interval when the class interval is equal to 4 is less than that when the lag class interval is equal to 1. Although both experimental semi-variogram models indicate that the variance stabilize at lag 5, but when the class interval is equal to 4 the experimental semi-variogram model overestimates the variance estimated from the data (Figure 17), when the class interval is equal to 1, the model estimation of the variance is more accurate.

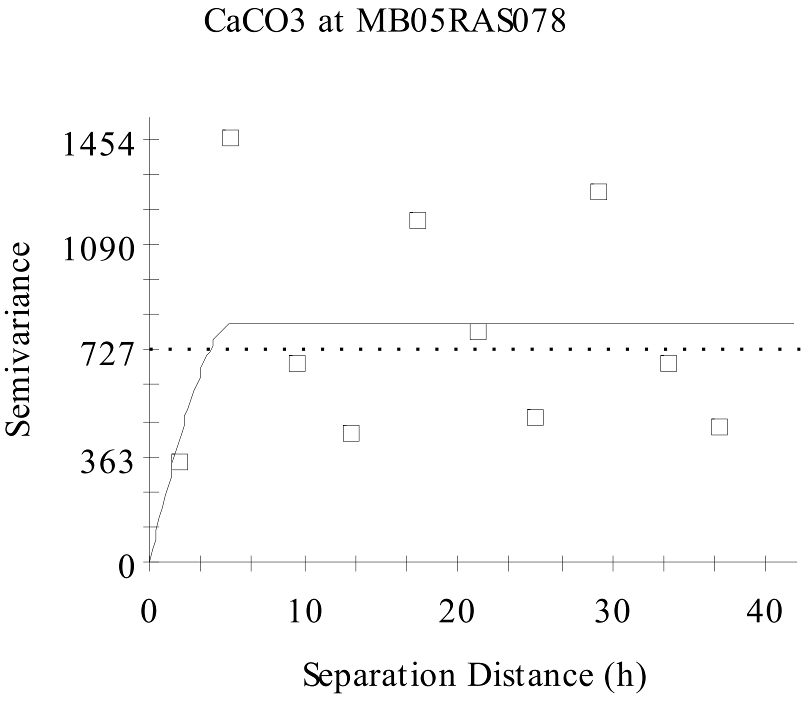


Figure 17. Variogram for CaCO3 measured at MB05RAS078 (lag class interval =4)

The variogram analysis is carried out for the 13 water quality variables measured at each of the seven selected sampling sites. It is concluded that the limited number of observations and irregularity of the sampling frequency affect seriously the variogram analysis. In general, the variogram analysis indicates that the variance is almost constant for large lag distances. However, by increasing the number of available records, the lag distance at which variance stabilize may be specified accurately. Thus, with available limited data assessment of sampling frequency using variogram analysis is not possible.

4.4.2 Confidence Interval approach

The confidence interval approach, as described in the methodology, was applied for assessment of the sampling frequency of water quality variables measured at the seven selected sampling sites. The estimation of the variance of the stationary component requires removal of the trend and seasonal components if they exist. Due to the limited number of records and the irregularly applied sampling frequency, only the trend assessment was used in this study. Water quality variables that showed a significant trend were de-trended before applying the confidence interval approach. For water quality variables showing no significant temporal trend, the approach was applied using estimates from the raw data. In addition, due to the limited number of records and the irregular sampling frequency, it was assumed that water quality observations were independent in time. Rosner`s test was applied to verify outliers labelled by boxplots. Outliers detected by Rosner`s test were removed.

Using the confidence interval as a criterion to assess the sampling frequency requires identification of the desired half-confidence interval around the mean. As described in the methodology, half the confidence interval is considered the error for estimating the mean. In this study, this error was expressed as a percentage of the sample mean. Several error percentages are considered, from 10% to 60% of the sample mean. The error percentages were plotted against the number of samples obtained by applying Equation 30. Three confidence levels were considered: 95%, 90% and 80%.

Figure 18 shows the number of samples required to estimate the mean value of $CaCO_3$ measured at MB05SBS126 for the three confidence levels considered. For smaller error percentages, a dramatic reduction in the number of samples caused a minor increase in the expected error. For instance, in the case of the 95% confidence level, reducing the number of samples from 27 to 13 increased the expected error for estimating the $CaCO_3$ mean value from 10% to 15%, while further reduction in the number of samples to eight increased the expected error to 20%. On the other hand, for relatively large expected error percentages, a minor reduction in the number of samples seriously affected the expected error percentage. For instance, when the number of samples was reduced from 7 to 5, the expected error increased from 25% to 35%.

In addition, Figure 18 shows the effect of the confidence level on the number of samples required. For the range of low error percentages, the number of samples is reduced dramatically when the confidence level is reduced. For higher error percentages, the reduction in the number of samples is not that significant. For example, when the confidence level was reduced from 95% to 90%, the number of samples was reduced from 27 to 19, with 10% expected error; for the 80% confidence level, the number of samples required was 12. However, when the expected error was 30%, the number of samples required was 6, 4 and 3 for the 95%, 90% and 85% confidence levels, respectively. Figures illustrating the number of samples versus the expected error percentages for water quality variables measured at the seven selected sampling sites are presented in Appendix 6.

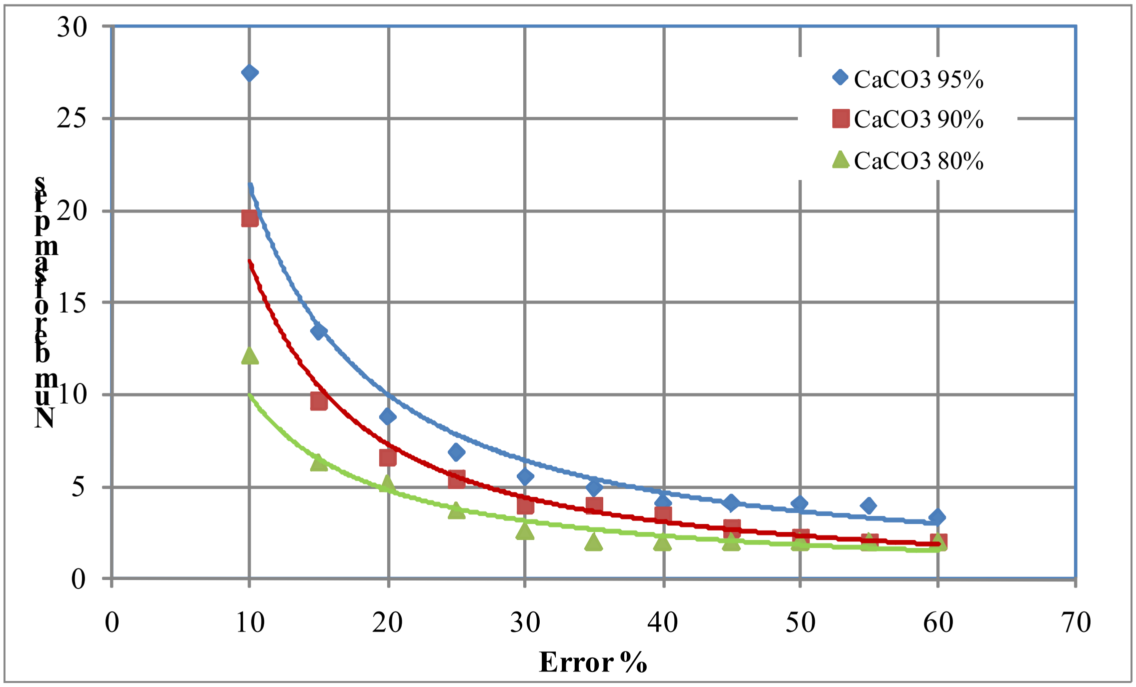


Figure 18. Number of samples vs expected error for $CaCO_3$ measured at MB05SBS126

To this point, assessment of the sampling frequency was performed for each variable at each of the selected sampling sites. If uniform sampling is to be performed, Tables 11-14 show the error expected for the estimation of mean values for 6, 12, 26 and 52 samples per year. These numbers represent sampling on a bimonthly, monthly, biweekly and weekly basis, respectively. For $CaCO_3$, HCO_3, Cl, Colour, EC, H.$CaCO_3$, pH, Na, SO_4 and TDS, sampling on a bimonthly basis would be acceptable. However, sampling on a monthly basis would allow accurate estimation of mean values. For Temp, TSS and Turb, sampling on a biweekly basis would allow acceptable estimation of the mean values.

Table 11. Error percentage expected based on collecting 6 samples per year

Sampling site	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
MB05RAS078	32	32	50	39	39	32	6	47	57	58	35	67	78
MB05RBS003	19	18	32	37	21	23	4	29	41	70	20	55	47
MB05SAS004	7	7	17	29	7	9	4	15	30	69	23	81	71
MB05SBS126	26	23	37	33	31	31	5	43	50	76	26	65	76
MB05SCS005	22	23	44	31	32	30	5	42	49	78	30	54	58
MB05SHS004	7	7	20	44	10	15	3	11	9	51	9	77	65
MB05SHS014	6	8	19	36	8	11	4	11	10	73	9	71	67

Table 12. Error percentage expected based on collecting 12 samples per year

Sampling site	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
MB05RAS078	20	20	32	24	25	20	4	30	36	37	22	42	49
MB05RBS003	12	11	20	23	13	15	2	18	26	44	13	35	30
MB05SAS004	4	4	11	18	5	6	2	9	19	44	15	51	45
MB05SBS126	16	15	23	21	19	19	3	27	31	48	16	41	48
MB05SCS005	14	15	28	19	20	19	3	26	31	49	19	34	37
MB05SHS004	5	5	12	28	6	10	2	7	6	32	6	48	41
MB05SHS014	4	5	12	22	5	7	3	7	6	46	6	44	42

Table 13. Error percentage expected based on collecting 26 samples per year

Sampling site	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
MB05RAS078	13	13	20	16	16	13	2	19	23	24	14	27	32
MB05RBS003	8	7	13	15	9	9	1	12	17	28	8	22	19
MB05SAS004	3	3	7	12	3	4	2	6	12	28	9	33	29
MB05SBS126	10	9	15	13	12	12	2	17	20	30	11	26	31
MB05SCS005	9	9	18	12	13	12	2	17	20	31	12	22	23
MB05SHS004	3	3	8	18	4	6	1	5	4	21	4	31	26
MB05SHS014	2	3	8	14	3	4	2	4	4	29	4	29	27

Table 14. Error percentage expected based on collecting 52 samples per year

Sampling site	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
MB05RAS078	9	9	14	11	11	9	2	13	16	16	10	19	22
MB05RBS003	5	5	9	10	6	6	1	8	12	20	6	15	13
MB05SAS004	2	2	5	8	2	2	1	4	8	19	6	23	20
MB05SBS126	7	7	10	9	9	9	1	12	14	21	7	18	21
MB05SCS005	6	6	12	9	9	8	1	12	14	22	8	15	16
MB05SHS004	2	2	5	12	3	4	1	3	3	14	2	21	18
MB05SHS014	2	2	5	10	2	3	1	3	3	20	3	20	19

Subsequently, for the case where the monitoring objectives are the determination of ambient water quality conditions and the assessment of annual trends, biweekly sampling is recommended. Following this design, the number of samples to be collected is 26, with analysis of the 13 water quality variables taking place in 12 of these samples, while the Temp, TSS and Turb are measured in all 26 samples. If the monitoring budget does not allow 26 samples per year, monthly sampling would be acceptable.

Identification of the sampling frequency and laboratory analyses frequency is based on the monitoring objectives, cost per sample, cost per variable and cost per sampling trip. The monitoring objective as well as cost criteria would help to identify the affordable sampling frequency for each variable and each sampling site.

If the monitoring budget does not support a high sampling frequency, a possible solution to improve the network accuracy would be to reduce either the number of sampling sites or the number of water quality variables to measure. This adjustment would help to address the trade-off between the sampling frequency and the number of sampling sites and variables to measure. Thus, the decision would be either to increase the number of samples in favour of more monitoring sites or variables to measure, or to keep more sampling sites and variables, while reducing the number of samples.

5 CONCLUSIONS

5.1 Discussion

In this study, different statistical approaches for the assessment of Lake Winnipeg monitoring activities were applied. The assessment of the number and configuration of isotope stations was performed using co-kriging. Co-kriging allows the evaluation of overall model performance, the identification of a parsimonious predictive model and the identification of individual stations that contribute unique information. The best kriging model was developed using two dimensions from an NMDS analysis of UTM coordinates and distance to the Red River. Good kriging models could be developed for both δ^2H and δ^{18}O. The spatial distribution of δ^{18}O was easier to model than for δ^2H. Some individual stations were identified as contributing unique information while others were identified as usually being redundant with their neighbours.

The assessment and selection of the optimal combination of water quality variables to measure was performed using criteria developed from record-augmentation procedures and CA, allowing identification of highly correlated variables. An information performance index was applied to identify the optimal combinations of variables to be continuously measured or discontinued. Simple linear regression and the KTRL2 technique were applied to reconstitute information about discontinued variables. The confidence interval width around the mean was used as a criterion to assess the sampling frequency, assuming that the monitoring objectives were the determination of ambient water quality conditions and assessment of annual trends.

The proposed approach for assessment of the number of water quality variables to measure provides the decision maker with the optimal combinations of variables to discontinue from a statistical point of view. However, various qualitative and managerial criteria can be integrated when deciding which variables to discontinue and which variables to continuously measure.

The empirical experiment performed to compare the regression and KTRL2 methods with respect to reconstitution of information about discontinued variables shows that regression falls substantially short of achieving the desired result of creating a realistic extended record. Regression 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 KTRL2 is better than regression. Regression estimates substantially underestimated the variance. Consequently, the frequency of extreme events, such as the exceedance of standards, would be underestimated by regression. On the other hand, KTRL2 reduced the bias exhibited by regression for the estimation of the variance and reduced the bias in the estimation of extreme percentiles.

The assessment of the sampling frequency for the 13 water quality variables considered in this study showed that monthly sampling would allow acceptable estimation of the mean values for most of the variables. Due to the relatively high variability in the Temp, TSS and Turb records, more samples may be required for more accurate estimation of the mean values of these variables.

5.2 Future analysis

There is no set of stations that should unambiguously be removed from or retained from the network. The ideal number of stations depends on other objectives outside this analysis such as the incremental budget implications of a station and the level of spatial resolution that is desired. One analysis that can assist with this decision is the determination of the information content at various numbers and configurations of stations. Unfortunately, the number of possible combinations of stations is very large and practically impossible to evaluate.

One solution is to break the set of stations into small groups of proximate stations. For each group, the kriging variance can be calculated for a subset of combinations. This will give an indication of the information at various numbers of stations. It will also identify the ideal configuration for a few combinations.

5.3 Recommendations

Given that the design of water quality monitoring networks is based mainly on the monitoring objectives, it is recommended that these objectives are reviewed and precisely defined for the Lake Winnipeg monitoring network.

It may be more appropriate to identify monitoring objectives for each sampling site. For instance, assessment of temporal trends may be one of the objectives; however, assessment of trends may be performed only at specific locations.

Given the irregularly applied sampling frequency, autocorrelation analysis and seasonal characteristics of the water quality variable cannot be assessed. It is recommended that intensive sampling be applied this year for a few sampling sites that represent the three main regions (the narrows, the south and north basins) in the lake. It is believed that some information on optimal sampling frequency could be gained by analyzing the data gathered by the ADCP deployed in Lake Winnipeg. In addition, intensive sampling should be applied at major inflows, such as the Winnipeg River, Saskatchewan River and the Red River, as well as the major outflow to the Nelson River. Representative smaller rivers should also be included.

6 REFERENCES

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

Anderson AJB. 1971. Ordination Methods in Ecology. Journal of Ecology 59(3): 713-726.

Armstrong M. 1998. Basic linear geostatistics. New York: Springer.

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.

Cressie N. 1990. The origins of kriging. Mathematical Geology 22(3): 239-252.

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.

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

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

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

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.

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.

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.

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.

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

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

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

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.

Journel AG, Huijbregts CJ. 1978. Mining Geostatistics. London: Academic Press.

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.

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.

Khalil B., T.B.M.J. Ouarda, A. St-Hilaire and F. Chebana (2010a). A statistical approach for the rationalization of water quality indicators in surface water quality monitoring networks, Journal of Hydrology, (in press) doi:10.1016/j.jhydrol.2010.03.019.

Khalil, B., T.B.M.J. Ouarda and A. St-Hilaire (2010b). Comparison of record-extension techniques for water quality variables, Submitted to Journal of Hydrology.

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.

Legendre P, Legendre L. 1998. Numerical ecology. Amsterdam: Elsevier

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

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.

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

Mann, H. B. (1945). Nonparametric test against trend, Econometrica, 13, 245-259.

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

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

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

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 Quebec, Quebec, Canada, 75 p.

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

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.

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.

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.

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.

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.

Weisberg S. 2005. Applied Linear Regression, third edition. Hoboken NJ: John Wiley.

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

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.

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

Appendix 1. Kriging

Kriging predicts the value of a random variable at an unmeasured location using the spatial dependency between points. Experimental semivariograms quantify this spatial dependency through:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z_{x_{i+h}} - z_{x_i})^2 \tag{1}$$

Where *N(h)* is the number of observations pairs separated by distance *h* with values z_{x_i}.

A theoretical model is fitted to the semivariogram. The kriging estimator $\hat{z}(x_0)$ at a given point x_0 is the best linear unbiased estimator of mean parameters (Cressie 1990):

$$\hat{z}(x_0) = \sum_{i=1}^{N(h)} \lambda_i z(x_i) \tag{2}$$

Where $z(x_i)$ is the weighted average of the observed values at the i^{th} location, and λ_i is the unknown weight for the measured value at the i^{th} location. The weights fulfill the unbiasedness condition:

There is an additional constraint in ordinary kriging:

$$\sum_{i=1}^{N(h)} \lambda_i = 1 \tag{3}$$

This requires an additional parameter in the system of weights called the Lagrange multiplier which accounts for this constraint in the system of equations.

Ordinary kriging assumes local stationarity of the mean, i.e. the mean is invariant under translation (Armstrong 1998), and an unknown, constant trend component.

Appendix 2. Principal Component Analysis

PCA creates linear combinations of the original variables x_i (Principal Components) that extract the maximum amount of variance and are orthogonal to one another.

$$PC1 = \sum_{i=1}^{p} a_{1i} x_i \tag{1}$$

a_{1i} is the weight of variable x_i. The variance explained by PC_1 is maximised under the constraint that the sum of the squared weights is equal to one.

$$\sum_{i=1}^{p} a_{1i}^{2} = 1 \tag{2}$$

The greatest variance of any projection of the data are explained by the first principal component with successive components explaining decreasing proportions of variance. It creates a multivariate space in which the original variables (or cases) can be projected and their relative positions visualized in a smaller number of dimensions than the original dataset. PCA is based on Pearson's *r*, which captures only the linear relationship between variables.

Appendix 3. Non-Metric Multidimensional Scaling

Non-Metric Multidimensional Scaling (NMDS) is a rank-based ordination technique that is able to detect non-linear relationships. NMDS tries to maintain dissimilarity relations, in a low dimension (usually two or three) configuration, or plot, of points. The primary assumption of NMDS is that the data are independent and identically distributed (Legendre and Legendre 1998).

Stress is the measure of how well sample relations in the original dissimilarity matrix are preserved in the low-dimensional configuration. The configuration is iteratively optimised in a direction of decreasing stress. I used the most popular definition of stress.

$$Stress = \sqrt{\frac{\sum_j \sum_k \left(d_{jk} - \hat{d}_{jk}\right)^2}{\sum_j \sum_k d_{jk}^2}} \tag{1}$$

where $\hat{d}_{jk}$ is the predicted distance between points *j* and *k*, d_{jk} is the distance between points *i* and *j* (from the dissimilarity matrix), and d_{jk}^2 is a scaling factor.

Appendix 4. Data Preliminary analyses

Location MB05RAS078

Table 15 Location MB05RAS078, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	32	32	27	32	33	31	33	31	29	22	31	24	32
Mean	85.57	104.49	8.00	31.59	250.76	100.90	7.77	9.78	31.48	13.69	168.29	8.29	11.73
Median	83.00	101.30	8.14	30.00	242.00	103.00	7.79	9.80	28.80	14.55	158.00	6.00	9.65
Std. Deviation	27.10	33.04	4.04	12.18	98.58	32.82	.44	4.65	18.08	8.02	59.18	5.54	9.21
Variance	734.36	1091.76	16.29	148.38	9718.38	1077.05	0.20	21.60	326.87	64.26	3502.21	30.65	84.89
Skewness	0.283	0.285	0.235	0.640	0.095	-0.064	-0.31	0.045	0.342	-0.634	0.047	1.116	0.967
Std. Error of Skewness	0.414	0.414	0.448	0.414	0.409	0.421	0.409	0.421	0.434	0.491	0.421	0.472	0.414
Kurtosis	-0.121	-0.113	-0.146	0.335	-0.868	-0.578	-0.553	-0.638	-0.252	-0.633	-0.866	0.351	0.001
Std. Error of Kurtosis	0.809	0.809	0.872	0.809	0.798	0.821	0.798	0.821	0.845	0.953	0.821	0.918	0.809
Range	108.00	131.80	15.70	50.00	340.00	125.00	1.77	16.70	71.78	24.00	223.00	19.00	30.70
Minimum	42.00	51.20	1.50	10.00	106.00	46.00	6.74	2.60	4.02	0.00	65.00	3.00	2.30
Maximum	150.00	183.00	17.20	60.00	446.00	171.00	8.51	19.30	75.80	24.00	288.00	22.00	33.00

Table 16 Location MB05RAS078, WQ variables Kolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	Absolute	0.099	0.099	0.127	0.115	0.128	0.111	0.136	0.112	0.095	0.173	0.089	0.202	0.165
	Positive	0.099	0.099	0.127	0.115	0.128	0.111	0.102	0.112	0.095	0.136	0.089	0.202	0.165
	Negative	-0.078	-0.077	-0.078	-0.070	-0.080	-0.074	-0.136	-0.061	-0.083	-0.173	-0.087	-0.170	-0.153
Kolmogorov-Smirnov Z		0.560	0.562	0.659	0.652	0.735	0.619	0.782	0.624	0.509	0.813	0.495	0.991	0.932
Asymp. Sig. (2-tailed)		0.913	0.910	0.778	0.788	0.653	0.838	0.574	0.831	0.958	0.524	0.967	0.280	0.350

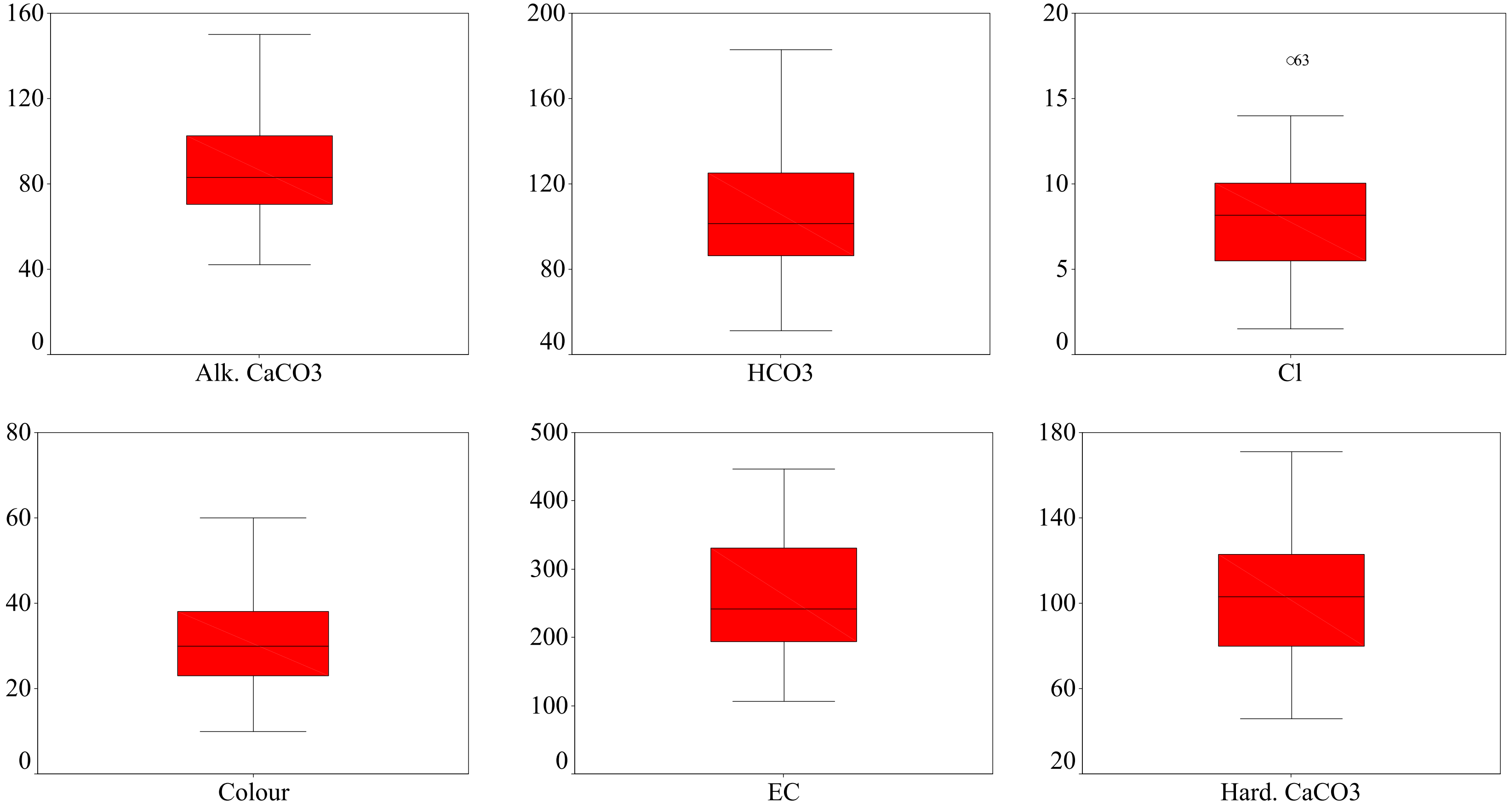


Figure 19. Location MB05RAS078, WQ variables Box-plots

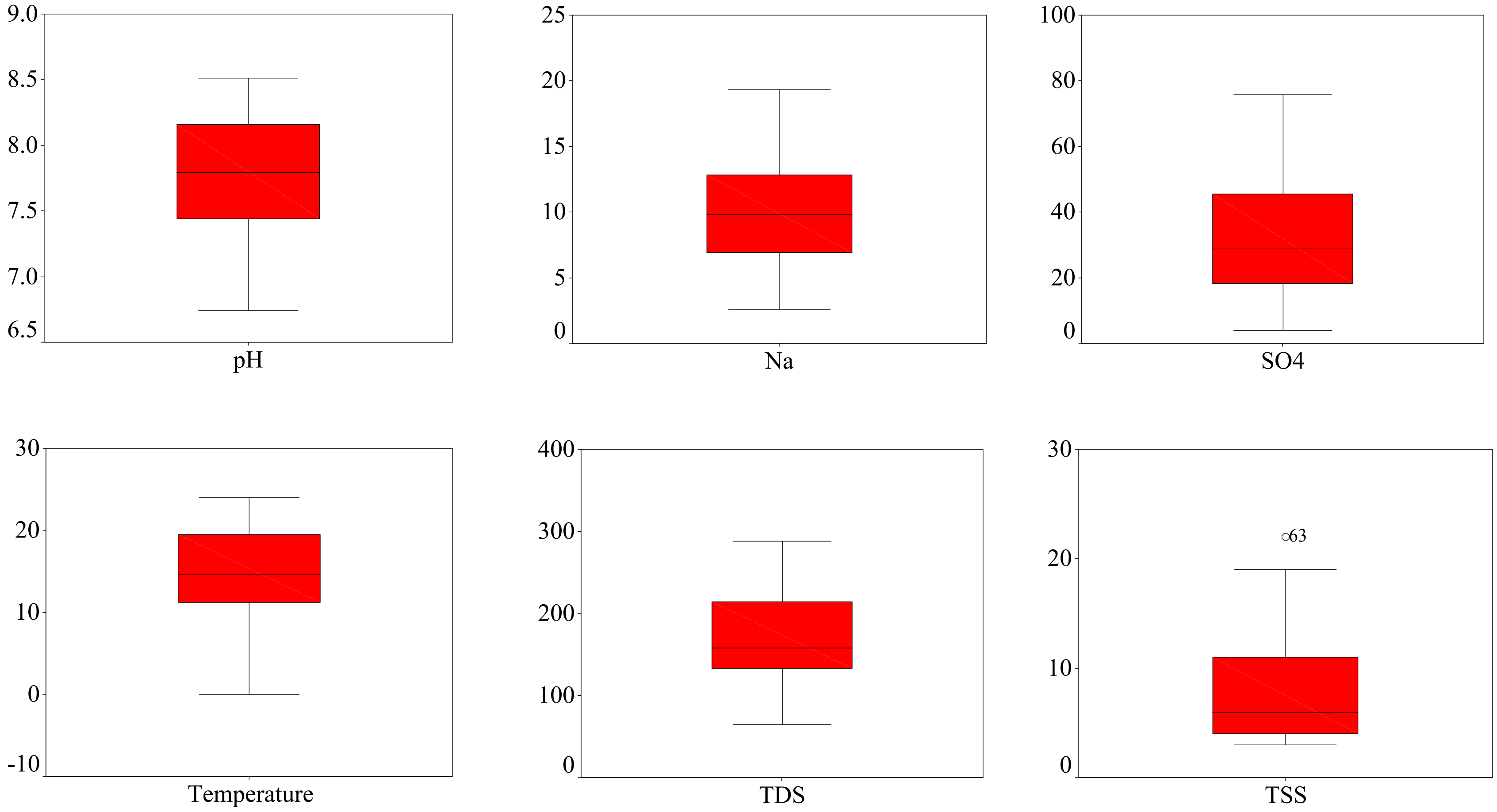


Figure 19 (cont.) Location MB05RAS078, WQ variables Box-plots

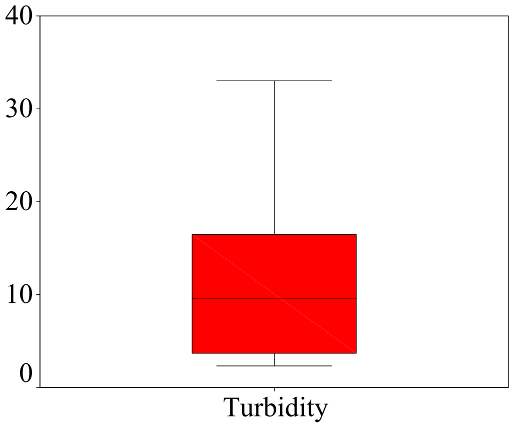


Figure 19 (cont.) Location MB05RAS078, WQ variables Box-plots

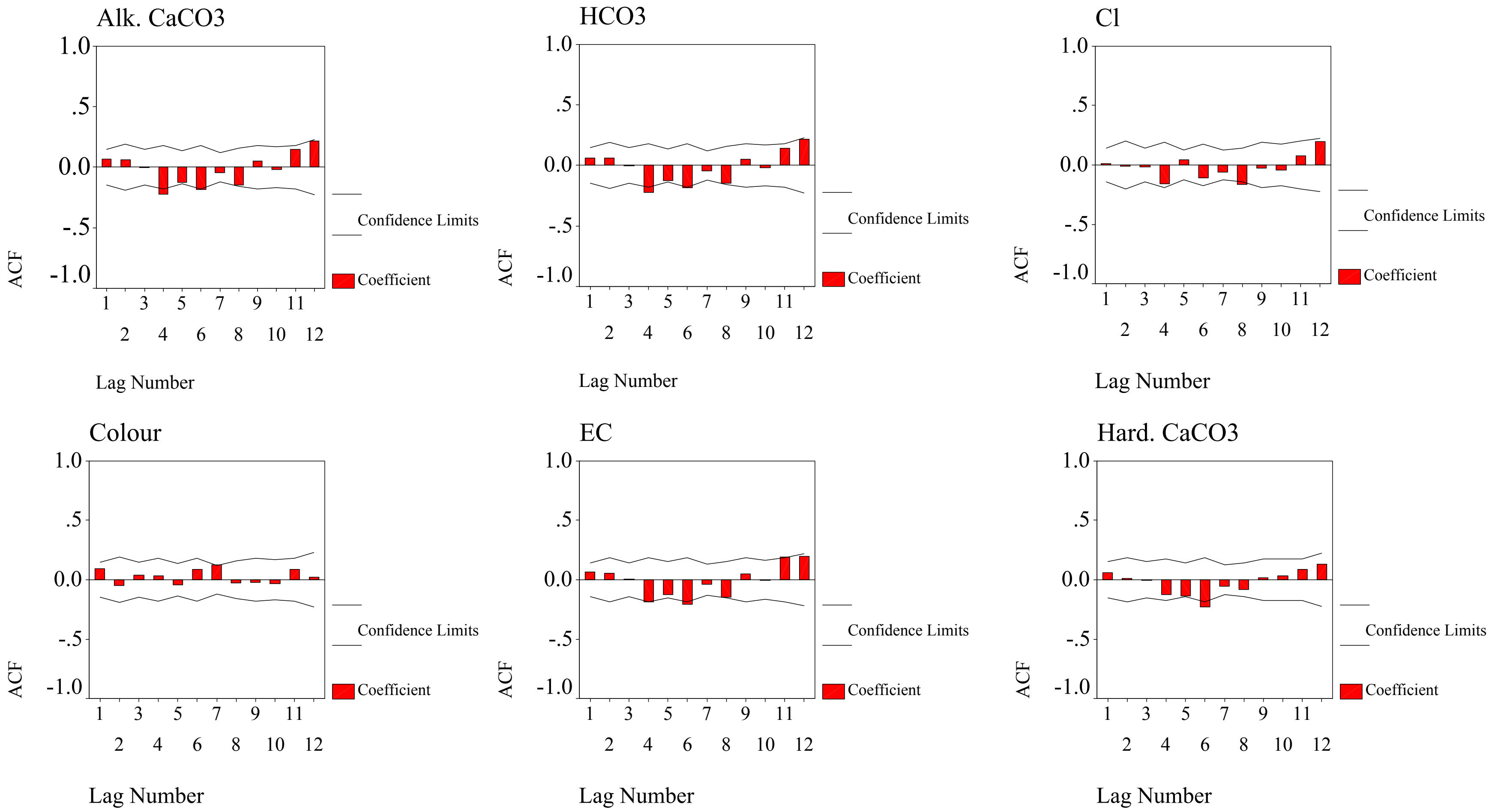


Figure 20. Location MB05RAS078, WQ variables Corellogram

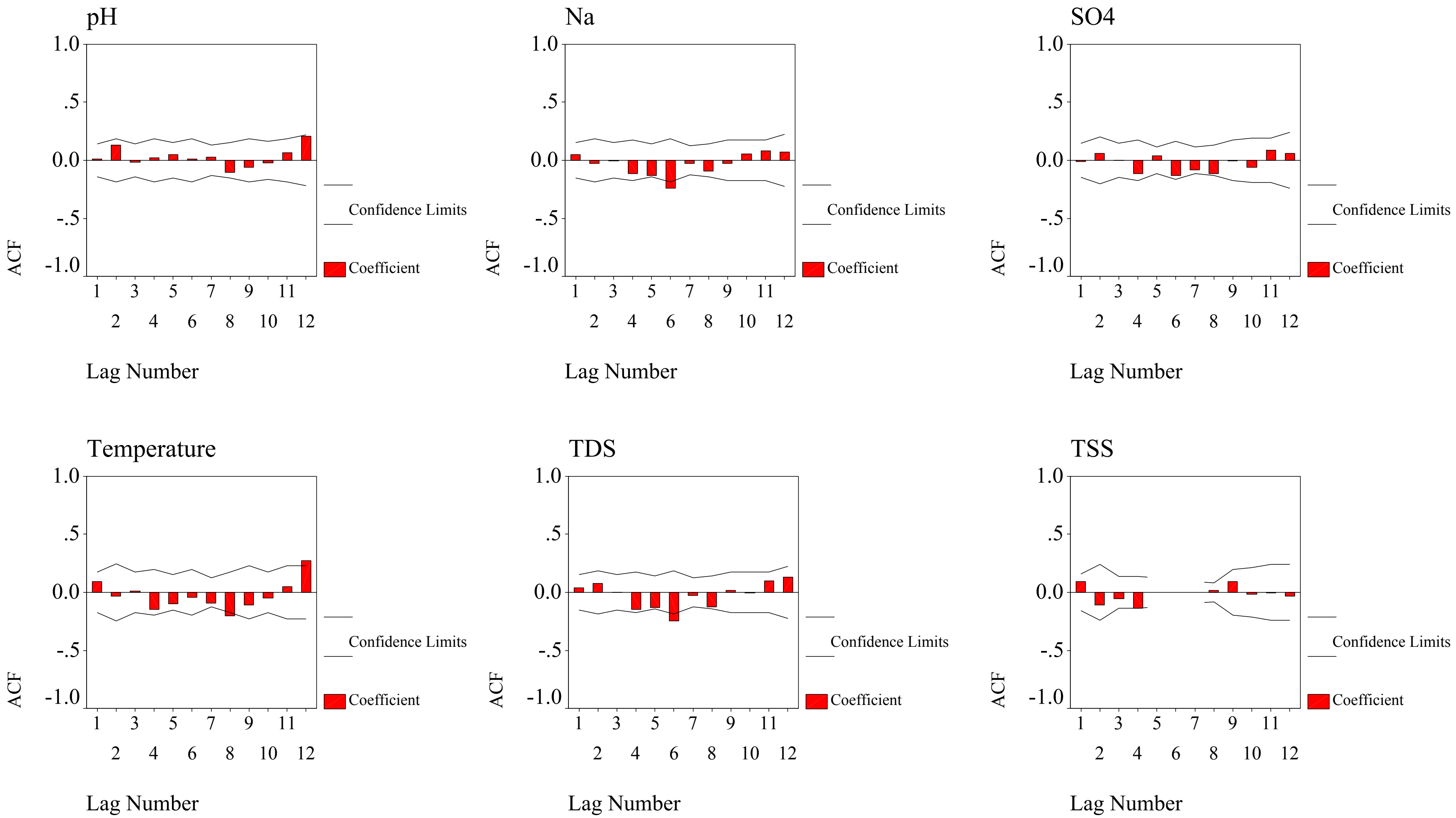


Figure 20 (cont.) Location MB05RAS078, WQ variables Corellogram

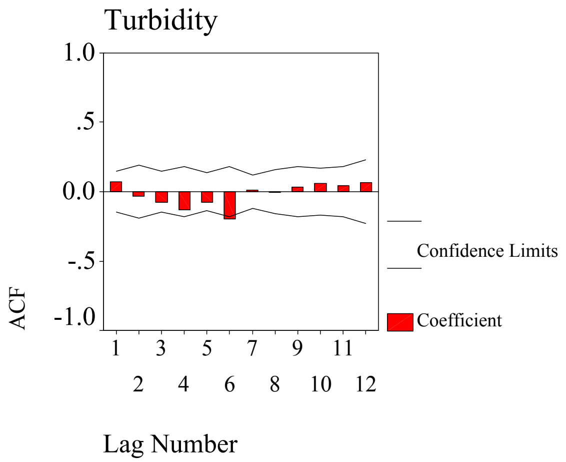


Figure 20 (cont.) Location MB05RAS078, WQ variables Corellogram

Location MB05RBAS0038

Table 17 Location MB05RBS003, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	40	40	34	39	41	40	41	35	34	24	39	37	40
Mean	82.59	100.75	7.16	34.26	237.15	97.58	7.85	9.21	27.57	12.00	159.05	8.86	14.31
Median	81.50	99.45	6.75	35.00	223.00	95.90	7.89	9.19	25.40	13.20	150.00	8.00	14.55
Std. Deviation	15.60	19.06	2.38	12.54	53.76	22.83	0.26	2.81	11.67	9.01	34.73	6.15	7.30
Variance	243.29	363.14	5.64	157.35	2890.43	521.24	0.07	7.88	136.15	81.14	1206.21	37.79	53.31
Skewness	0.27	0.28	0.44	0.14	0.39	0.60	-0.24	0.52	0.64	-0.14	0.60	1.53	0.61
Std. Error of Skewness	0.37	0.37	0.40	0.38	0.37	0.37	0.37	0.40	0.40	0.47	0.38	0.39	0.37
Kurtosis	-0.63	-0.60	-0.41	-1.00	-0.56	-0.37	0.75	-0.19	-0.24	-1.23	0.26	3.22	0.76
Std. Error of Kurtosis	0.73	0.73	0.79	0.74	0.72	0.73	0.72	0.78	0.79	0.92	0.74	0.76	0.73
Range	62.70	76.90	9.38	48.00	203.00	80.00	1.32	11.20	47.92	27.00	163.00	29.00	34.80
Minimum	53.30	65.10	2.62	10.00	143.00	64.00	7.15	4.50	9.08	0.00	87.00	1.00	2.20
Maximum	116.00	142.00	12.00	58.00	346.00	144.00	8.47	15.70	57.00	27.00	250.00	30.00	37.00

Table 18 Location MB05RBS003, WQ variables Kolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	Absolute	0.084	0.083	0.109	0.126	0.119	0.106	0.115	0.121	0.112	0.196	0.119	0.175	0.125
	Positive	0.084	0.083	0.109	0.103	0.119	0.106	0.080	0.121	0.112	0.196	0.119	0.175	0.125
	Negative	-0.080	-0.082	-0.066	-0.126	-0.063	-0.087	-0.115	-0.063	-0.093	-0.110	-0.077	-0.105	-0.076
Kolmogorov-Smirnov Z		0.530	0.524	0.633	0.787	0.760	0.671	0.734	0.718	0.656	0.958	0.741	1.064	0.790
Asymp. Sig. (2-tailed)		0.941	0.946	0.817	0.565	0.610	0.758	0.654	0.682	0.783	0.318	0.642	0.207	0.561

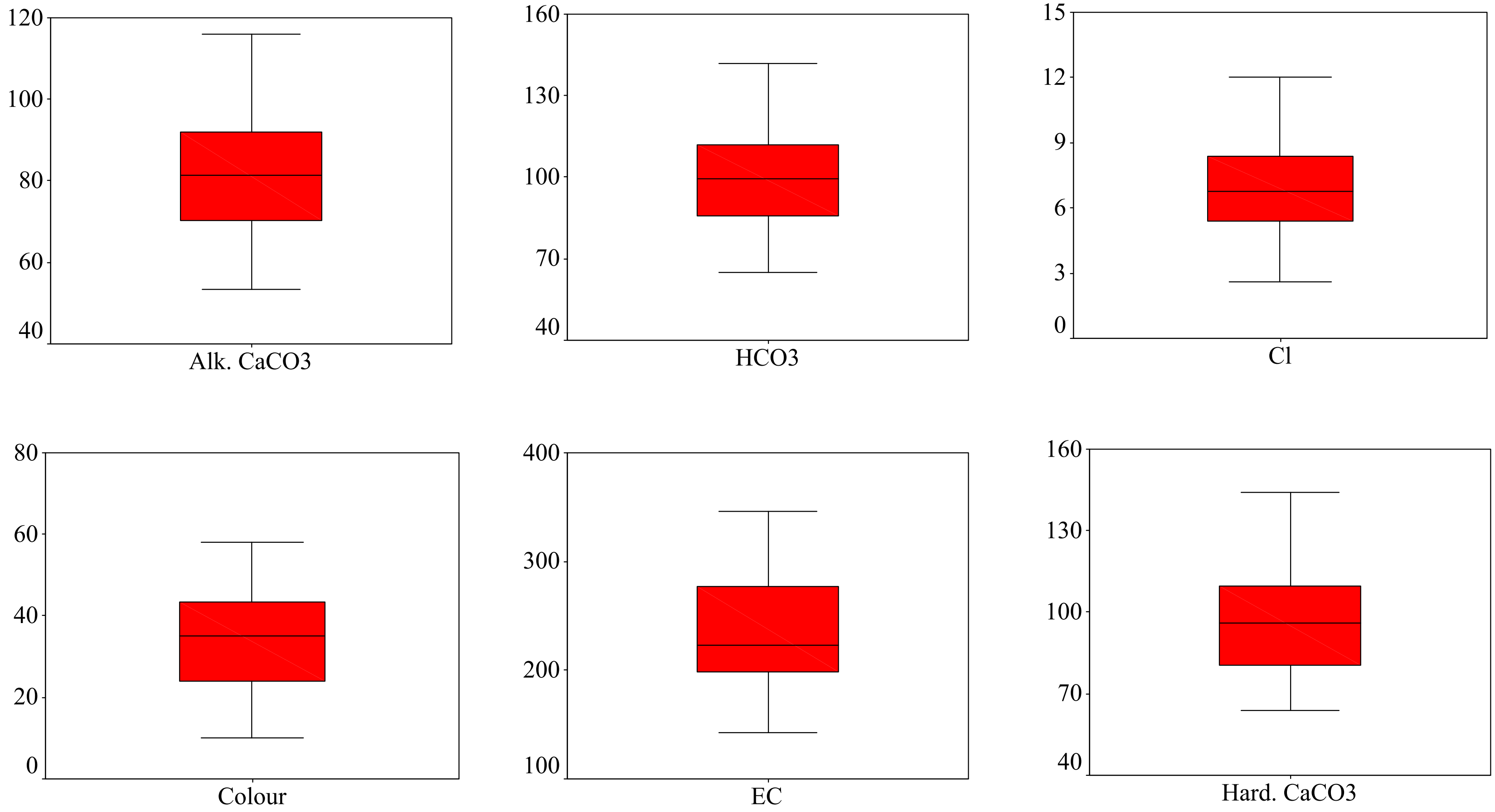


Figure 21. Location MB05RBS003, WQ variables Box-plots

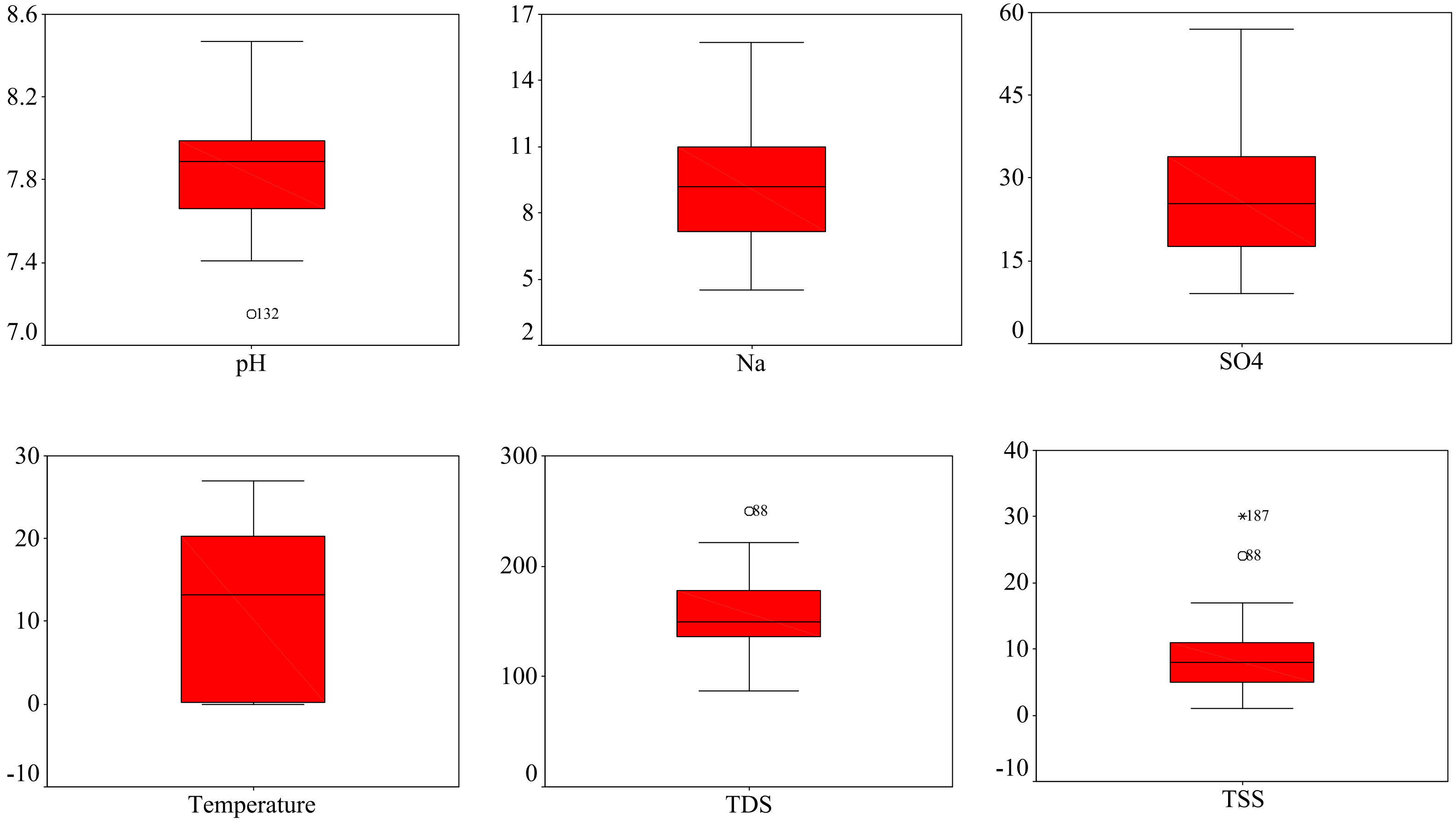


Figure 21 (cont.) Location MB05RBS003, WQ variables Box-plots

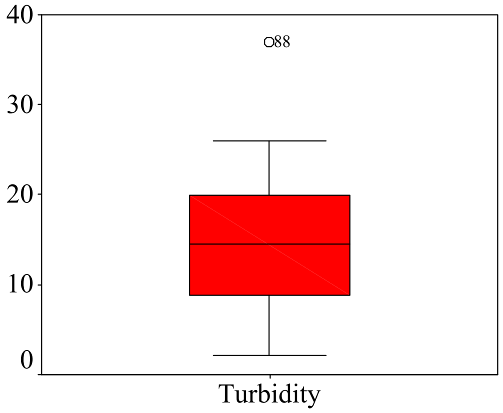


Figure 21 (cont.) Location MB05RBS003, WQ variables Box-plots

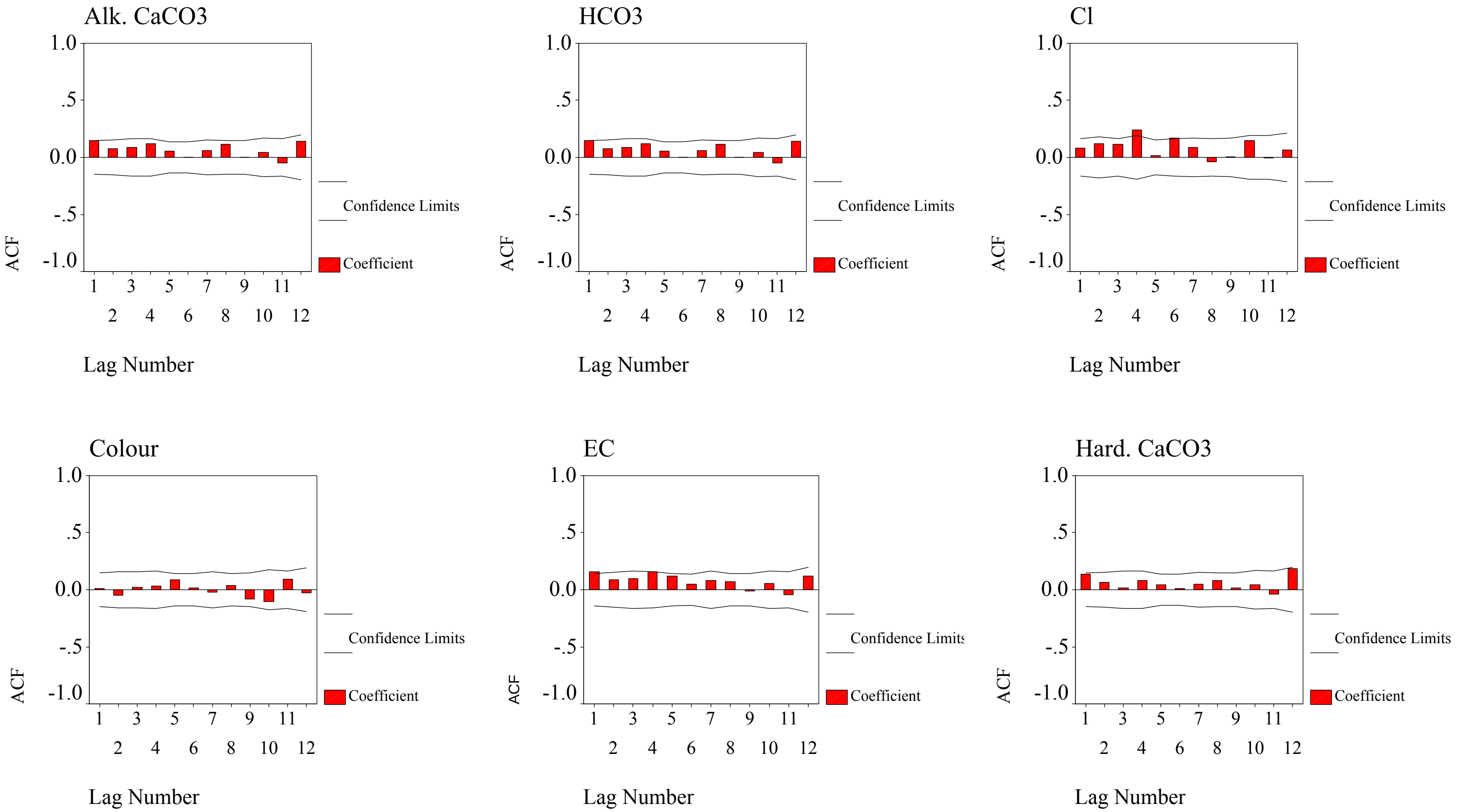


Figure 22. Location MB05RBS003, WQ variables Corellogram

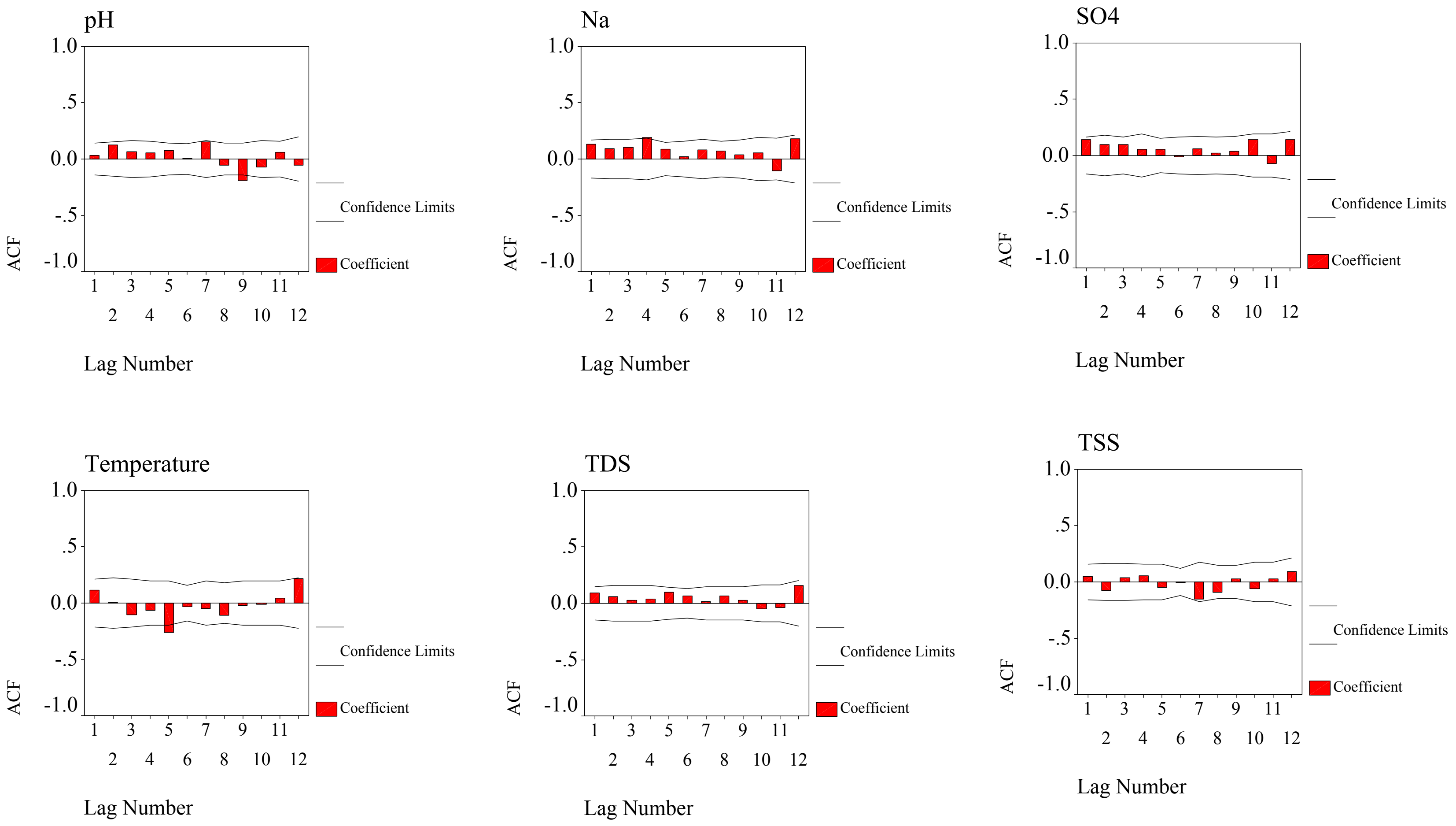


Figure 22 (cont.) Location MB05RBS003, WQ variables Corellogram

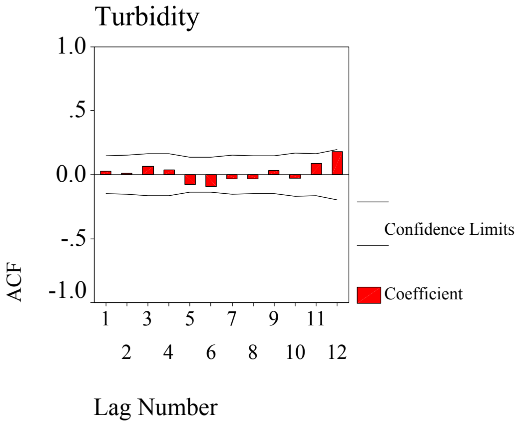


Figure 22 (cont.) Location MB05RBS003, WQ variables Corellogram

Location MB05SAS004

Table 19 Location MB05SAS004, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	30	30	25	30	31	29	31	30	26	19	30	29	30
Mean	49.52	60.36	2.39	40.47	117.74	52.86	7.55	3.18	7.11	12.00	93.03	18.76	16.99
Median	47.30	57.30	1.70	39.50	110.00	50.00	7.57	2.80	4.40	13.20	86.00	10.00	8.05
Std. Deviation	7.02	8.57	1.71	11.61	23.70	10.57	0.29	1.39	5.28	8.33	25.98	25.66	21.56
Variance	49.27	73.50	2.91	134.74	561.60	111.79	0.08	1.94	27.88	69.33	674.93	658.26	464.90
Skewness	2.06	2.07	2.60	0.24	2.38	2.24	-0.35	2.63	1.57	-0.37	1.30	2.54	2.60
Std. Error of Skewness	0.43	0.43	0.46	0.43	0.42	0.43	0.42	0.43	0.46	0.52	0.43	0.43	0.43
Kurtosis	4.71	4.71	7.05	-0.96	5.03	5.13	0.22	6.57	1.05	-1.05	2.81	6.29	6.78
Std. Error of Kurtosis	0.83	0.83	0.90	0.83	0.82	0.85	0.82	0.83	0.89	1.01	0.83	0.85	0.83
Range	32.6	39.7	7.25	40	96	46	1.24	5.9	17.23	24	129	107	91.9
Minimum	39.4	48.1	1.38	20	98	40	6.93	2.3	3.27	0	49	1	1.6
Maximum	72	87.8	8.63	60	194	86	8.17	8.2	20.5	24	178	108	93.5

Table 20 Location MB05SAS004, WQ variables Kolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	Absolute	0.229	0.228	0.356	0.128	0.252	0.222	0.100	0.328	0.292	0.188	0.177	0.312	0.247
	Positive	0.229	0.228	0.356	0.116	0.252	0.222	0.060	0.328	0.292	0.188	0.177	0.312	0.247
	Negative	-0.160	-0.162	-0.277	-0.128	-0.202	-0.194	-0.100	-0.263	-0.234	-0.179	-0.102	-0.244	-0.238
Kolmogorov-Smirnov Z		1.252	1.249	1.778	0.699	1.402	1.196	0.556	1.798	1.487	0.819	0.968	1.679	1.351
Asymp. Sig. (2-tailed)		0.087	0.088	0.004	0.713	0.039	0.114	0.917	0.003	0.024	0.514	0.305	0.007	0.052

Table 21 Location MB05SAS004, transformed WQ variables descriptive statistics

WQ variable	Ln Cl	Ln EC	Ln Na	Ln SO_4	Ln TSS
Measuring unit	**mg/l**	**mg/l**	**mg/l**	**CU**	**US/cm**
n	25	31	29	26	29
Mean	0.725	4.753	1.062	1.766	2.262
Median	0.531	4.700	1.030	1.481	2.303
Std. Deviation	0.485	0.170	0.262	0.588	1.198
Variance	0.236	0.029	0.069	0.345	1.436
Skewness	1.806	2.088	2.116	1.141	0.011
Std. Error of Skewness	0.464	0.421	0.434	0.456	0.434
Kurtosis	2.467	3.929	4.936	-0.107	-0.164
Std. Error of Kurtosis	0.902	0.821	0.845	0.887	0.845
Range	1.833	0.683	1.039	1.836	4.682
Minimum	0.322	4.585	0.833	1.185	0.000
Maximum	2.155	5.268	1.872	3.020	4.682

Table 22 Location MB05SAS004, transformed WQ variables Kolmogrov-Smirnov test

Test parameters		Ln Cl	Ln EC	Ln Na	Ln SO_4	Ln TSS
Most Extreme Differences	**Absolute**	0.270	0.216	0.212	0.211	0.107
	Positive	0.270	0.216	0.212	0.211	0.107
	Negative	-0.203	-0.162	-0.191	-0.161	-0.086
Kolmogorov-Smirnov Z		1.349	1.205	1.142	1.073	0.575
Asymp. Sig. (2-tailed)		0.053	0.109	0.147	0.199	0.896

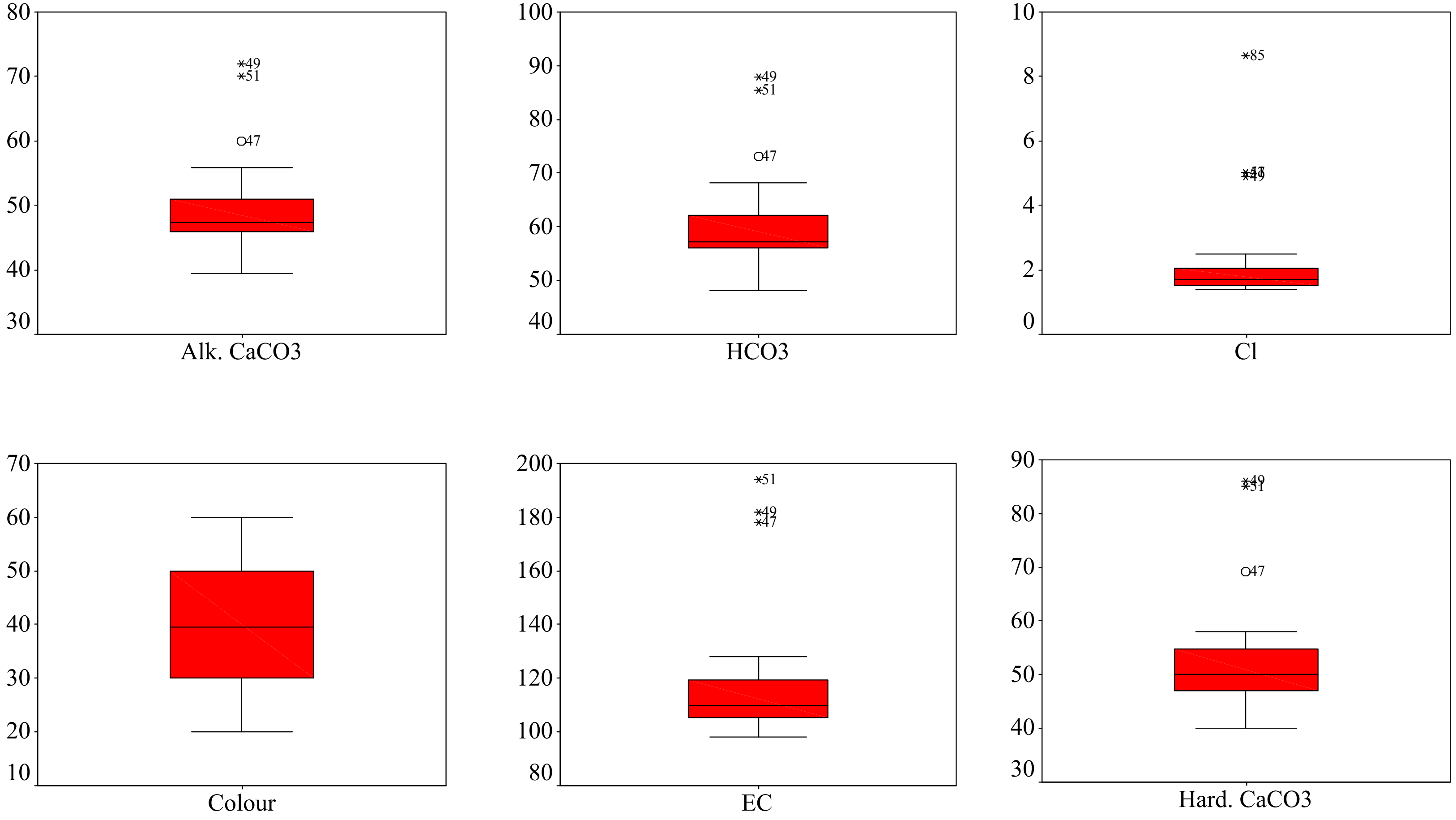


Figure 23. Location MB05SAS004, WQ variables Box-plots

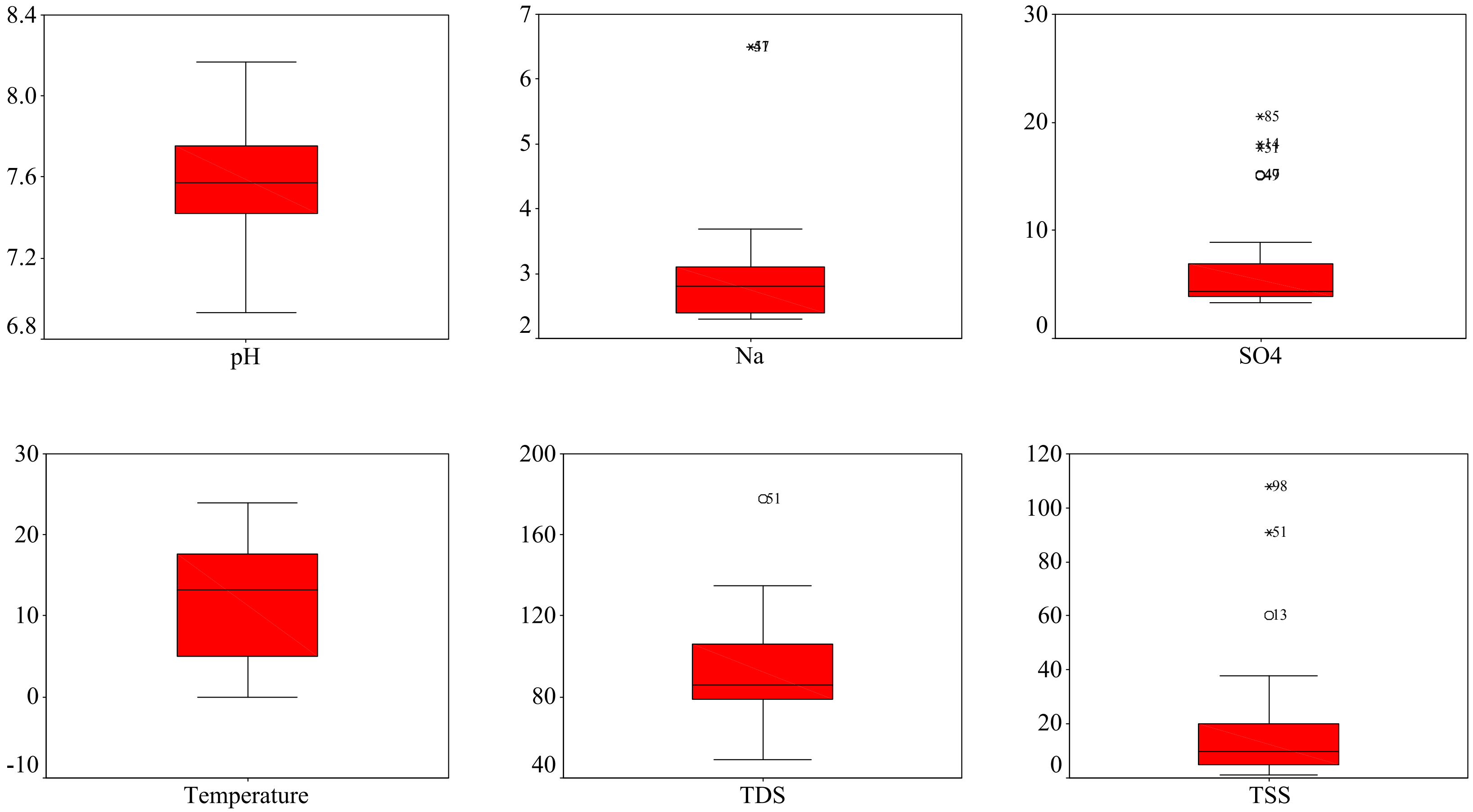


Figure 23 (cont.) Location MB05SAS004, WQ variables Box-plots

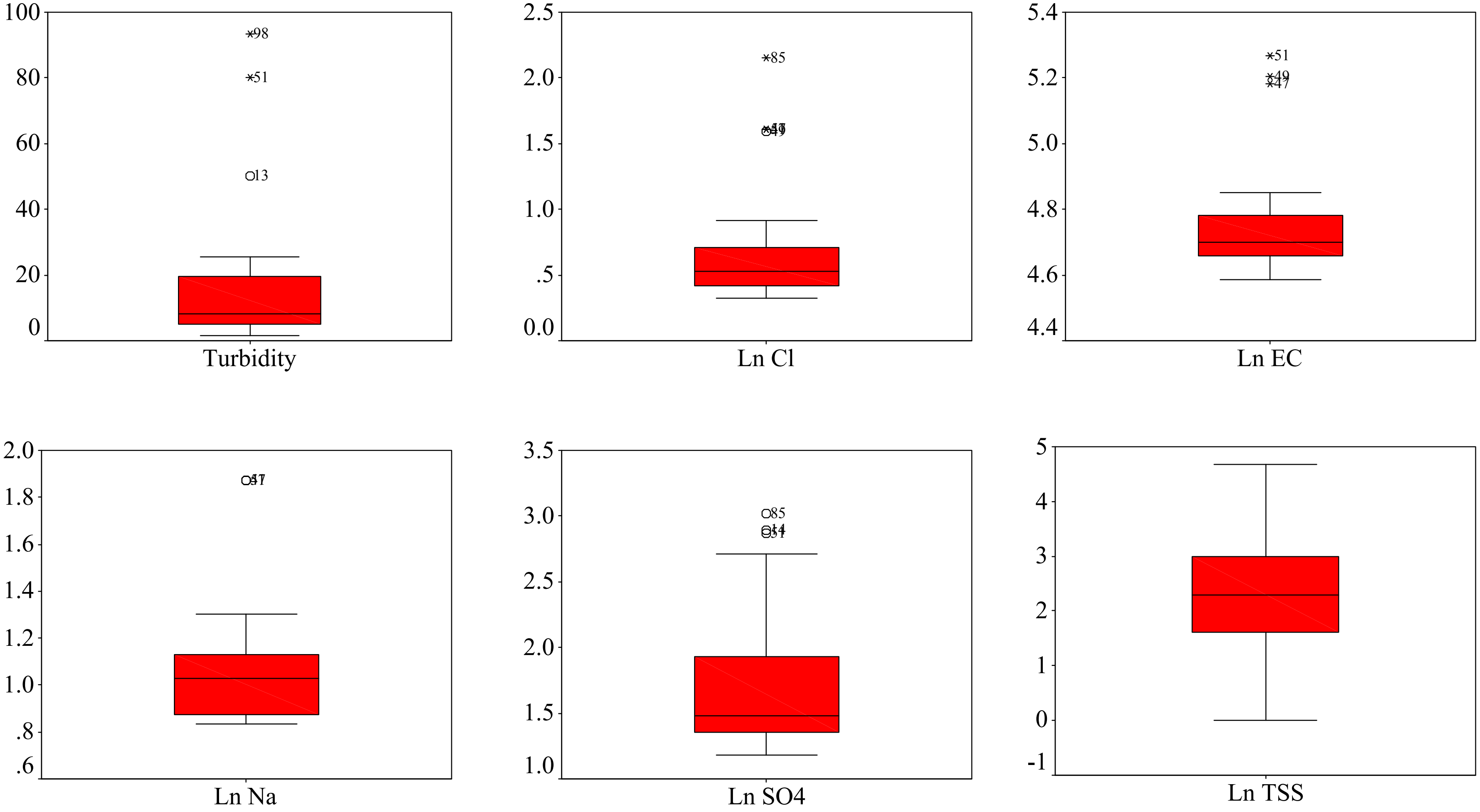


Figure 23 (cont.) Location MB05SAS004, WQ variables Box-plots

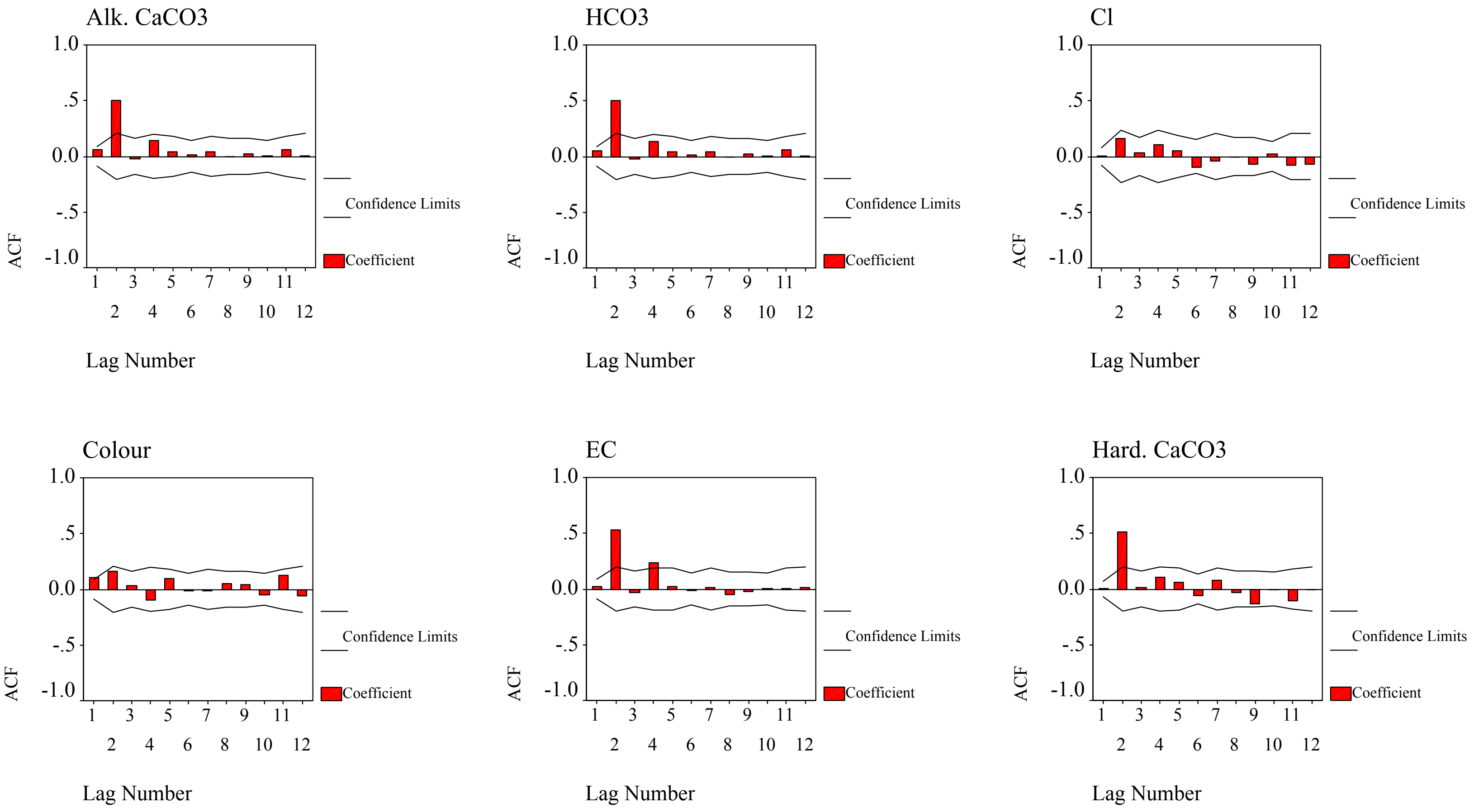


Figure 24. Location MB05SAS004, WQ variables Corellogram

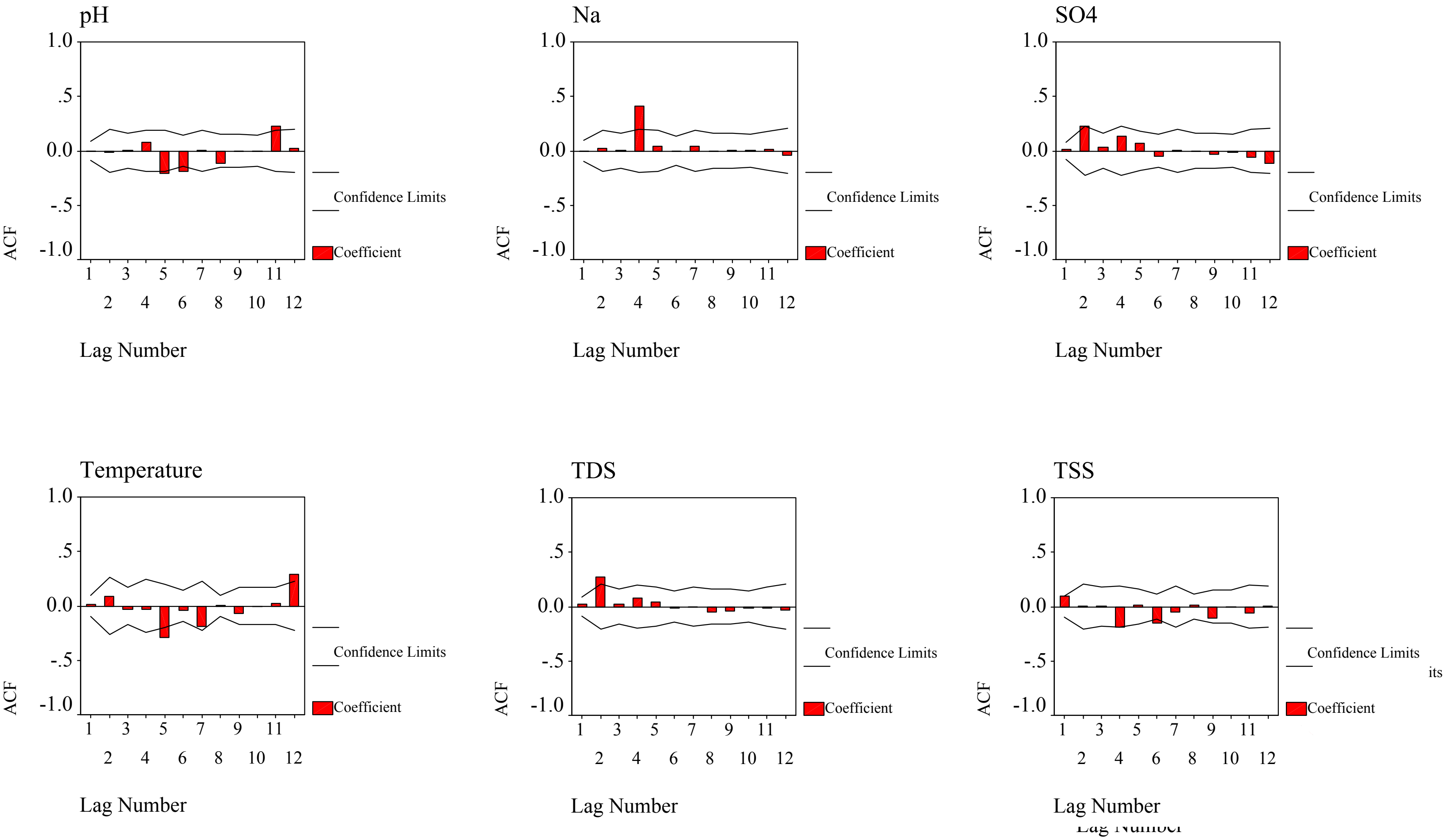


Figure 24 (cont.) Location MB05SAS004, WQ variables Corellogram

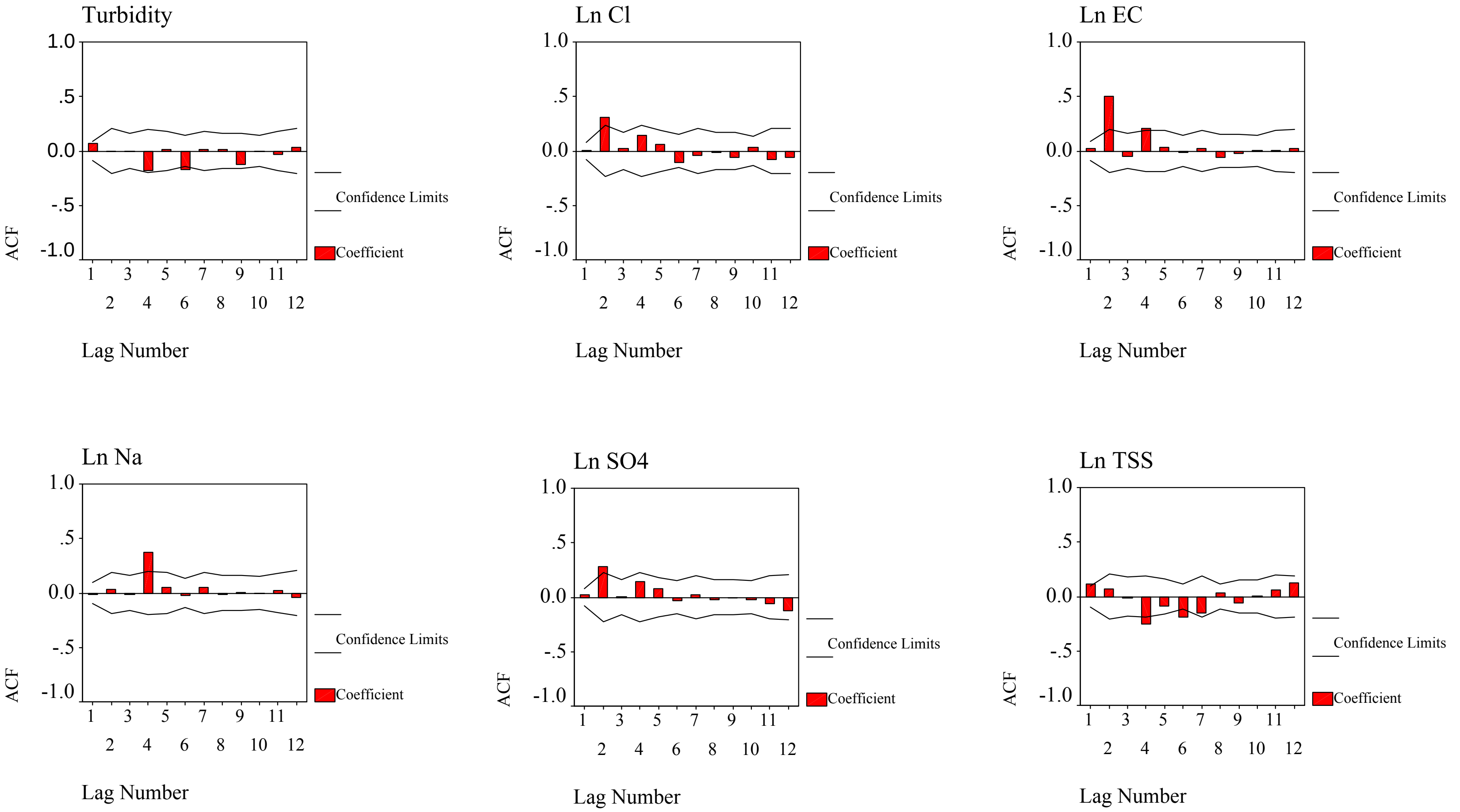


Figure 24 (cont.) Location MB05SAS004, WQ variables Corellogram

Location MB05SBS126

Table 23. Location MB05SBS126, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	**mg/l**	**mg/l**	**mg/l**	**CU**	**US/cm**	**mg/l**	**pH units**	**mg/l**	**mg/l**	**deg C**	**mg/l**	**mg/l**	**NTU**
n	33	33	33	33	35	31	35	31	33	21	33	29	33
Mean	111.13	134.14	11.07	40.70	347.89	141.14	7.95	14.37	50.68	11.45	235.97	19.97	23.35
Median	112.00	137.00	11.50	35.00	334.00	135.00	8.02	13.90	45.00	12.60	210.00	11.00	17.10
Std. Deviation	28.37	33.51	4.35	24.40	113.44	45.62	0.36	6.49	26.61	8.14	78.81	18.56	19.88
Variance	805.04	1123.09	18.91	595.28	12869.75	2080.76	0.13	42.07	708.35	66.21	6211.03	344.53	395.32
Skewness	0.22	0.29	-0.21	3.85	0.11	0.22	-1.03	0.11	0.22	-0.12	0.97	1.77	1.69
Std. Error of Skewness	0.41	0.41	0.41	0.41	0.40	0.42	0.40	0.42	0.41	0.50	0.41	0.43	0.41
Kurtosis	-0.39	-0.10	-1.01	18.44	-1.02	-0.79	1.25	-0.60	-1.20	-1.06	1.06	2.59	3.07
Std. Error of Kurtosis	0.80	0.80	0.80	0.80	0.78	0.82	0.78	0.82	0.80	0.97	0.80	0.85	0.80
Range	108.90	133.40	14.52	142.00	385.00	164.00	1.58	26.40	87.80	24.00	344.00	72.00	87.10
Minimum	61.10	74.60	3.48	18.00	159.00	68.00	6.90	1.60	9.30	0.00	122.00	2.00	3.90
Maximum	170.00	208.00	18.00	160.00	544.00	232.00	8.48	28.00	97.10	24.00	466.00	74.00	91.00

Table 24. Location MB05SBS126, WQ variablesKolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	**Absolute**	0.074	0.072	0.080	0.230	0.114	0.074	0.168	0.074	0.137	0.157	0.156	0.226	0.209
	Positive	0.074	0.072	0.080	0.230	0.078	0.072	0.075	0.074	0.132	0.157	0.156	0.226	0.209
	Negative	-0.057	-0.055	-0.076	-0.176	-0.114	-0.074	-0.168	-0.065	-0.137	-0.118	-0.074	-0.176	-0.164
Kolmogorov-Smirnov Z		0.426	0.414	0.462	1.323	0.676	0.411	0.991	0.409	0.788	0.720	0.898	1.218	1.202
Asymp. Sig. (2-tailed)		0.993	0.995	0.983	0.060	0.750	0.996	0.279	0.996	0.565	0.677	0.395	0.103	0.111

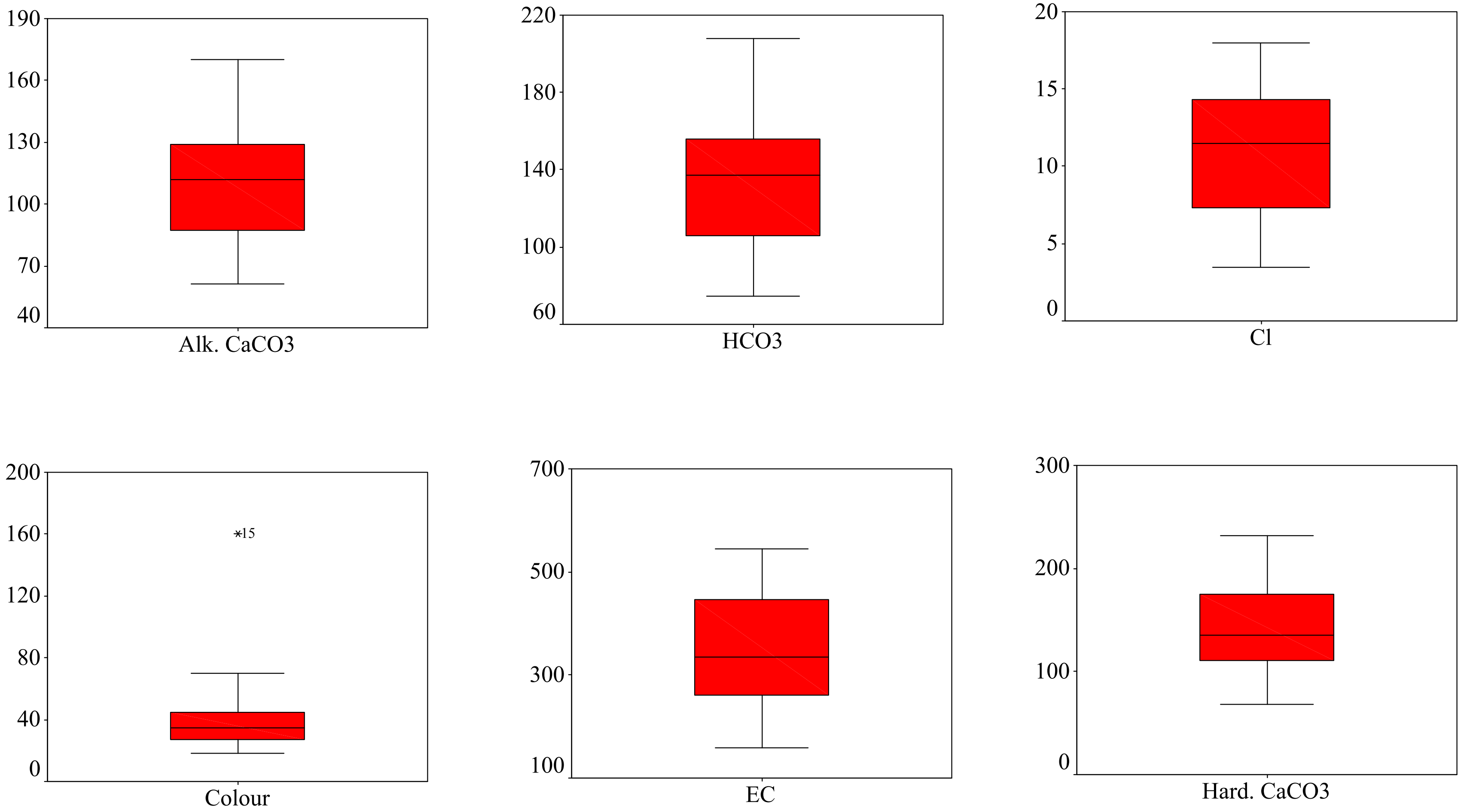


Figure 25. Location MB05SBS126, WQ variables Box-plots

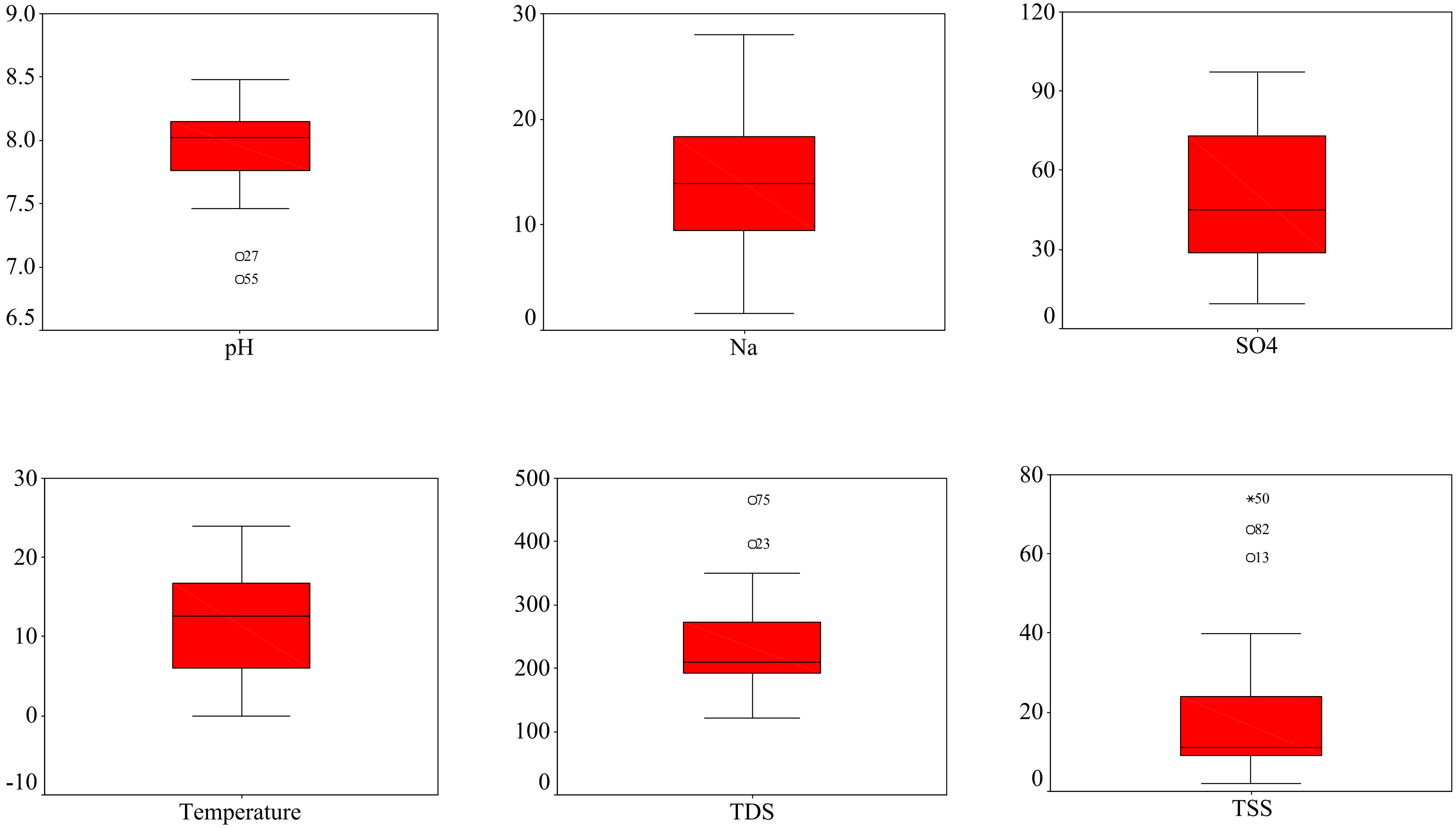


Figure 25 (cont.) Location MB05SBS126, WQ variables Box-plots

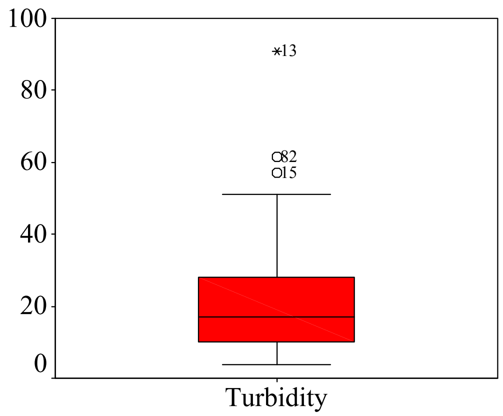


Figure 25 (cont.) Location MB05SBS126, WQ variables Box-plots

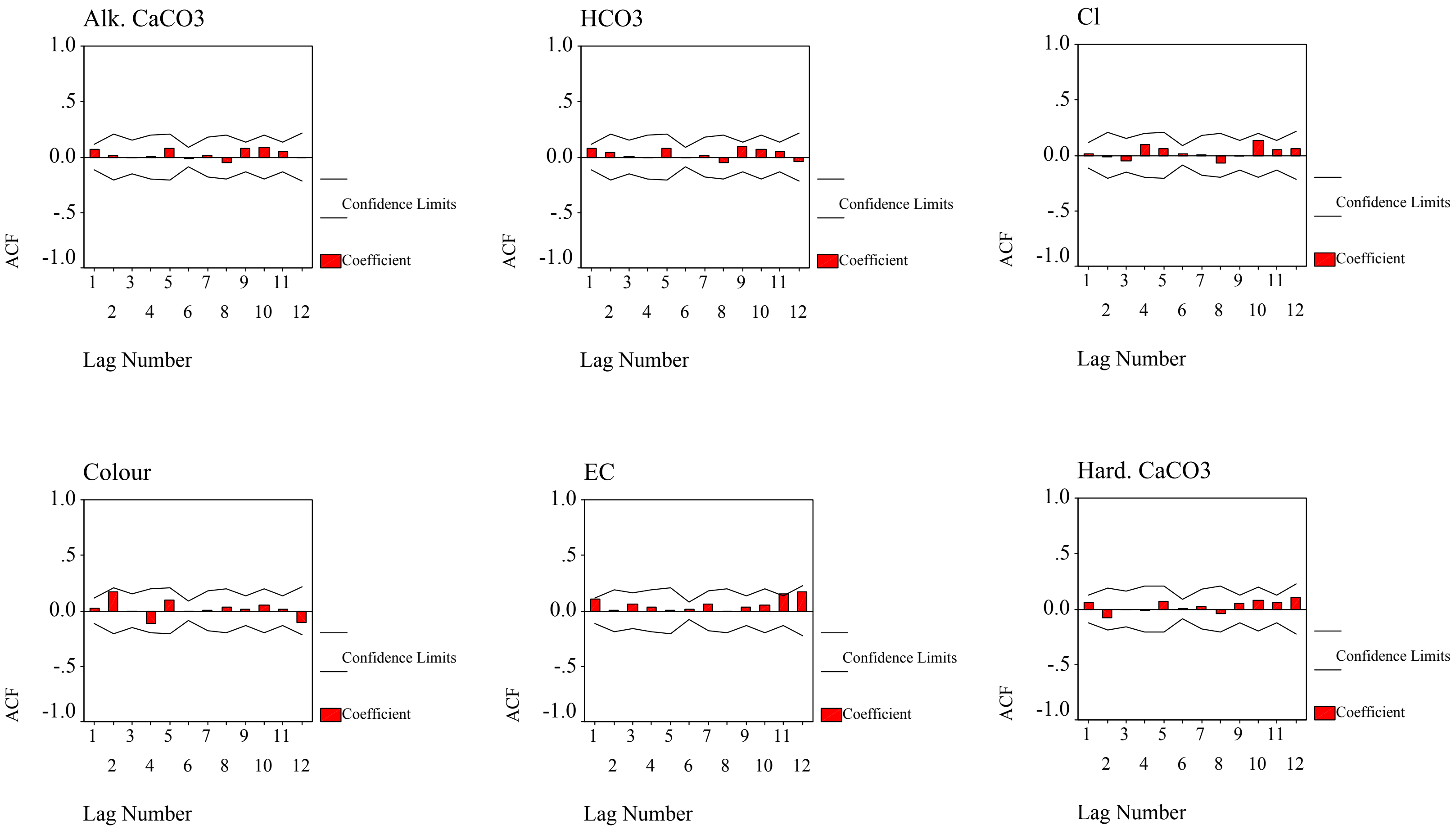


Figure 26. Location MB05SBS126, WQ variables Corellogram

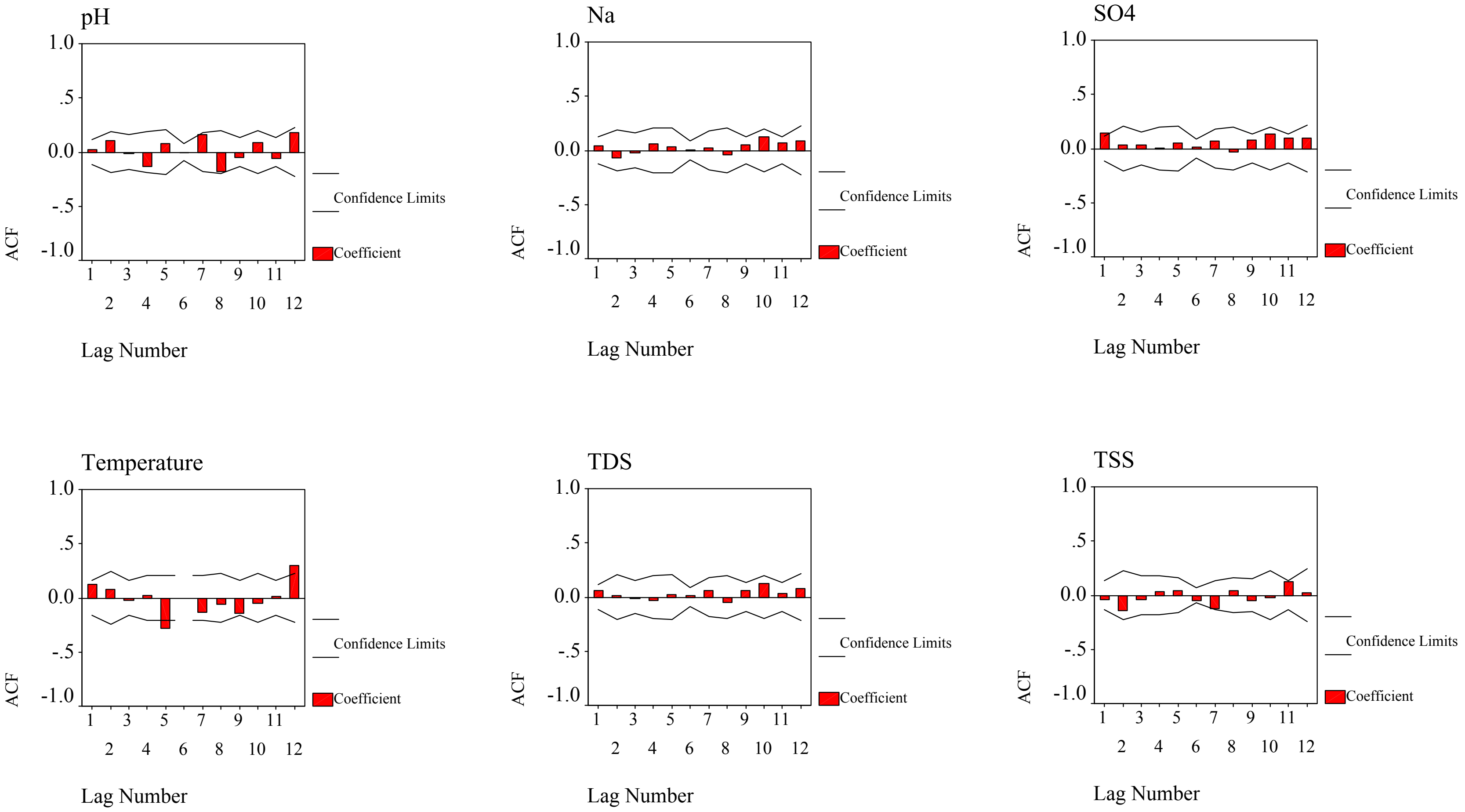


Figure 26 (cont.) Location MB05SBS126, WQ variables Corellogram

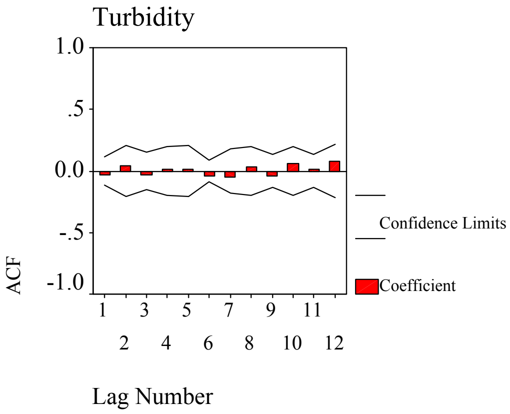


Figure 26 (cont.) Location MB05SBS126, WQ variables Corellogram

Location MB05SCS005

Table 25. Location MB05SCS005, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	41	41	35	39	42	41	42	41	35	26	40	31	41
Mean	95.12	115.69	9.00	32.10	279.10	113.65	7.92	11.27	37.38	11.54	186.18	6.65	10.83
Median	97.00	118.00	9.60	30.00	282.50	114.00	7.97	11.30	39.10	13.25	190.00	6.00	8.60
Std. Deviation	21.32	25.72	3.86	11.94	84.81	31.59	0.35	4.81	17.62	8.27	53.01	3.87	7.24
Variance	454.59	661.42	14.87	142.52	7193.41	997.75	0.12	23.11	310.35	68.47	2809.53	14.97	52.35
Skewness	-0.26	-0.26	-0.29	0.81	-0.19	-0.27	-1.06	0.18	-0.20	-0.25	-0.26	1.01	1.22
Std. Error of Skewness	0.37	0.37	0.40	0.38	0.37	0.37	0.37	0.37	0.40	0.46	0.37	0.42	0.37
Kurtosis	-0.42	-0.35	-0.54	0.37	-0.67	-0.50	2.23	0.00	-0.47	-1.43	-0.45	2.05	1.20
Std. Error of Kurtosis	0.72	0.72	0.78	0.74	0.72	0.72	0.72	0.72	0.78	0.89	0.73	0.82	0.72
Range	83.90	102.70	14.40	51.00	331.00	121.80	1.78	20.51	65.08	23.00	215.00	18.00	29.90
Minimum	51.10	62.30	1.60	12.00	113.00	48.20	6.76	2.59	4.72	0.00	74.00	1.00	2.10
Maximum	135.00	165.00	16.00	63.00	444.00	170.00	8.54	23.10	69.80	23.00	289.00	19.00	32.00

Table 26. Location MB05SCS005, WQ variablesKolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	H.$CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	Absolute	0.100	0.101	0.081	0.148	0.093	0.076	0.126	0.094	0.099	0.148	0.079	0.141	0.210
	Positive	0.061	0.061	0.071	0.148	0.066	0.073	0.086	0.094	0.067	0.148	0.075	0.141	0.210
	Negative	-0.100	-0.101	-0.081	-0.067	-0.093	-0.076	-0.126	-0.087	-0.099	-0.129	-0.079	-0.079	-0.114
Kolmogorov-Smirnov Z		0.642	0.650	0.481	0.922	0.600	0.489	0.816	0.603	0.587	0.756	0.497	0.784	1.343
Asymp. Sig. (2-tailed)		0.804	0.793	0.975	0.363	0.865	0.971	0.518	0.860	0.880	0.616	0.966	0.570	0.054

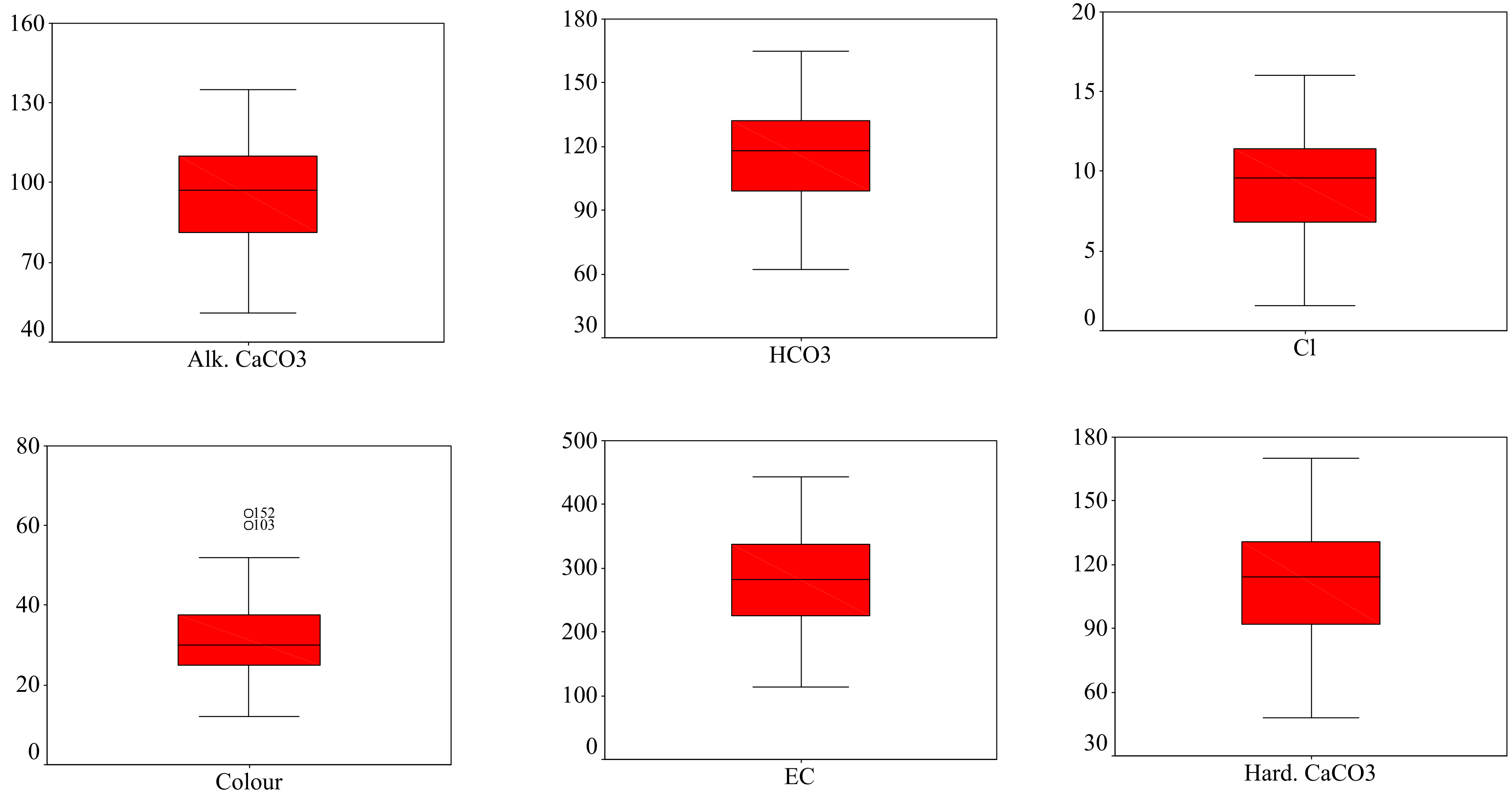


Figure 27. Location MB05SCS005, WQ variables Box-plots

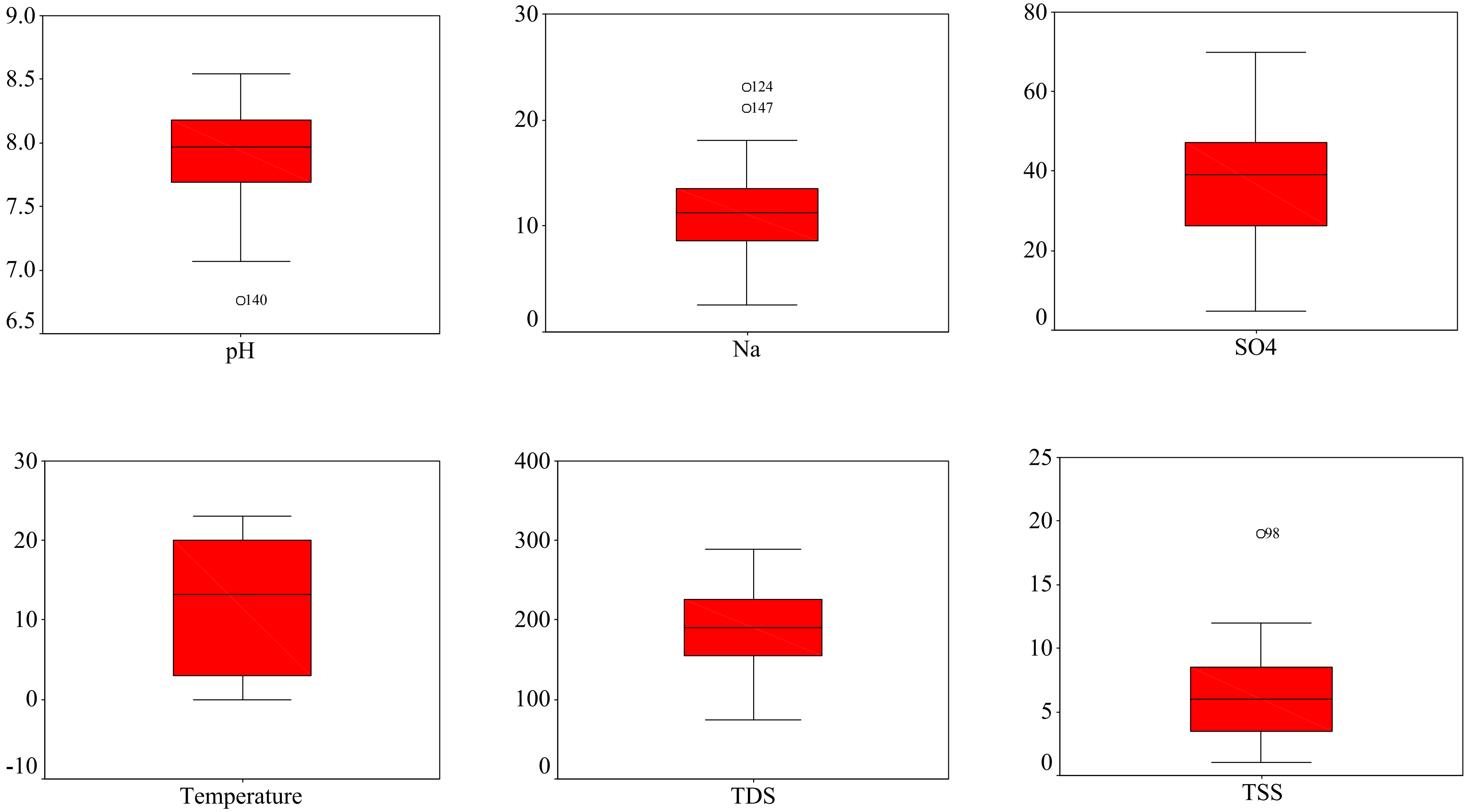


Figure 27 (cont.) Location MB05SCS005, WQ variables Box-plots

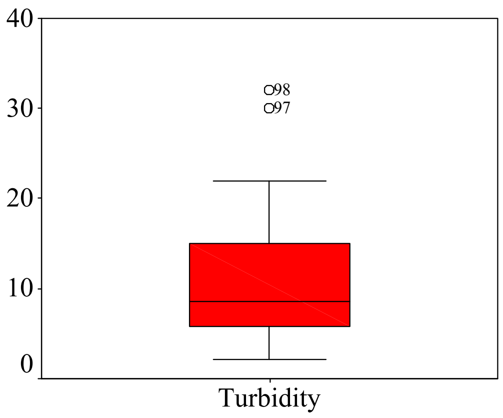


Figure 27 (cont.) Location MB05SCS005, WQ variables Box-plots

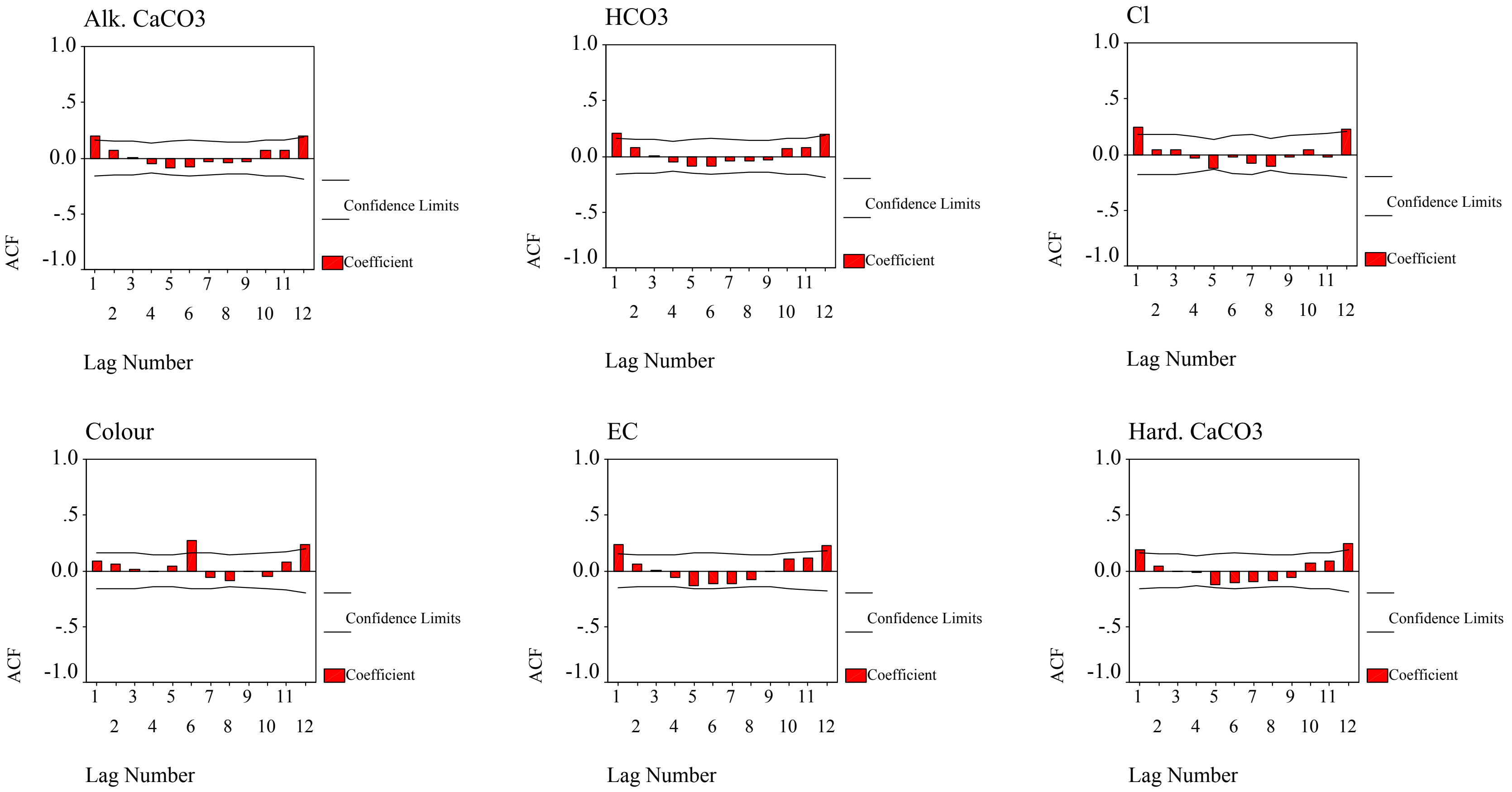


Figure 28. Location MB05SCS005, WQ variables Corellogram

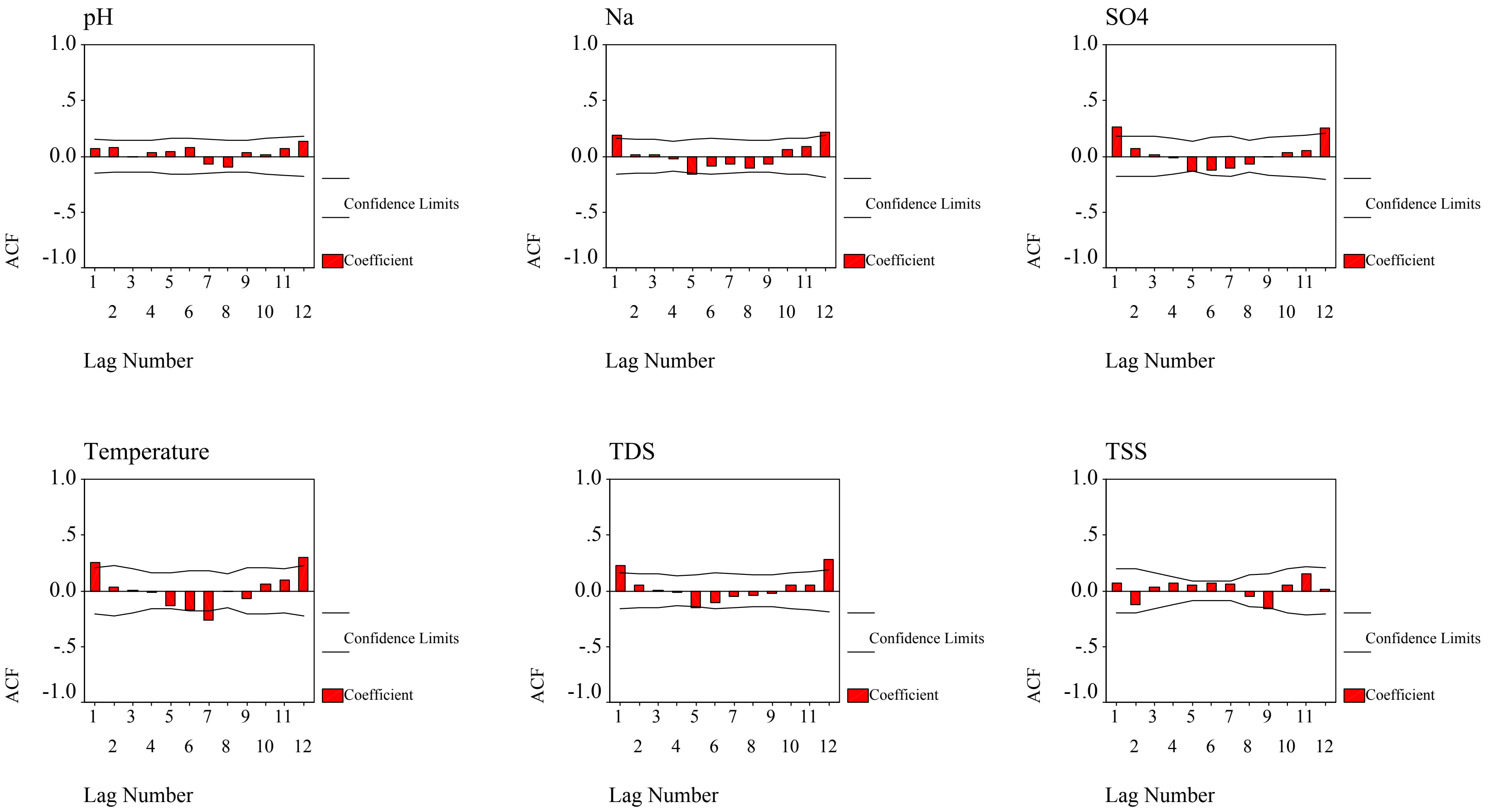


Figure 28 (cont.) Location MB05SCS005, WQ variables Corellogram

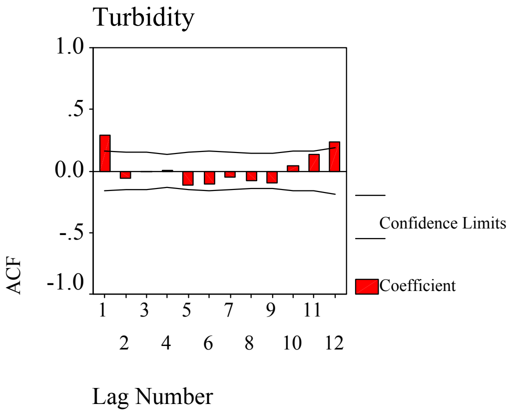


Figure 28 (cont.) Location MB05SCS005, WQ variables Corellogram

Location MB05SHS004

Table 27. Location MB05SHS004, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	36	36	30	29	37	36	37	36	30	19	36	26	36
Mean	119.67	143.94	19.34	8.79	357.19	140.28	8.06	18.35	34.04	12.54	207.67	5.58	4.16
Median	117.50	140.00	18.85	8.00	356.00	138.00	8.13	18.40	32.80	13.40	209.00	5.00	2.65
Std. Deviation	14.68	16.86	3.59	3.83	35.49	20.95	0.27	2.42	4.29	6.75	21.37	4.10	3.80
Variance	215.60	284.34	12.91	14.67	1259.38	438.83	0.07	5.86	18.44	45.52	456.63	16.81	14.43
Skewness	1.12	1.32	-0.18	0.51	0.79	0.69	-0.86	0.66	1.39	-0.79	-0.24	0.63	1.81
Std. Error of Skewness	0.39	0.39	0.43	0.43	0.39	0.39	0.39	0.39	0.43	0.52	0.39	0.46	0.39
Kurtosis	1.06	2.04	-0.51	-0.97	0.08	-0.45	1.08	1.44	2.72	-0.18	0.43	-0.68	3.04
Std. Error of Kurtosis	0.77	0.77	0.83	0.85	0.76	0.77	0.76	0.77	0.83	1.01	0.77	0.89	0.77
Range	63.00	77.00	14.30	12.00	135.00	75.00	1.31	12.00	19.80	21.90	99.00	13.00	16.21
Minimum	98.00	120.00	11.70	5.00	308.00	112.00	7.32	14.00	28.20	0.00	151.00	1.00	0.79
Maximum	161.00	197.00	26.00	17.00	443.00	187.00	8.63	26.00	48.00	21.90	250.00	14.00	17.00

Table 28. Location MB05SHS004, WQ variablesKolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	Absolute	0.187	0.160	0.135	0.253	0.112	0.122	0.160	0.083	0.131	0.158	0.123	0.155	0.239
	Positive	0.187	0.160	0.135	0.253	0.112	0.122	0.100	0.083	0.131	0.125	0.123	0.155	0.239
	Negative	-0.090	-0.091	-0.126	-0.161	-0.085	-0.089	-0.160	-0.049	-0.107	-0.158	-0.080	-0.132	-0.188
Kolmogorov-Smirnov Z		1.121	0.960	0.737	1.361	0.684	0.729	0.972	0.497	0.719	0.690	0.739	0.788	1.432
Asymp. Sig. (2-tailed)		0.162	0.316	0.648	0.049	0.737	0.663	0.301	0.966	0.679	0.729	0.645	0.563	0.033

Table 29. Location MB05SHS004, transformed WQ variables descriptive statistics

WQ variable	Colour square	Ln Turb.
Measuring unit	**CU**	**NTU**
n	29	36
Mean	91.48	1.10
Median	64.00	0.97
Std. Deviation	75.32	0.79
Variance	5673.76	0.63
Skewness	1.01	0.44
Std. Error of Skewness	0.43	0.39
Kurtosis	0.23	-0.52
Std. Error of Kurtosis	0.85	0.77
Range	264.00	3.07
Minimum	25.00	-0.24
Maximum	289.00	2.83

Table 30. Location MB05SHS004, transformed WQ variables Kolmogrov-Smirnov test

Test parameters		Colour square	Ln Turb.
Most Extreme Differences	**Absolute**	0.225	0.102
	Positive	0.225	0.102
	Negative	-0.189	-0.078
Kolmogorov-Smirnov Z		1.212	0.611
Asymp. Sig. (2-tailed)		0.106	0.850

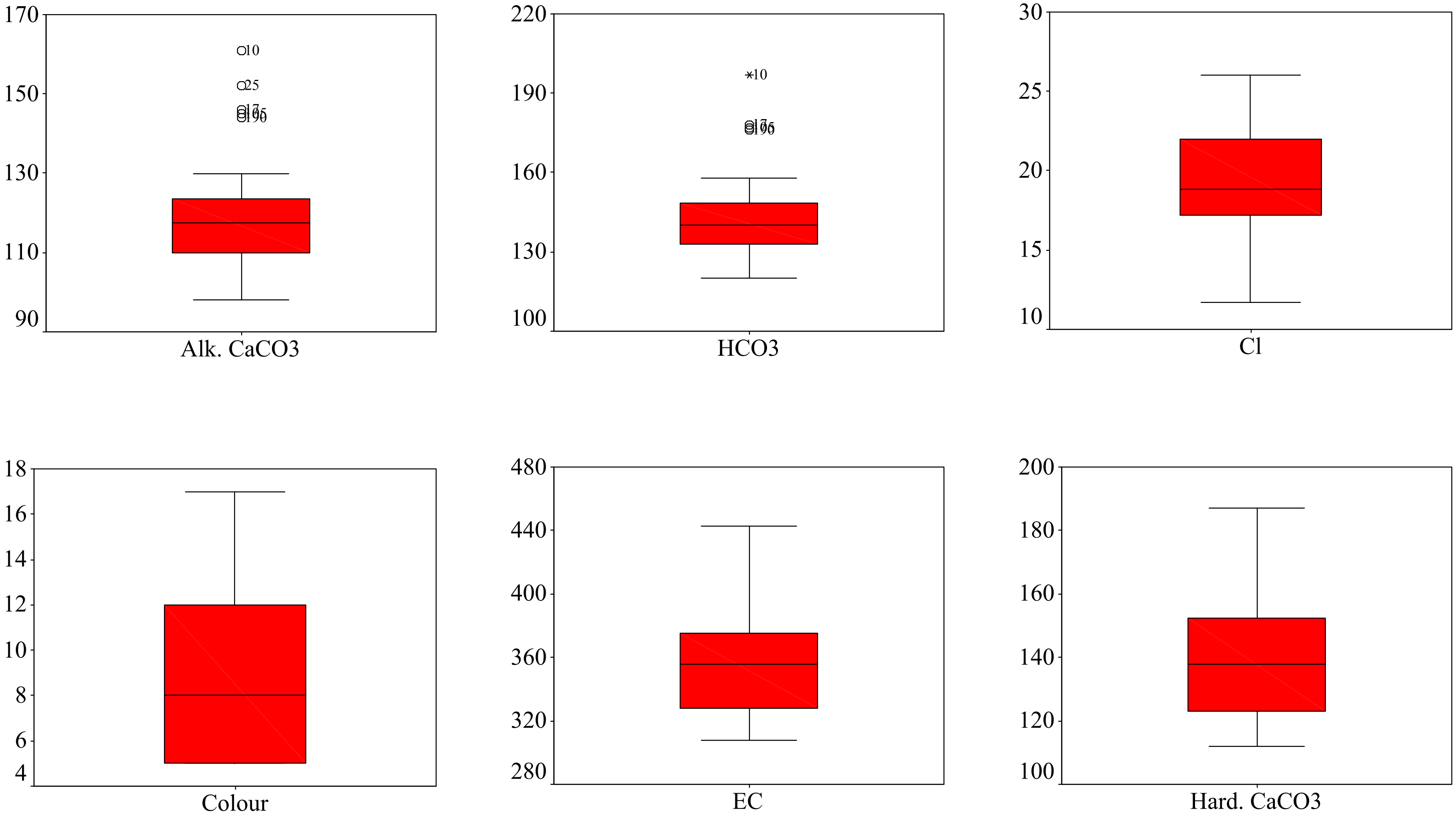


Figure 29. Location MB05SHS004, WQ variables Box-plots

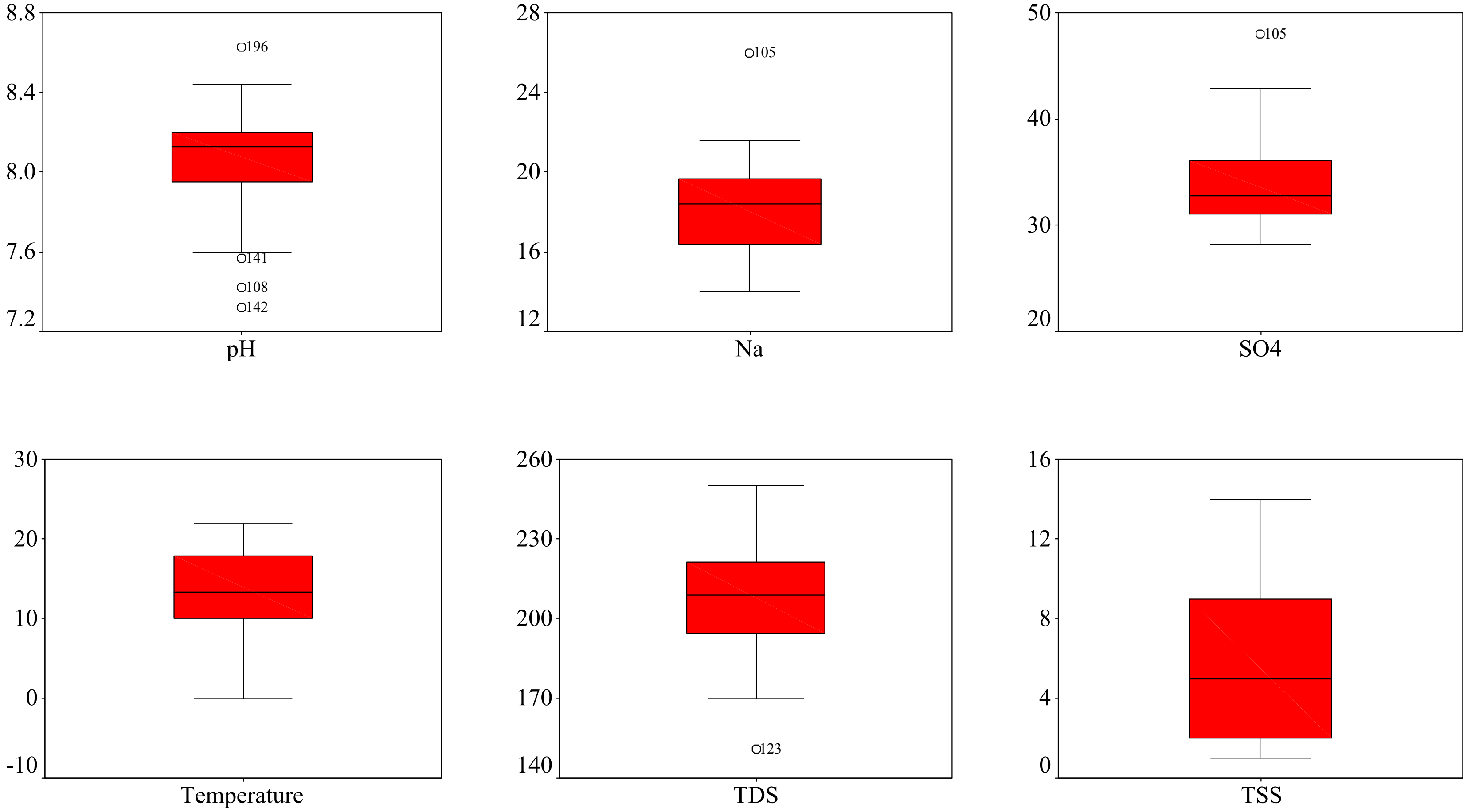


Figure 29 (cont.) Location MB05SHS004, WQ variables Box-plots

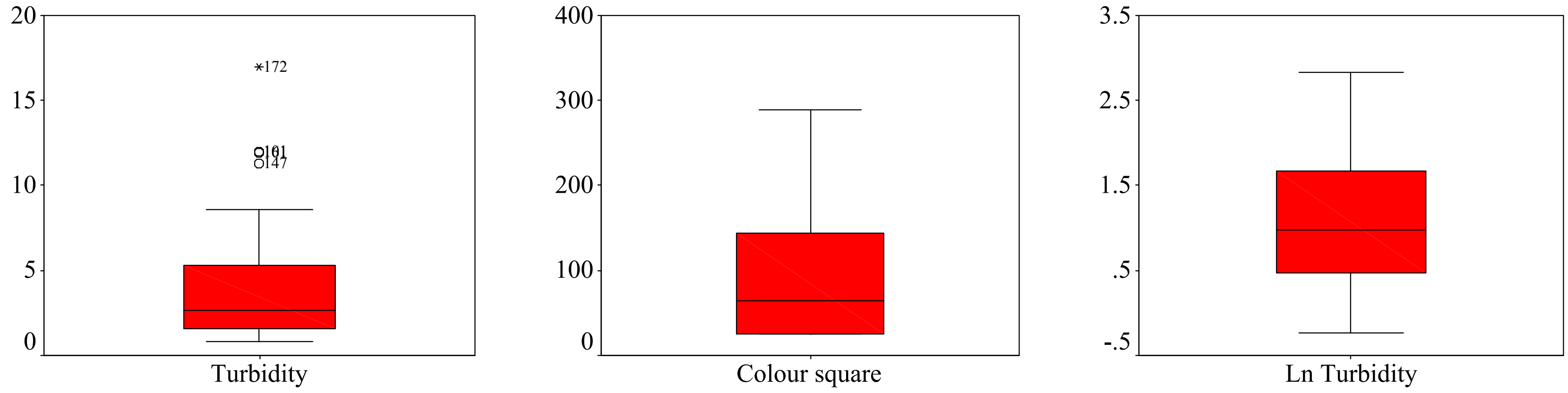


Figure 29 (cont.) Location MB05SHS004, WQ variables Box-plots

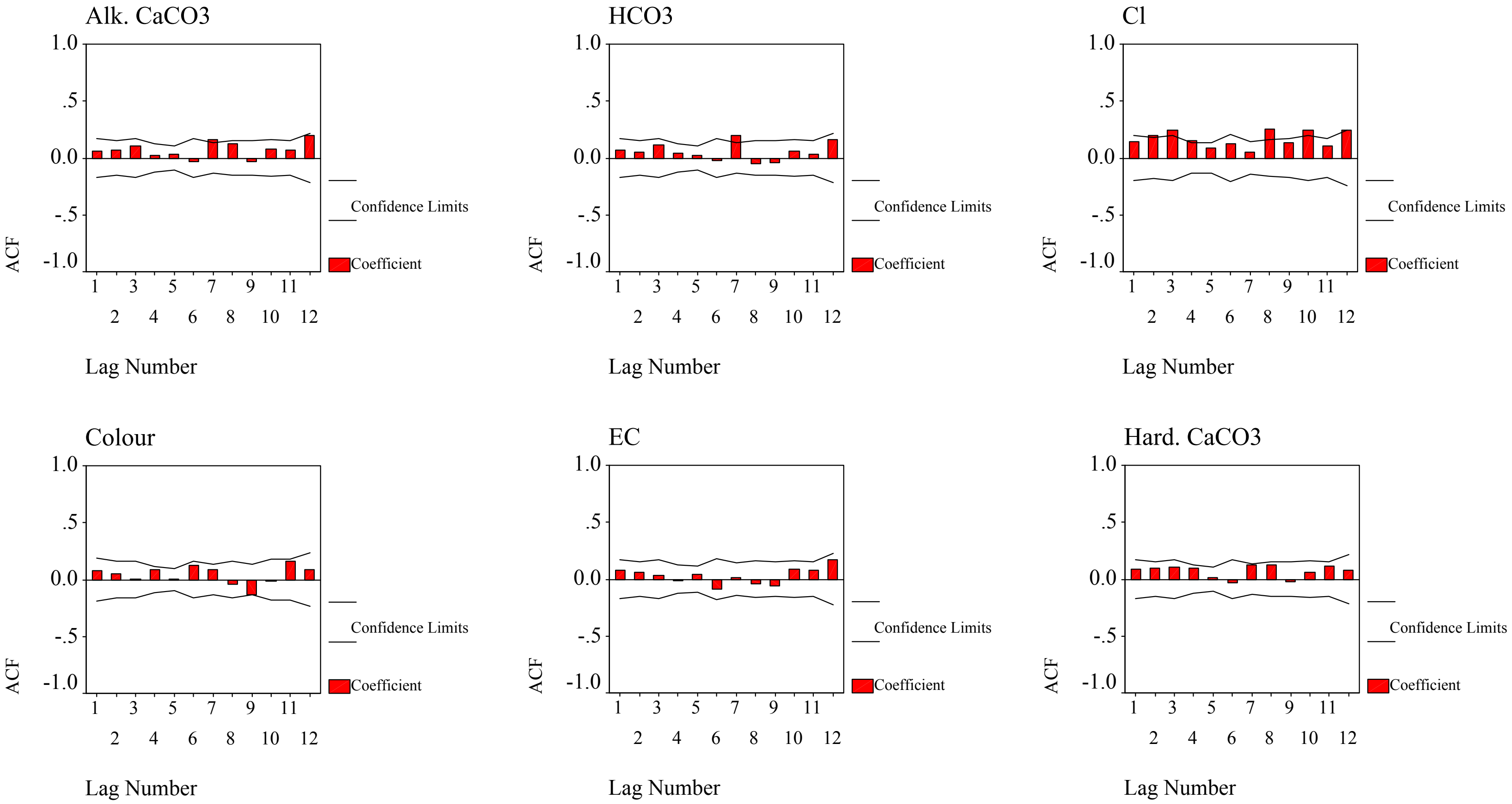


Figure 30. Location MB05SHS004, WQ variables Corellogram

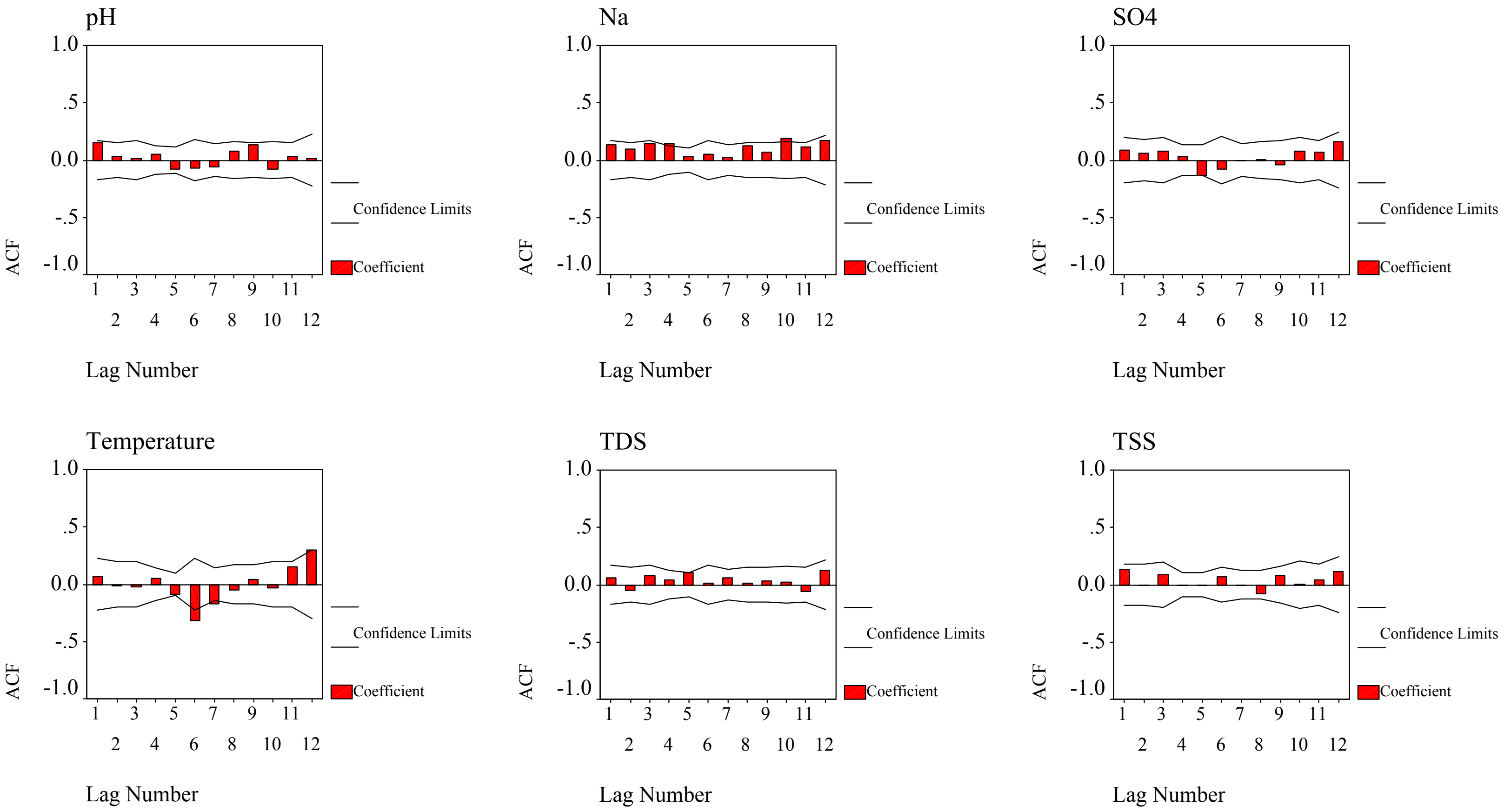
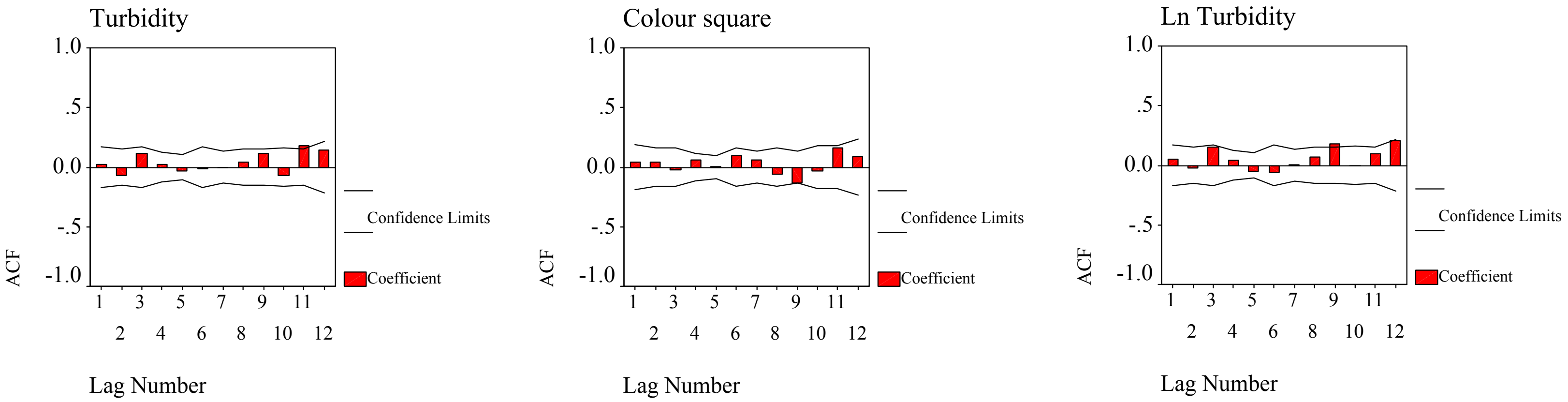


Figure 30 (cont.) Location MB05SHS004, WQ variables Corellogram


Turbidity

Colour square

Ln Turbidity

Figure 30 (cont.) Location MB05SHS004, WQ variables Corellogram

Location MB05SHS014

Table 31. Location MB05SHS014, WQ variables descriptive statistics

WQ variable	$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Measuring unit	mg/l	mg/l	mg/l	CU	US/cm	mg/l	pH units	mg/l	mg/l	deg C	mg/l	mg/l	NTU
n	30	30	30	28	31	30	31	30	30	20	30	25	30
Mean	105.79	128.00	21.06	10.04	333.87	122.62	7.90	18.48	31.99	10.65	196.10	3.92	3.56
Median	106.00	129.00	21.10	10.00	329.00	119.50	8.01	18.30	32.70	10.50	194.00	3.00	2.30
Std. Deviation	9.26	11.64	4.00	3.58	29.03	13.21	0.35	2.47	3.15	7.77	18.08	3.26	2.96
Variance	85.74	135.52	16.01	12.85	842.92	174.55	0.12	6.10	9.90	60.43	326.78	10.66	8.75
Skewness	-0.31	-0.37	0.20	0.31	0.29	-0.10	-0.57	0.24	-0.45	-0.23	0.00	1.37	2.16
Std. Error of Skewness	0.43	0.43	0.43	0.44	0.42	0.43	0.42	0.43	0.43	0.51	0.43	0.46	0.43
Kurtosis	0.89	0.83	0.51	-0.33	0.12	-0.82	-0.29	1.11	-0.28	-1.42	-0.70	1.16	5.64
Std. Error of Kurtosis	0.83	0.83	0.83	0.86	0.82	0.83	0.82	0.83	0.83	0.99	0.83	0.90	0.83
Range	44.00	54.00	18.80	13.00	121.00	49.00	1.40	11.90	12.00	21.60	65.00	11.00	13.60
Minimum	82.00	100.00	12.60	5.00	278.00	98.00	7.17	12.80	25.50	0.00	164.00	1.00	1.00
Maximum	126.00	154.00	31.40	18.00	399.00	147.00	8.57	24.70	37.50	21.60	229.00	12.00	14.60

Table 32. Location MB05SHS014, WQ variables Kolmogrov-Smirnov test

Test parameters		$CaCO_3$	HCO_3	Cl	Color	EC	$H.CaCO_3$	pH	Na	SO_4	Temp.	TDS	TSS	Turb.
Most Extreme Differences	**Absolute**	0.125	0.132	0.086	0.139	0.120	0.145	0.146	0.103	0.147	0.164	0.078	0.251	0.255
	Positive	0.125	0.107	0.086	0.111	0.120	0.112	0.094	0.102	0.078	0.164	0.078	0.251	0.255
	Negative	-0.108	-0.132	-0.084	-0.139	-0.091	-0.145	-0.146	-0.103	-0.147	-0.128	-0.068	-0.186	-0.193
Kolmogorov-Smirnov Z		0.682	0.725	0.469	0.735	0.668	0.795	0.816	0.562	0.807	0.733	0.426	1.255	1.397
Asymp. Sig. (2-tailed)		0.741	0.669	0.980	0.653	0.763	0.552	0.519	0.910	0.533	0.656	0.993	0.086	0.040

Table 33. Location MB05SHS014, transformed WQ variables descriptive statistics

WQ variable	Ln Turb.
Measuring unit	NTU
n	30
Mean	1.02
Median	0.83
Std. Deviation	0.68
Variance	0.46
Skewness	0.65
Std. Error of Skewness	0.43
Kurtosis	-0.22
Std. Error of Kurtosis	0.83
Range	2.68
Minimum	0.00
Maximum	2.68

Table 34. Location MB05SHS014, transformed WQ variables Kolmogrov-Smirnov test

Test parameters		Ln Turb.
Most Extreme Differences	Absolute	0.145
	Positive	0.145
	Negative	-0.088
Kolmogorov-Smirnov Z		0.793
Asymp. Sig. (2-tailed)		0.556

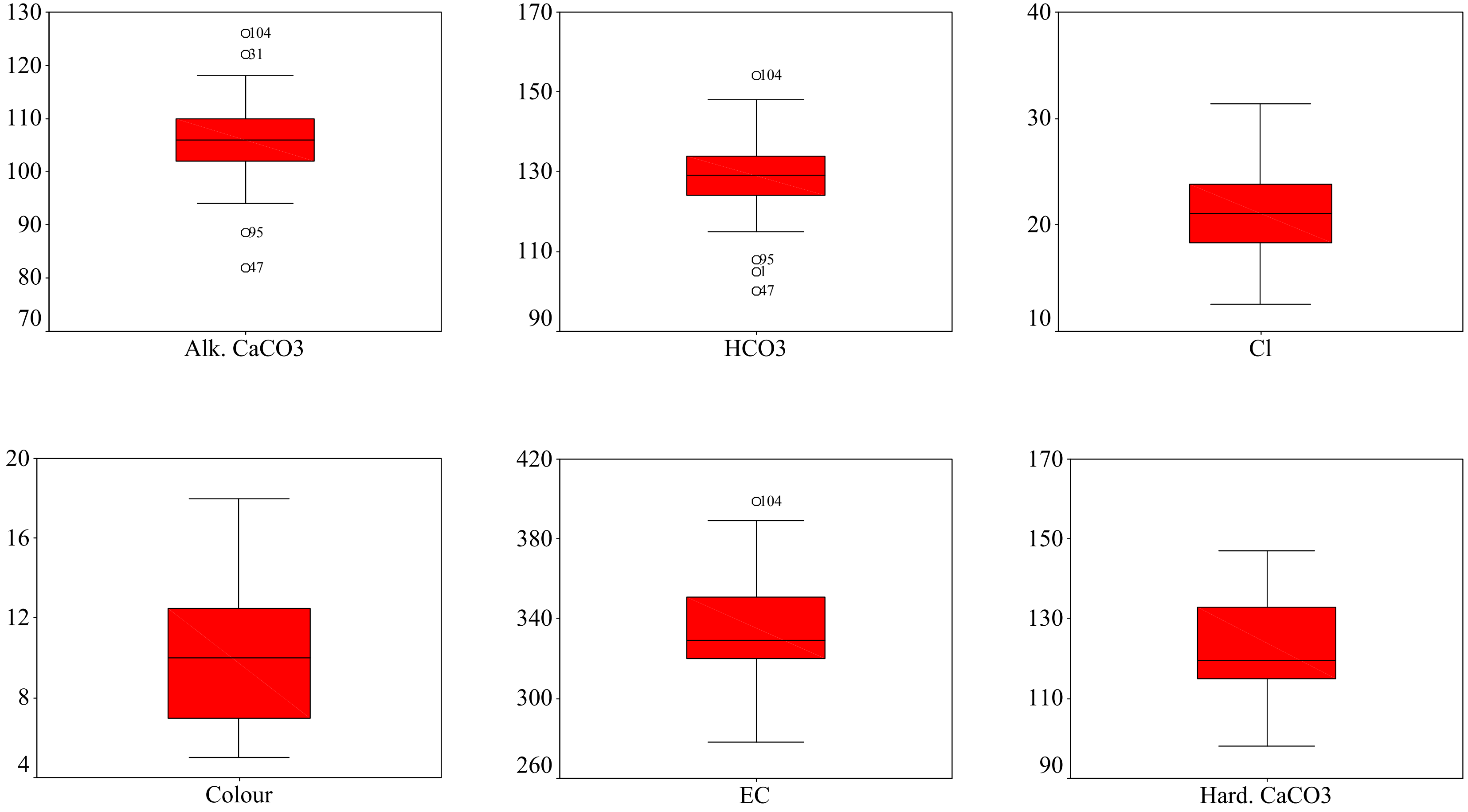


Figure 31. Location MB05SHS014, WQ variables Box-plots

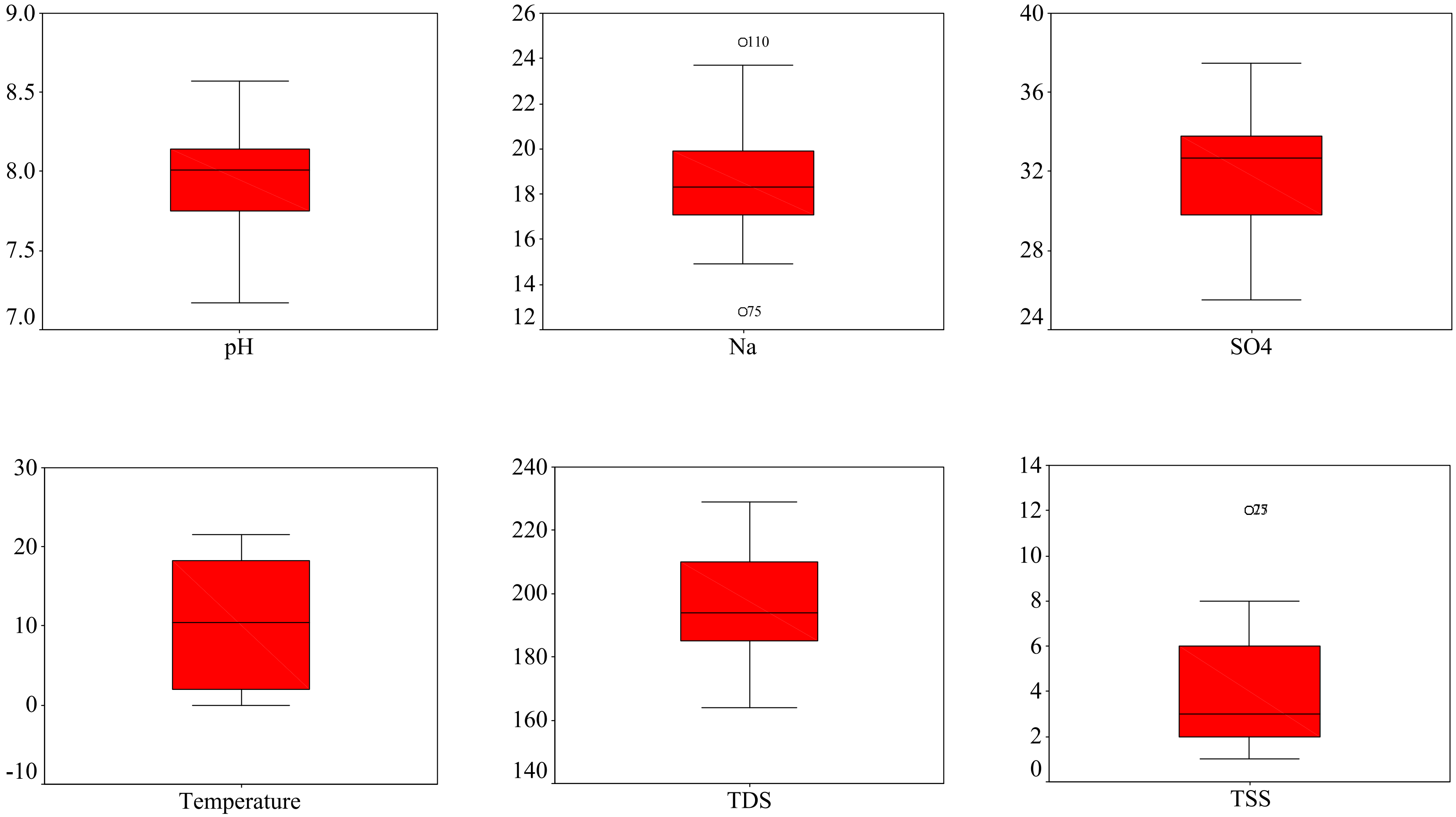


Figure 31 (cont.) Location MB05SHS014, WQ variables Box-plots

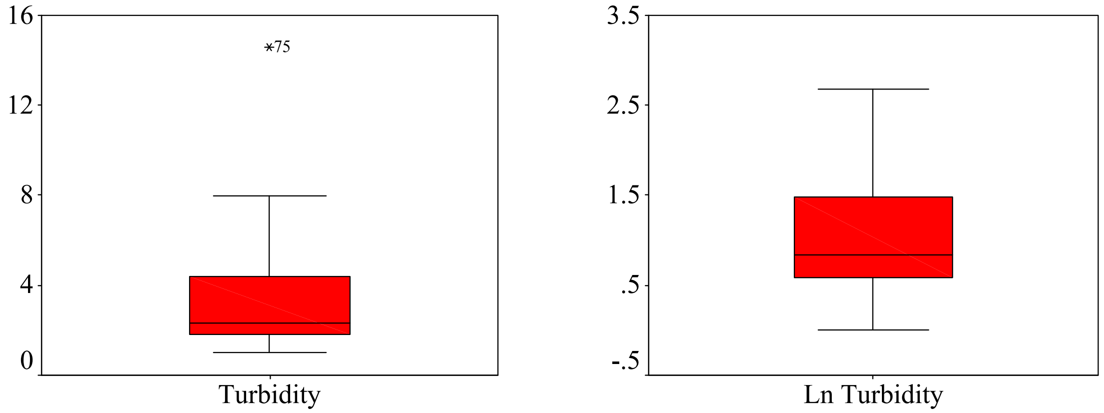


Figure 31 (cont.) Location MB05SHS014, WQ variables Box-plots

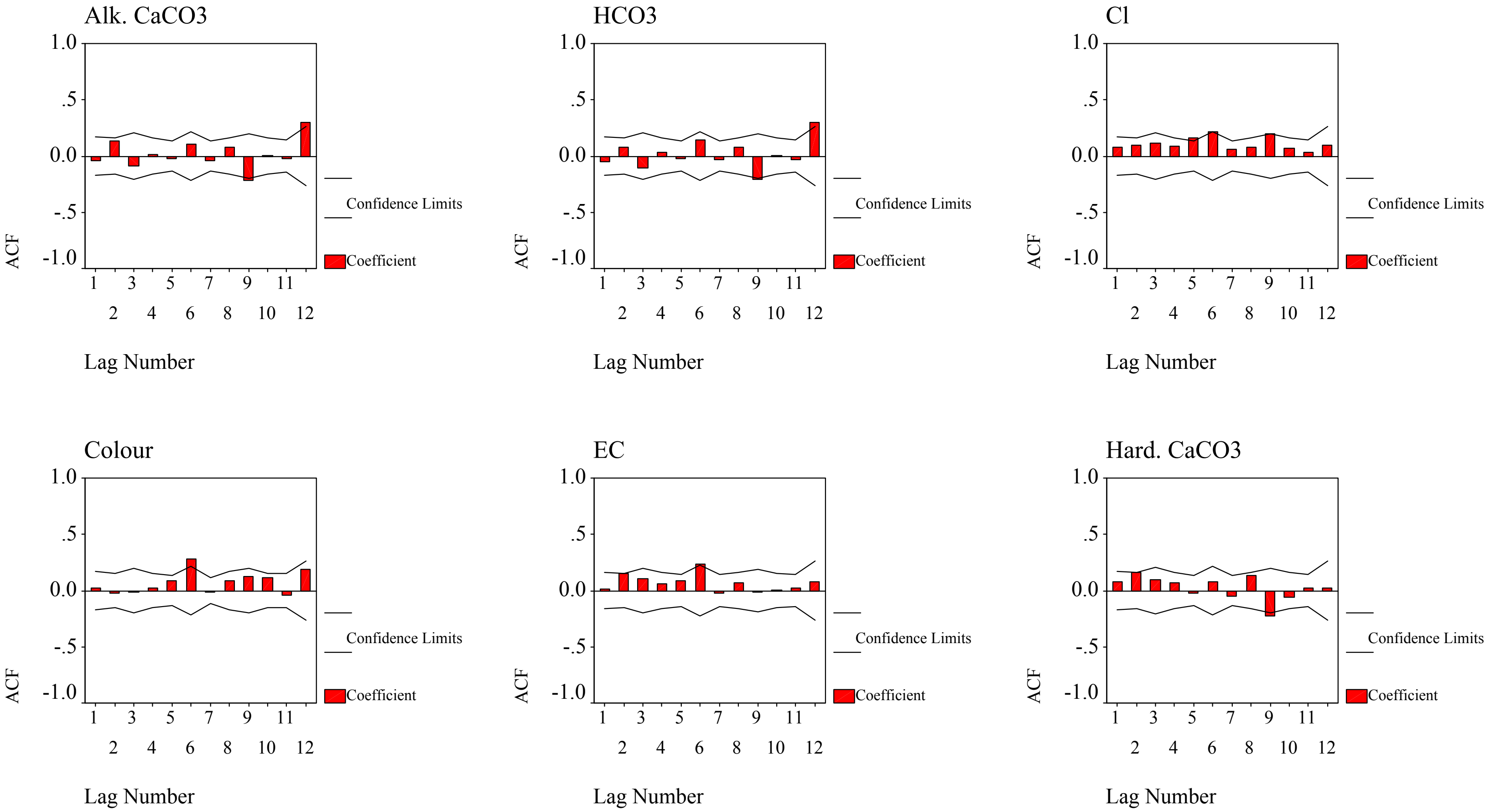


Figure 32. Location MB05SHS014, WQ variables Corellogram

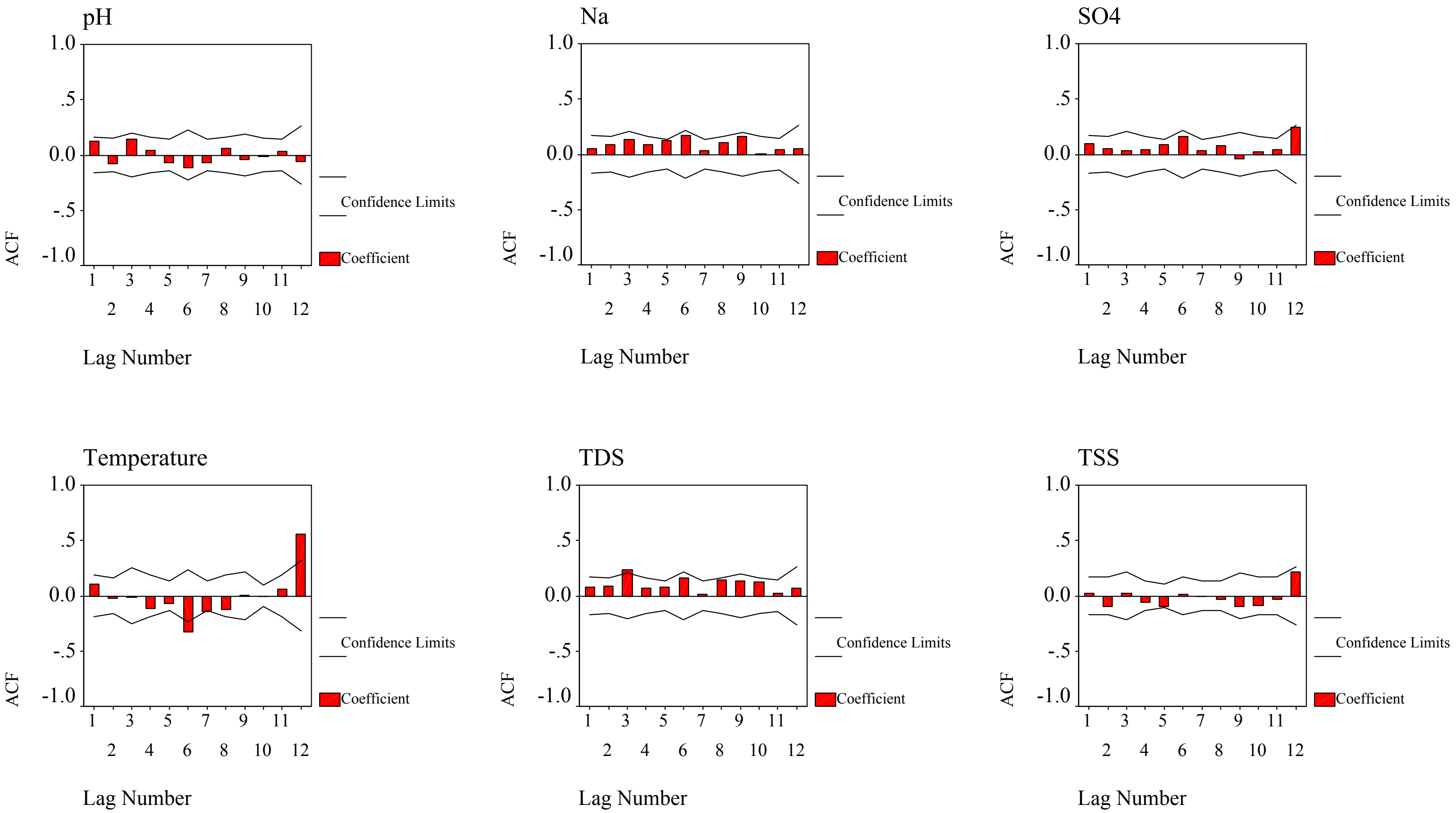


Figure 32 (cont.) Location MB05SHS014, WQ variables Corellogram

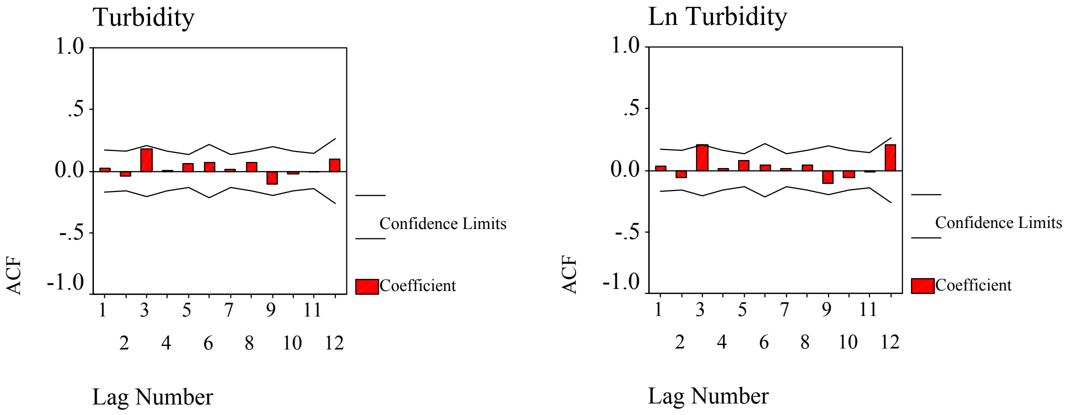


Figure 32 (cont.) Location MB05SHS014, WQ variables Corellogram

Table 35. Mann-Kendall nonparametric trend test results

Location	MB05RAS078		MB05RBS003		MB05SAS004		MB05SBS126		MB05SCS005		MB05SHS004		MB05SHS014	
WQ variable	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**	**Z**	**p-value**
$CaCO_3$	0.41	0.68	-0.12	0.91	-1.57	0.12	-0.19	0.85	0.49	0.62	-2.29	**0.02**	0.38	0.71
HCO_3	0.36	0.72	-0.08	0.93	-1.39	0.16	0.00	1.00	0.56	0.57	-1.23	0.22	1.14	0.25
Cl	-0.56	0.57	-1.36	0.17	-1.87	0.06	-1.16	0.25	-1.34	0.18	-2.75	**0.01**	-1.12	0.26
Colour	-1.16	0.25	0.74	0.46	-0.32	0.75	-1.04	0.30	-1.53	0.13	1.22	0.22	2.27	**0.02**
EC	0.25	0.80	-0.10	0.92	-1.65	0.10	0.41	0.68	0.61	0.54	-1.88	0.06	0.75	0.45
H. $CaCO_3$	0.24	0.81	-0.09	0.93	-0.75	0.45	0.27	0.79	0.25	0.80	-2.94	**0.003**	0.02	0.99
Ph	1.47	0.14	-0.76	0.44	0.94	0.35	0.31	0.75	0.83	0.40	-1.58	0.11	0.56	0.57
Na	0.15	0.88	-0.51	0.61	-1.86	0.06	0.32	0.75	0.46	0.65	-1.38	0.17	-0.98	0.33
SO_4	-0.13	0.90	-0.31	0.76	-2.49	**0.01**	0.22	0.83	-0.03	0.98	-1.54	0.12	-1.27	0.21
Temp.	-0.31	0.76	0.27	0.78	0.53	0.60	1.51	0.13	1.44	0.15	0.25	0.81	0.20	0.85
TDS	0.17	0.86	0.58	0.56	-1.64	0.10	0.05	0.96	0.50	0.62	-2.27	**0.02**	-0.32	0.75
TSS	-1.08	0.28	-2.13	**0.03**	0.04	0.97	-1.22	0.22	-1.28	0.20	0.18	0.86	0.33	0.74
Turb.	-1.07	0.28	-1.24	0.22	-0.18	0.86	-1.69	0.09	-1.31	0.19	1.17	0.24	1.32	0.19

Appendix 5. Detailed results for the assessment of water quality variables

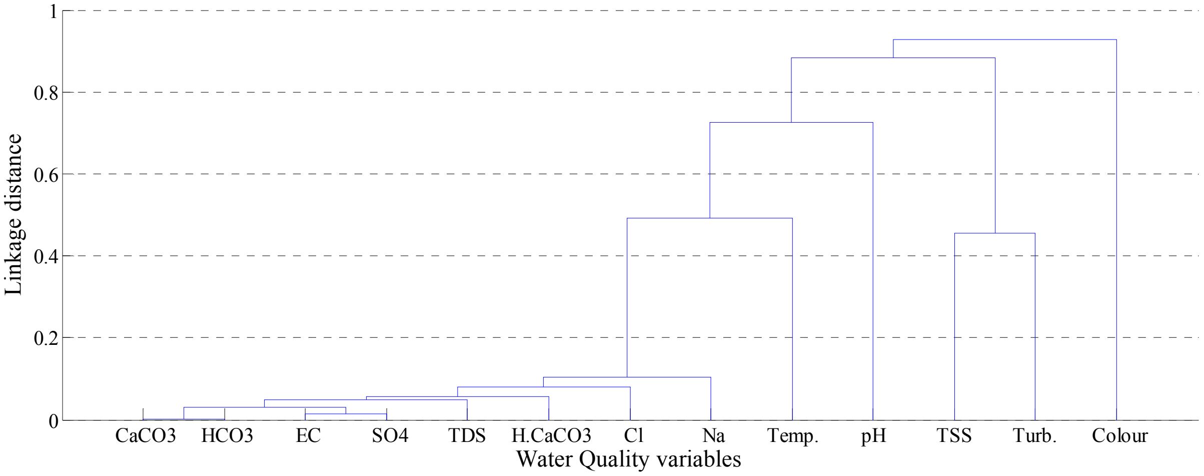


Figure 33. Cluster tree for water quality variables at MB05RAS078

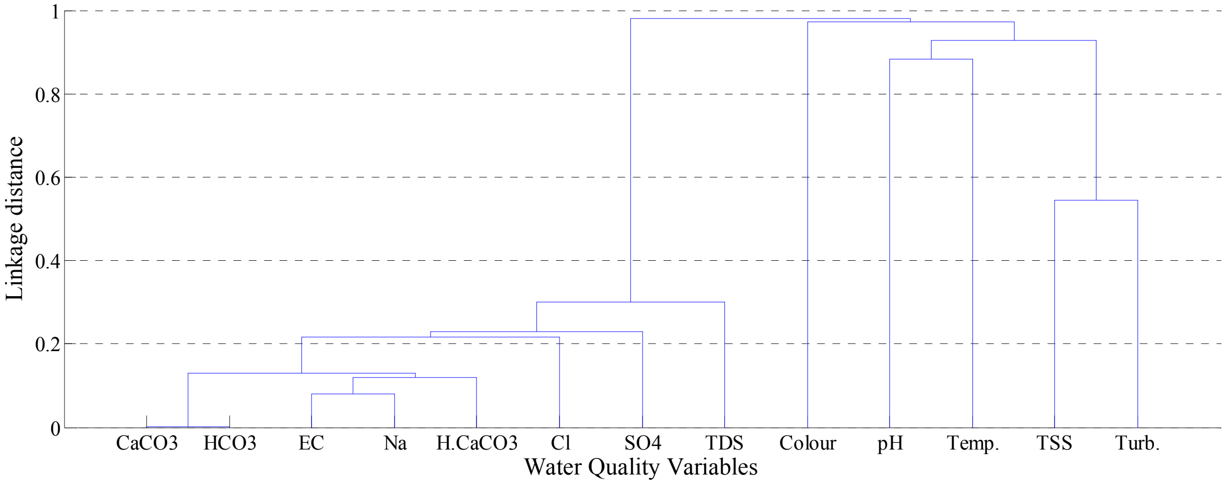


Figure 34. Cluster tree for water quality variables at MB05RBS003

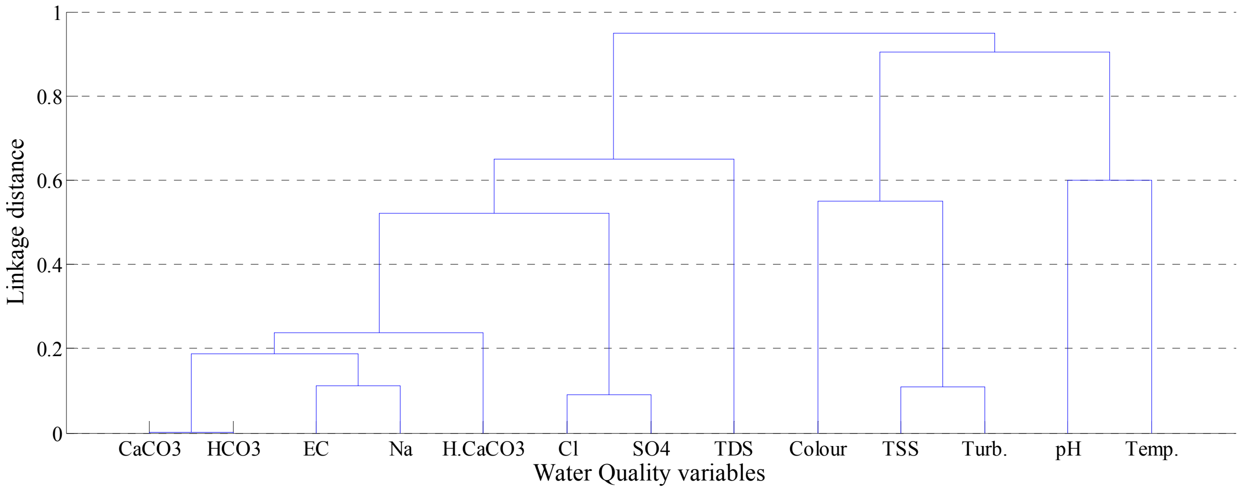


Figure 35. Cluster tree for water quality variables at MB05SAS004

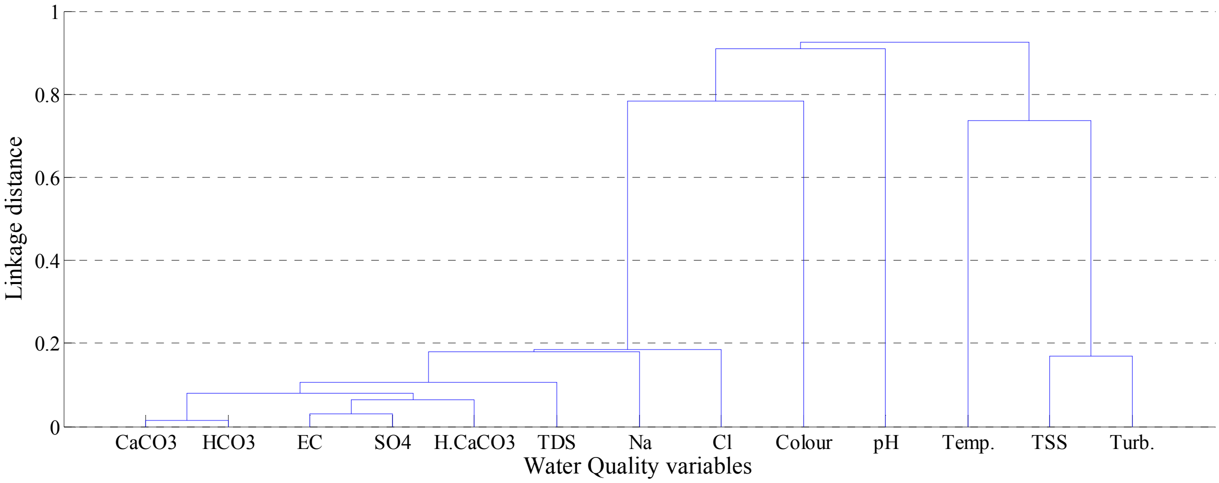


Figure 36. Cluster tree for water quality variables at MB05SBS126

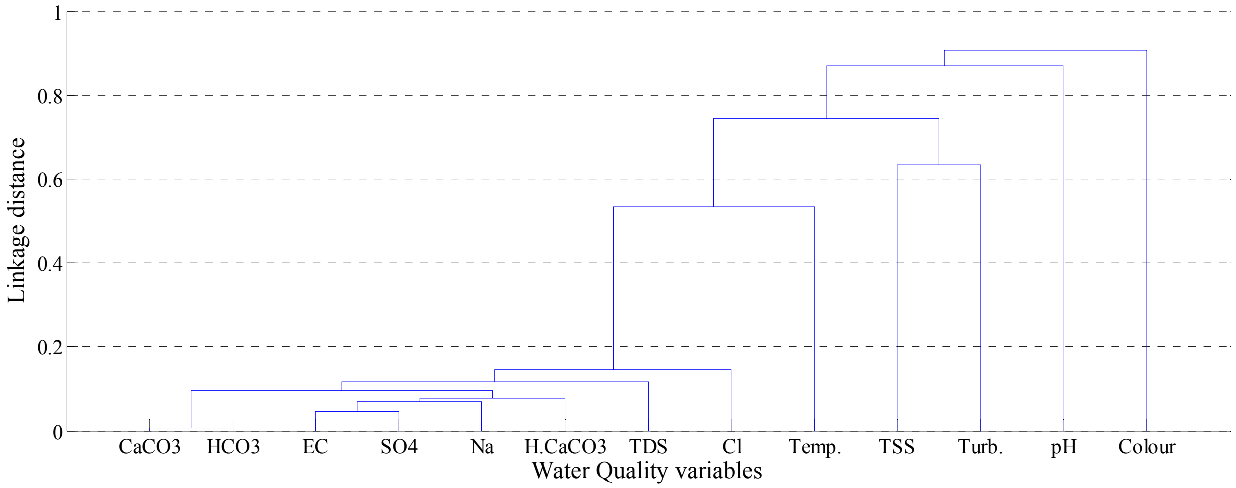


Figure 37. Cluster tree for water quality variables at MB05SCS005

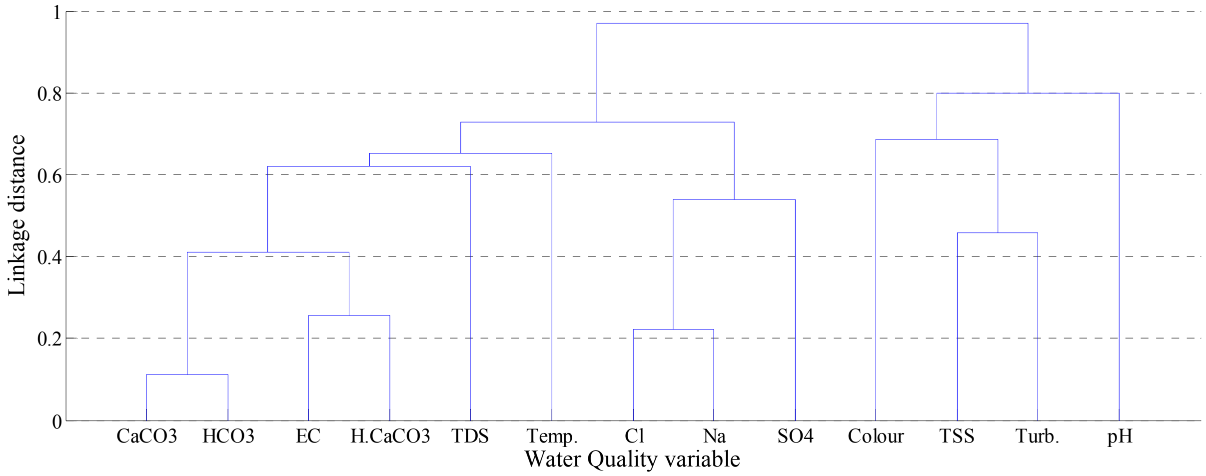


Figure 38. Cluster tree for water quality variables at MB05SHS004

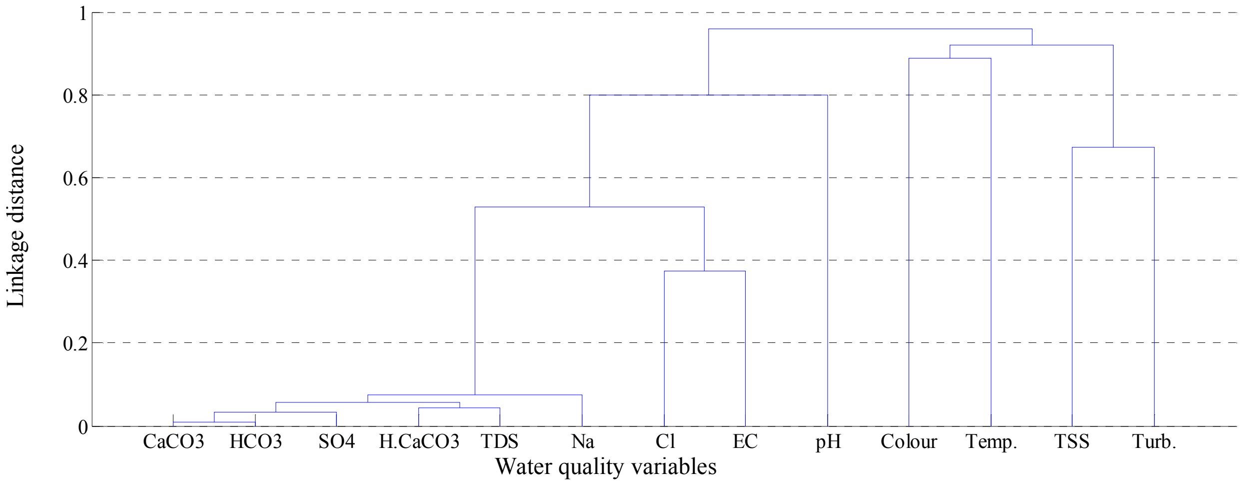


Figure 39. Cluster tree for water quality variables at MB05SHS014

Table 36. Variable to discontinue and its best auxiliary variable (MB05RAS078)

Discontinued variable	**Best auxiliary variable**	r	**Number of samples**	$Var(\hat{\mu})$	I_a
Na	$H.CaCO_3$	0.93	31	0.14	2.03
$H.CaCO_3$	Na	0.93	31	0.14	2.03
TDS	Na	0.94	31	0.15	2.04
HCO_3	Na	0.95	31	0.15	2.04
$CaCO_3$	Na	0.95	31	0.15	2.04
EC	Na	0.96	31	0.16	2.06
SO_4	Na	0.96	29	0.17	2.07
Cl	TDS	0.95	27	0.18	2.07

Table 37. Variable to discontinue and its best auxiliary variable (MB05RBS003)

Discontinued variable	Best auxiliary variable	r	Number of samples	$Var(\hat{\mu})$	I_a
TDS	Na	0.84	33	0.001	1.63
Na	TDS	0.84	33	0.001	1.63
Cl	SO_4	0.84	33	0.01	1.67
SO_4	Cl	0.84	33	0.01	1.67
$CaCO_3$	Na	0.91	34	0.01	1.7
HCO3	Na	0.91	34	0.01	1.7
$H.CaCO_3$	SO_4	0.85	33	0.02	1.71
EC	TDS	0.85	33	0.02	1.72

Table 38. Variable to discontinue and its best auxiliary variable (MB05SAS004)

Discontinued variable	Best auxiliary variable	r	Number of samples	$Var(\hat{\mu})$	I_a
Na	$CaCO_3$	0.88	30	0.13	2.07
$CaCO_3$	Na	0.88	30	0.13	2.07
HCO_3	Na	0.88	30	0.13	2.07
$H.CaCO_3$	$CaCO_3$	0.85	29	0.14	2.07
EC	$CaCO_3$	0.92	30	0.15	2.09
TSS	Turb.	0.94	28	0.17	2.12
Turb.	TSS	0.94	28	0.17	2.12
Cl	SO_4	0.95	25	0.19	2.13
SO_4	Cl	0.95	25	0.19	2.13

Table 39. Variable to discontinue and its best auxiliary variable (MB05SBS126)

Discontinued variable	Best auxiliary variable	r	Number of samples	$Var(\hat{\mu})$	I_a
HCO_3	Cl	0.88	32	0.09	1.93
Cl	HCO_3	0.88	32	0.09	1.93
TDS	Cl	0.88	32	0.09	1.94
$CaCO_3$	Cl	0.9	32	0.1	1.95
SO_4	Cl	0.92	32	0.11	1.98
Na	TDS	0.87	30	0.13	1.99
EC	Cl	0.93	32	0.12	1.99
$H.CaCO_3$	Cl	0.91	30	0.14	2.01
TSS	Turb.	0.91	27	0.18	2.05
Turb.	TSS	0.91	27	0.18	2.06

Table 40. Variable to discontinue and its best auxiliary variable (MB05SCS005)

Discontinued variable	**Best auxiliary variable**	r	**Number of samples**	$Var(\hat{\mu})$	I_a
Cl	HCO_3	0.89	34	-0.02	1.56
EC	HCO_3	0.95	35	-0.01	1.56
Na	TDS	0.94	35	-0.04	1.56
$H.CaCO_3$	TDS	0.94	35	-0.04	1.56
$CaCO_3$	TDS	0.93	35	-0.1	1.56
TDS	HCO_3	0.92	35	-0.13	1.56
HCO_3	TDS	0.92	35	-0.13	1.56
SO_4	HCO_3	0.94	34	0.07	1.82

Table 41. Variable to discontinue and its best auxiliary variable (MB05SHS004)

Discontinued variable	**Best auxiliary variable**	r	**Number of samples**	$Var(\hat{\mu})$	I_a
$H.CaCO_3$	EC	0.86	30	0.13	2.06
EC	$H.CaCO_3$	0.86	30	0.13	2.06
Na	Cl	0.88	29	0.15	2.09
Cl	Na	0.88	29	0.15	2.09
$CaCO_3$	HCO_3	0.94	29	0.17	2.11
HCO_3	$CaCO_3$	0.94	29	0.17	2.11

Table 42. Variable to discontinue and its best auxiliary variable (MB05SHS014)

Discontinued variable	**Best auxiliary variable**	r	**Number of samples**	$Var(\hat{\mu})$	I_a
HCO_3	TDS	0.96	31	0.16	2.06
TDS	HCO_3	0.96	31	0.16	2.06
Na	HCO_3	0.95	30	0.16	2.06
$CaCO_3$	TDS	0.97	31	0.16	2.07
SO_4	TDS	0.97	31	0.16	2.07
$H.CaCO_3$	HCO_3	0.97	31	0.16	2.07

Table 43. Combinations of two variables to discontinue

MB05RAS078		I_a	**MB05RBS003**		I_a	**MB05SAS004**		I_a
$H.CaCO_3$	TDS	1.44	TDS	HCO_3	1.1	$H.CaCO_3$	TSS	1.51
$H.CaCO_3$	HCO_3	1.45	TDS	$H.CaCO_3$	1.1	$CaCO_3$	Turb.	1.51
$H.CaCO_3$	$CaCO_3$	1.45	Na	$H.CaCO_3$	1.1	Na	Turb.	1.51
TDS	HCO_3	1.45	SO_4	$CaCO_3$	1.11	HCO_3	Turb.	1.51
TDS	$CaCO_3$	1.45	Cl	$CaCO_3$	1.11	$H.CaCO_3$	Turb.	1.51
HCO_3	$CaCO_3$	1.46	SO_4	HCO_3	1.11	Na	Cl	1.51
$H.CaCO_3$	Cl	1.46	Cl	HCO_3	1.11	$CaCO_3$	Cl	1.51
$H.CaCO_3$	SO_4	1.46	Cl	$H.CaCO_3$	1.11	HCO_3	Cl	1.51
TDS	EC	1.46	Na	EC	1.11	$H.CaCO_3$	Cl	1.51
TDS	SO_4	1.47	SO_4	EC	1.11	$CaCO_3$	SO_4	1.51
HCO_3	EC	1.47	Cl	EC	1.11	Na	SO_4	1.51
$H.CaCO_3$	Cl	1.47	$CaCO_3$	HCO_3	1.11	HCO_3	SO_4	1.51
Na	Cl	1.47	$CaCO_3$	$H.CaCO_3$	1.12	$H.CaCO_3$	SO_4	1.52
$CaCO_3$	EC	1.47	HCO_3	$H.CaCO_3$	1.12	EC	TSS	1.52
HCO_3	SO_4	1.47	$CaCO_3$	EC	1.12	EC	Turb.	1.52
$CaCO_3$	SO_4	1.47	HCO_3	EC	1.12	EC	Cl	1.53
HCO_3	Cl	1.48	$H.CaCO_3$	EC	1.12	EC	SO_4	1.53
$CaCO_3$	Cl	1.48	**MB05SAS004**		I_a	TSS	Cl	1.55
EC	SO_4	1.48	$CaCO_3$	HCO_3	1.47	Turb.	Cl	1.55
EC	Cl	1.49	Na	$H.CaCO_3$	1.47	TSS	SO_4	1.55
SO_4	Cl	1.49	HCO_3	$H.CaCO_3$	1.47	Turb.	SO_4	1.55
MB05RBS003		I_a	Na	EC	1.48	**MB05SBS126**		I_a
TDS	SO_4	1.09	HCO_3	EC	1.48	HCO_3	TDS	1.32
TDS	Cl	1.09	$H.CaCO_3$	EC	1.48	HCO_3	$CaCO_3$	1.33
Na	SO_4	1.09	$CaCO_3$	TSS	1.5	TDS	$CaCO_3$	1.33
Na	Cl	1.09	Na	TSS	1.5	HCO_3	SO_4	1.34
TDS	$CaCO_3$	1.1	HCO_3	TSS	1.5	TDS	SO_4	1.35

Table 43 (cont). Combinations of two variables to discontinue

MB05SBS126		I_a	**MB05SBS126**		I_a	**MB05SCS005**		I_a
Cl	Na	1.35	SO_4	Turb.	1.43	Cl	SO_4	1.12
HCO_3	Na	1.35	EC	TSS	1.43	EC	SO_4	1.13
$CaCO_3$	SO_4	1.35	Na	Turb.	1.44	**MB05SHS004**		I_a
HCO_3	EC	1.36	EC	Turb.	1.44	H.$CaCO_3$	Na	1.47
TDS	EC	1.36	H.$CaCO_3$	TSS	1.44	EC	Na	1.48
$CaCO_3$	Na	1.36	H.$CaCO_3$	Turb.	1.46	H.$CaCO_3$	Cl	1.48
HCO_3	H.$CaCO_3$	1.37	**MB05SCS005**		I_a	EC	Cl	1.48
$CaCO_3$	EC	1.37	HCO_3	$CaCO_3$	0.84	H.$CaCO_3$	HCO_3	1.49
TDS	H.$CaCO_3$	1.37	HCO_3	H.$CaCO_3$	0.9	H.$CaCO_3$	$CaCO_3$	1.49
SO_4	Na	1.38	HCO_3	Na	0.9	EC	$CaCO_3$	1.49
$CaCO_3$	H.$CaCO_3$	1.38	$CaCO_3$	H.$CaCO_3$	0.92	EC	HCO_3	1.49
SO_4	EC	1.38	TDS	Cl	0.92	Na	$CaCO_3$	1.51
EC	Na	1.39	$CaCO_3$	Na	0.93	Na	HCO_3	1.51
Cl	TSS	1.39	TDS	EC	0.93	Cl	HCO_3	1.51
HCO_3	TSS	1.39	$CaCO_3$	Cl	0.95	Cl	$CaCO_3$	1.51
SO_4	H.$CaCO_3$	1.39	$CaCO_3$	EC	0.96	**MB05SHS014**		I_a
TDS	TSS	1.4	H.$CaCO_3$	Na	0.99	TDS	Na	1.48
Na	H.$CaCO_3$	1.4	H.$CaCO_3$	Cl	1.01	HCO_3	$CaCO_3$	1.48
Cl	Turb.	1.4	TDS	SO_4	1.01	Na	$CaCO_3$	1.48
HCO_3	Turb.	1.4	Na	Cl	1.01	HCO_3	SO_4	1.48
$CaCO_3$	TSS	1.41	H.$CaCO_3$	EC	1.02	Na	SO_4	1.48
TDS	Turb.	1.41	Na	EC	1.02	TDS	H.$CaCO_3$	1.49
EC	H.$CaCO_3$	1.41	$CaCO_3$	SO_4	1.03	Na	H.$CaCO_3$	1.49
$CaCO_3$	Turb.	1.42	Cl	EC	1.04	$CaCO_3$	SO_4	1.49
SO_4	TSS	1.42	H.$CaCO_3$	SO_4	1.09	$CaCO_3$	H.$CaCO_3$	1.49
Na	TSS	1.43	Na	SO_4	1.1	SO_4	H.$CaCO_3$	1.49

Table 44. Combinations of three variables to discontinue

MB05RAS078			I_a	MB05RAS078			I_a
H.CaCO$_3$	TDS	HCO$_3$	1.28	HCO$_3$	SO$_4$	Cl	1.32
H.CaCO$_3$	TDS	CaCO$_3$	1.28	CaCO$_3$	SO$_4$	Cl	1.32
H.CaCO$_3$	HCO$_3$	CaCO$_3$	1.28	EC	SO$_4$	Cl	1.34
TDS	HCO$_3$	CaCO$_3$	1.29	**MB05RBS003**			I_a
H.CaCO$_3$	TDS	EC	1.29	TDS	SO$_4$	CaCO$_3$	0.79
H.CaCO$_3$	TDS	SO$_4$	1.29	TDS	Cl	CaCO$_3$	0.79
H.CaCO$_3$	HCO$_3$	EC	1.29	TDS	SO$_4$	HCO$_3$	0.8
H.CaCO$_3$	CaCO3	EC	1.29	TDS	Cl	HCO$_3$	0.8
H.CaCO$_3$	HCO$_3$	SO$_4$	1.3	TDS	Cl	H.CaCO$_3$	0.8
TDS	HCO$_3$	EC	1.3	Na	Cl	H.CaCO$_3$	0.8
H.CaCO$_3$	CaCO$_3$	SO$_4$	1.3	TDS	CaCO$_3$	HCO$_3$	0.8
TDS	CaCO$_3$	EC	1.3	Na	SO$_4$	EC	0.8
TDS	HCO$_3$	SO$_4$	1.3	Na	Cl	EC	0.8
TDS	CaCO$_3$	SO$_4$	1.3	TDS	CaCO$_3$	H.CaCO$_3$	0.81
H.CaCO$_3$	HCO$_3$	Cl	1.3	TDS	HCO3	H.CaCO$_3$	0.81
HCO$_3$	CaCO$_3$	EC	1.3	SO$_4$	CaCO$_3$	HCO$_3$	0.81
H.CaCO$_3$	CaCO$_3$	Cl	1.3	Cl	CaCO$_3$	HCO$_3$	0.81
HCO$_3$	CaCO$_3$	SO$_4$	1.31	Cl	CaCO$_3$	H.CaCO$_3$	0.81
H.CaCO$_3$	EC	SO$_4$	1.31	Cl	HCO$_3$	H.CaCO$_3$	0.81
TDS	EC	SO$_4$	1.31	Na	H.CaCO$_3$	EC	0.82
HCO$_3$	CaCO$_3$	Cl	1.31	SO$_4$	CaCO$_3$	EC	0.82
H.CaCO$_3$	EC	Cl	1.31	Cl	CaCO$_3$	EC	0.82
HCO$_3$	EC	SO$_4$	1.32	SO$_4$	HCO$_3$	EC	0.82
H.CaCO$_3$	SO$_4$	Cl	1.32	Cl	HCO$_3$	EC	0.82
CaCO$_3$	EC	SO$_4$	1.32	Cl	H.CaCO$_3$	EC	0.82
HCO$_3$	EC	Cl	1.32	CaCO$_3$	HCO$_3$	H.CaCO$_3$	0.82
CaCO$_3$	EC	Cl	1.32	CaCO$_3$	HCO$_3$	EC	0.82

Table 44 (cont). Combinations of three variables to discontinue

MB05RBS003			I_a	**MB05SAS004**			I_a
$CaCO_3$	H.$CaCO_3$	EC	0.83	HCO_3	EC	SO_4	1.34
HCO_3	H.$CaCO_3$	EC	0.83	H.$CaCO_3$	EC	SO_4	1.35
MB05SAS004			I_a	$CaCO_3$	TSS	Cl	1.36
Na	H.$CaCO_3$	EC	1.3	Na	TSS	Cl	1.36
HCO_3	H.$CaCO_3$	EC	1.3	HCO_3	TSS	Cl	1.36
$CaCO_3$	HCO_3	TSS	1.32	H.$CaCO_3$	TSS	Cl	1.36
Na	H.$CaCO_3$	TSS	1.32	$CaCO_3$	Turb.	Cl	1.36
HCO_3	H.$CaCO_3$	TSS	1.32	Na	Turb.	Cl	1.36
$CaCO_3$	HCO_3	Turb.	1.32	$CaCO_3$	TSS	SO_4	1.37
Na	H.$CaCO_3$	Turb.	1.32	Na	TSS	SO_4	1.37
HCO_3	H.$CaCO_3$	Turb.	1.32	HCO_3	Turb.	Cl	1.37
$CaCO_3$	HCO_3	Cl	1.32	HCO_3	TSS	SO_4	1.37
Na	H.$CaCO_3$	Cl	1.33	H.$CaCO_3$	Turb.	Cl	1.37
HCO_3	H.$CaCO_3$	Cl	1.33	H.$CaCO_3$	TSS	SO_4	1.37
$CaCO_3$	HCO_3	SO_4	1.33	$CaCO_3$	Turb.	SO_4	1.37
Na	H.$CaCO_3$	SO_4	1.33	Na	Turb.	SO_4	1.37
HCO_3	H.$CaCO_3$	SO_4	1.33	HCO_3	Turb.	SO_4	1.37
Na	EC	TSS	1.33	H.$CaCO_3$	Turb.	SO_4	1.37
HCO_3	EC	TSS	1.33	EC	TSS	Cl	1.38
H.$CaCO_3$	EC	TSS	1.33	EC	Turb.	Cl	1.38
Na	EC	Turb.	1.33	EC	TSS	SO_4	1.38
HCO_3	EC	Turb.	1.34	EC	Turb.	SO_4	1.38
H.$CaCO_3$	EC	Turb.	1.34	**MB05SBS126**			I_a
Na	EC	Cl	1.34	HCO_3	TDS	$CaCO_3$	1.1
HCO_3	EC	Cl	1.34	HCO_3	TDS	SO_4	1.12
H.$CaCO_3$	EC	Cl	1.34	HCO_3	$CaCO_3$	SO_4	1.13
Na	EC	SO_4	1.34	TDS	$CaCO_3$	SO_4	1.13

Table 44 (cont). Combinations of three variables to discontinue

MB05SBS126			I_a	MB05SBS126			I_a
HCO_3	TDS	EC	1.13	TDS	$CaCO_3$	Turb.	1.19
HCO_3	$CaCO_3$	Na	1.13	$CaCO_3$	EC	H.$CaCO_3$	1.19
HCO_3	$CaCO_3$	EC	1.14	HCO_3	SO_4	TSS	1.19
HCO_3	TDS	H.$CaCO_3$	1.14	TDS	SO_4	TSS	1.2
TDS	$CaCO_3$	EC	1.14	Cl	Na	TSS	1.2
HCO_3	SO_4	Na	1.15	HCO_3	Na	TSS	1.2
HCO_3	$CaCO_3$	H.$CaCO_3$	1.15	SO_4	Na	H.$CaCO_3$	1.2
TDS	$CaCO_3$	H.$CaCO_3$	1.15	HCO_3	SO_4	Turb.	1.2
HCO_3	SO_4	EC	1.16	$CaCO_3$	SO_4	TSS	1.2
TDS	SO_4	EC	1.16	HCO_3	EC	TSS	1.21
$CaCO_3$	SO_4	Na	1.16	TDS	SO_4	Turb.	1.21
HCO_3	EC	Na	1.16	SO_4	EC	H.$CaCO_3$	1.21
HCO_3	SO_4	H.$CaCO_3$	1.17	TDS	EC	TSS	1.21
$CaCO_3$	SO_4	EC	1.17	Cl	Na	Turb.	1.21
HCO_3	TDS	TSS	1.17	HCO_3	Na	Turb.	1.21
TDS	SO_4	H.$CaCO_3$	1.17	$CaCO_3$	Na	TSS	1.21
HCO_3	Na	H.$CaCO_3$	1.17	EC	Na	H.$CaCO_3$	1.21
$CaCO_3$	EC	Na	1.17	$CaCO_3$	SO_4	Turb.	1.22
HCO_3	$CaCO_3$	TSS	1.18	HCO_3	EC	Turb.	1.22
$CaCO_3$	SO_4	H.$CaCO_3$	1.18	HCO_3	H.$CaCO_3$	TSS	1.22
HCO_3	TDS	Turb.	1.18	$CaCO_3$	EC	TSS	1.22
HCO_3	EC	H.$CaCO_3$	1.18	TDS	EC	Turb.	1.22
TDS	$CaCO_3$	TSS	1.18	TDS	H.$CaCO_3$	TSS	1.22
TDS	EC	H.$CaCO_3$	1.18	$CaCO_3$	Na	Turb.	1.22
$CaCO_3$	Na	H.$CaCO_3$	1.19	HCO_3	H.$CaCO_3$	Turb.	1.23
HCO_3	$CaCO_3$	Turb.	1.19	SO_4	Na	TSS	1.23
SO_4	EC	Na	1.19	$CaCO_3$	EC	Turb.	1.23

Table 44 (cont). Combinations of three variables to discontinue

MB05SBS126			I_a	**MB05SCS005**			I_a
$CaCO_3$	H.$CaCO_3$	TSS	1.23	TDS	Cl	SO_4	0.68
TDS	H.$CaCO_3$	Turb.	1.23	$CaCO_3$	Na	SO_4	0.69
SO_4	EC	TSS	1.23	H.$CaCO_3$	Cl	EC	0.69
SO_4	Na	Turb.	1.24	TDS	EC	SO_4	0.69
$CaCO_3$	H.$CaCO_3$	Turb.	1.24	Na	Cl	EC	0.7
EC	Na	TSS	1.24	$CaCO_3$	Cl	SO_4	0.71
SO_4	EC	Turb.	1.24	$CaCO_3$	EC	SO_4	0.72
SO_4	H.$CaCO_3$	TSS	1.24	H.$CaCO_3$	Na	SO_4	0.74
EC	Na	Turb.	1.25	H.$CaCO_3$	Cl	SO_4	0.76
Na	H.$CaCO_3$	TSS	1.25	Na	Cl	SO_4	0.77
SO_4	H.$CaCO_3$	Turb.	1.26	H.$CaCO_3$	EC	SO_4	0.77
EC	H.$CaCO_3$	TSS	1.26	Na	EC	SO_4	0.78
Na	H.$CaCO_3$	Turb.	1.26	Cl	EC	SO_4	0.8
EC	H.$CaCO_3$	Turb.	1.27	**MB05SHS004**			I_a
MB05SCS005			I_a	H.$CaCO_3$	Na	HCO_3	1.32
HCO_3	$CaCO_3$	H.$CaCO_3$	0.49	H.$CaCO_3$	Na	$CaCO_3$	1.32
HCO_3	$CaCO_3$	Na	0.49	EC	Na	HCO_3	1.32
HCO_3	H.$CaCO_3$	Na	0.55	EC	Na	$CaCO_3$	1.32
$CaCO_3$	H.$CaCO_3$	Na	0.58	H.$CaCO_3$	Cl	$CaCO_3$	1.32
$CaCO_3$	H.$CaCO_3$	Cl	0.6	H.$CaCO_3$	Cl	HCO_3	1.32
$CaCO_3$	Na	Cl	0.6	EC	Cl	HCO_3	1.32
$CaCO_3$	H.$CaCO_3$	EC	0.61	EC	Cl	$CaCO_3$	1.32
TDS	Cl	EC	0.61	**MB05SHS014**			I_a
$CaCO_3$	Na	EC	0.61	TDS	Na	H.$CaCO_3$	1.32
$CaCO_3$	Cl	EC	0.63	HCO_3	$CaCO_3$	SO_4	1.33
H.$CaCO_3$	Na	Cl	0.66	Na	$CaCO_3$	SO_4	1.33
H.$CaCO_3$	Na	EC	0.67	Na	$CaCO_3$	H.$CaCO_3$	1.33
$CaCO_3$	H.$CaCO_3$	SO_4	0.68	Na	SO_4	H.$CaCO_3$	1.33
				$CaCO_3$	SO_4	H.$CaCO_3$	1.33

Table 45. Combinations of four variables to discontinue

MB05RAS078				I_a	**MB05RBS003**				I_a
$H.CaCO_3$	TDS	HCO_3	$CaCO_3$	1.11	TDS	Cl	HCO_3	$H.CaCO_3$	0.5
$H.CaCO_3$	TDS	HCO_3	EC	1.12	TDS	$CaCO_3$	HCO_3	$H.CaCO_3$	0.51
$H.CaCO_3$	TDS	$CaCO_3$	EC	1.12	Na	Cl	$H.CaCO_3$	EC	0.51
$H.CaCO_3$	TDS	HCO_3	SO_4	1.12	Cl	$CaCO_3$	HCO_3	$H.CaCO_3$	0.52
$H.CaCO_3$	TDS	$CaCO_3$	SO_4	1.12	SO_4	$CaCO_3$	HCO_3	EC	0.52
$H.CaCO_3$	HCO_3	$CaCO_3$	EC	1.12	Cl	$CaCO_3$	HCO_3	EC	0.52
$H.CaCO_3$	HCO_3	$CaCO_3$	SO_4	1.13	Cl	$CaCO_3$	$H.CaCO_3$	EC	0.52
TDS	HCO_3	$CaCO_3$	EC	1.13	Cl	HCO_3	$H.CaCO_3$	EC	0.52
TDS	HCO_3	$CaCO_3$	SO_4	1.13	$CaCO_3$	HCO_3	$H.CaCO_3$	EC	0.53
$H.CaCO_3$	TDS	EC	SO_4	1.13	**MB05SAS004**				I_a
$H.CaCO_3$	HCO_3	$CaCO_3$	Cl	1.13	Na	$H.CaCO_3$	EC	TSS	1.15
$H.CaCO_3$	HCO_3	EC	SO_4	1.14	HCO_3	$H.CaCO_3$	EC	TSS	1.15
$H.CaCO_3$	$CaCO_3$	EC	SO_4	1.14	Na	$H.CaCO_3$	EC	Turb.	1.15
TDS	HCO_3	EC	SO_4	1.14	HCO_3	$H.CaCO_3$	EC	Turb.	1.15
TDS	$CaCO_3$	EC	SO_4	1.14	Na	$H.CaCO_3$	EC	Cl	1.16
$H.CaCO_3$	HCO_3	EC	Cl	1.14	HCO_3	$H.CaCO_3$	EC	Cl	1.16
$H.CaCO_3$	$CaCO_3$	EC	Cl	1.14	Na	$H.CaCO_3$	EC	SO_4	1.16
$H.CaCO_3$	HCO_3	SO_4	Cl	1.15	HCO_3	$H.CaCO_3$	EC	SO_4	1.16
HCO_3	$CaCO_3$	EC	SO_4	1.15	$CaCO_3$	HCO_3	TSS	Cl	1.18
$H.CaCO_3$	$CaCO_3$	SO_4	Cl	1.15	Na	$H.CaCO_3$	TSS	Cl	1.18
HCO_3	$CaCO_3$	EC	Cl	1.15	HCO_3	$H.CaCO_3$	TSS	Cl	1.18
HCO_3	$CaCO_3$	SO_4	Cl	1.16	$CaCO_3$	HCO_3	Turb.	Cl	1.18
$H.CaCO_3$	EC	SO_4	Cl	1.16	$CaCO_3$	HCO_3	TSS	SO_4	1.18
HCO_3	EC	SO_4	Cl	1.17	Na	$H.CaCO_3$	Turb.	Cl	1.18
$CaCO_3$	EC	SO_4	Cl	1.17	Na	$H.CaCO_3$	TSS	SO_4	1.18
MB05RBS003				I_a	HCO_3	$H.CaCO_3$	Turb.	Cl	1.18
TDS	Cl	$CaCO_3$	HCO_3	0.5	HCO_3	$H.CaCO_3$	TSS	SO_4	1.18
TDS	SO_4	$CaCO_3$	HCO_3	0.5	$CaCO_3$	HCO_3	Turb.	SO_4	1.18
TDS	Cl	$CaCO_3$	$H.CaCO_3$	0.5	Na	$H.CaCO_3$	Turb.	SO_4	1.18

Table 45 (cont). Combinations of four variables to discontinue

MB05SAS004				I_a	**MB05SBS126**				I_a
HCO_3	$H.CaCO_3$	Turb.	SO_4	1.18	HCO_3	TDS	$CaCO_3$	Turb.	0.96
Na	EC	TSS	Cl	1.19	HCO_3	$CaCO_3$	EC	$H.CaCO_3$	0.96
HCO_3	EC	TSS	Cl	1.19	TDS	$CaCO_3$	EC	$H.CaCO_3$	0.97
$H.CaCO_3$	EC	TSS	Cl	1.19	HCO_3	TDS	SO_4	TSS	0.97
Na	EC	Turb.	Cl	1.19	HCO_3	SO_4	Na	$H.CaCO_3$	0.97
Na	EC	TSS	SO_4	1.19	$CaCO_3$	SO_4	EC	Na	0.97
HCO_3	EC	Turb.	Cl	1.19	HCO_3	$CaCO_3$	SO_4	TSS	0.98
HCO_3	EC	TSS	SO_4	1.2	HCO_3	TDS	SO_4	Turb.	0.98
$H.CaCO_3$	EC	Turb.	Cl	1.2	HCO_3	SO_4	EC	$H.CaCO_3$	0.98
$H.CaCO_3$	EC	TSS	SO_4	1.2	TDS	$CaCO_3$	SO_4	TSS	0.98
Na	EC	Turb.	SO_4	1.2	HCO_3	TDS	EC	TSS	0.98
HCO_3	EC	Turb.	SO_4	1.2	TDS	SO_4	EC	$H.CaCO_3$	0.98
$H.CaCO_3$	EC	Turb.	SO_4	1.2	HCO_3	$CaCO_3$	Na	TSS	0.98
MB05SBS126				I_a	$CaCO_3$	SO_4	Na	$H.CaCO_3$	0.99
HCO_3	TDS	$CaCO_3$	SO_4	0.9	HCO_3	EC	Na	$H.CaCO_3$	0.99
HCO_3	TDS	$CaCO_3$	EC	0.91	HCO_3	$CaCO_3$	SO_4	Turb.	0.99
HCO_3	TDS	$CaCO_3$	$H.CaCO_3$	0.93	HCO_3	$CaCO_3$	EC	TSS	0.99
HCO_3	TDS	SO_4	EC	0.93	TDS	$CaCO_3$	SO_4	Turb.	0.99
HCO_3	$CaCO_3$	SO_4	Na	0.93	$CaCO_3$	SO_4	EC	$H.CaCO_3$	0.99
HCO_3	$CaCO_3$	SO_4	EC	0.94	HCO_3	TDS	EC	Turb.	0.99
HCO_3	TDS	SO_4	$H.CaCO_3$	0.94	HCO_3	TDS	$H.CaCO_3$	TSS	0.99
TDS	$CaCO_3$	SO_4	EC	0.94	TDS	$CaCO_3$	EC	TSS	0.99
HCO_3	$CaCO_3$	EC	Na	0.95	HCO_3	$CaCO_3$	Na	Turb.	1
HCO_3	$CaCO_3$	SO_4	$H.CaCO_3$	0.95	$CaCO_3$	EC	Na	H.CaCO3	1
HCO_3	TDS	$CaCO_3$	TSS	0.95	HCO_3	SO_4	Na	TSS	1
TDS	$CaCO_3$	SO_4	$H.CaCO_3$	0.95	HCO_3	$CaCO_3$	EC	Turb.	1
HCO_3	TDS	EC	$H.CaCO_3$	0.95	HCO_3	$CaCO_3$	$H.CaCO_3$	TSS	1
HCO_3	$CaCO_3$	Na	$H.CaCO_3$	0.96	HCO_3	TDS	$H.CaCO_3$	Turb.	1
HCO_3	SO_4	EC	Na	0.96	TDS	$CaCO_3$	EC	Turb.	1

Table 45 (cont). Combinations of four variables to discontinue

MB05SBS126				I_a	MB05SBS126				I_a
TDS	$CaCO_3$	H.$CaCO_3$	TSS	1	SO_4	EC	Na	TSS	1.04
HCO_3	SO_4	EC	TSS	1.01	$CaCO_3$	EC	H.$CaCO_3$	TSS	1.04
TDS	SO_4	EC	TSS	1.01	TDS	EC	H.$CaCO_3$	Turb.	1.04
HCO_3	SO_4	Na	Turb.	1.01	$CaCO_3$	Na	H.$CaCO_3$	Turb.	1.05
HCO_3	$CaCO_3$	H.$CaCO_3$	Turb.	1.01	SO_4	EC	Na	Turb.	1.05
$CaCO_3$	SO_4	Na	TSS	1.01	SO_4	Na	H.$CaCO_3$	TSS	1.05
HCO_3	EC	Na	TSS	1.01	$CaCO_3$	EC	H.$CaCO_3$	Turb.	1.05
SO_4	EC	Na	H.$CaCO_3$	1.01	SO_4	EC	H.$CaCO_3$	TSS	1.06
TDS	$CaCO_3$	H.$CaCO_3$	Turb.	1.01	SO_4	Na	H.$CaCO_3$	Turb.	1.06
HCO_3	SO_4	EC	Turb.	1.02	EC	Na	H.$CaCO_3$	TSS	1.06
HCO_3	SO_4	H.$CaCO_3$	TSS	1.02	SO_4	EC	H.$CaCO_3$	Turb.	1.07
$CaCO_3$	SO_4	EC	TSS	1.02	EC	Na	H.$CaCO_3$	Turb.	1.08
TDS	SO_4	EC	Turb.	1.02	**MB05SCS005**				I_a
TDS	SO_4	H.$CaCO_3$	TSS	1.02	HCO_3	$CaCO_3$	H.$CaCO_3$	Na	0.14
$CaCO_3$	SO_4	Na	Turb.	1.02	$CaCO_3$	H.$CaCO_3$	Na	Cl	0.25
HCO_3	EC	Na	Turb.	1.02	$CaCO_3$	H.$CaCO_3$	Na	EC	0.26
HCO_3	Na	H.$CaCO_3$	TSS	1.02	$CaCO_3$	H.$CaCO_3$	Cl	EC	0.28
$CaCO_3$	EC	Na	TSS	1.02	$CaCO_3$	Na	Cl	EC	0.29
HCO_3	SO_4	H.$CaCO_3$	Turb.	1.03	$CaCO_3$	H.$CaCO_3$	Na	SO_4	0.33
$CaCO_3$	SO_4	EC	Turb.	1.03	H.$CaCO_3$	Na	Cl	EC	0.34
$CaCO_3$	SO_4	H.$CaCO_3$	TSS	1.03	$CaCO_3$	H.$CaCO_3$	Cl	SO_4	0.35
HCO_3	EC	H.$CaCO_3$	TSS	1.03	$CaCO_3$	Na	Cl	SO_4	0.36
TDS	SO_4	H.$CaCO_3$	Turb.	1.03	$CaCO_3$	H.$CaCO_3$	EC	SO_4	0.37
TDS	EC	H.$CaCO_3$	TSS	1.03	TDS	Cl	EC	SO_4	0.37
HCO_3	Na	H.$CaCO_3$	Turb.	1.03	$CaCO_3$	Na	EC	SO_4	0.37
$CaCO_3$	EC	Na	Turb.	1.04	$CaCO_3$	Cl	EC	SO_4	0.39
$CaCO_3$	Na	H.$CaCO_3$	TSS	1.04	H.$CaCO_3$	Na	Cl	SO_4	0.42
$CaCO_3$	SO_4	H.$CaCO_3$	Turb.	1.04	H.$CaCO_3$	Na	EC	SO_4	0.43
HCO_3	EC	H.$CaCO_3$	Turb.	1.04	H.$CaCO_3$	Cl	EC	SO_4	0.45
					Na	Cl	EC	SO_4	0.45

Table 46. Combinations of five variables to discontinue

MB05SAS004					I_a
Na	H.$CaCO_3$	EC	TSS	Cl	1.01
HCO_3	H.$CaCO_3$	EC	TSS	Cl	1.01
Na	H.$CaCO_3$	EC	Turb.	Cl	1.01
Na	H.$CaCO_3$	EC	TSS	SO_4	1.01
HCO_3	H.$CaCO_3$	EC	Turb.	Cl	1.01
HCO_3	H.$CaCO_3$	EC	TSS	SO_4	1.01
Na	H.$CaCO_3$	EC	Turb.	SO_4	1.01
HCO_3	H.$CaCO_3$	EC	Turb.	SO_4	1.01
MB05SBS126					I_a
HCO_3	TDS	$CaCO_3$	SO_4	EC	0.71
HCO_3	TDS	$CaCO_3$	SO_4	H.$CaCO_3$	0.73
HCO_3	TDS	$CaCO_3$	EC	H.$CaCO_3$	0.74
HCO_3	$CaCO_3$	SO_4	EC	Na	0.75
HCO_3	TDS	$CaCO_3$	SO_4	TSS	0.75
HCO_3	TDS	SO_4	EC	H.$CaCO_3$	0.75
HCO_3	$CaCO_3$	SO_4	Na	H.$CaCO_3$	0.76
HCO_3	TDS	$CaCO_3$	SO_4	Turb.	0.76
HCO_3	$CaCO_3$	SO_4	EC	H.$CaCO_3$	0.76
HCO_3	TDS	$CaCO_3$	EC	TSS	0.76
TDS	$CaCO_3$	SO_4	EC	H.$CaCO_3$	0.77
HCO_3	$CaCO_3$	EC	Na	H.$CaCO_3$	0.77
HCO_3	TDS	$CaCO_3$	EC	Turb.	0.78
HCO_3	TDS	$CaCO_3$	H.$CaCO_3$	TSS	0.78
HCO_3	TDS	SO_4	EC	TSS	0.78

Table 46 (cont). Combinations of five variables to discontinue

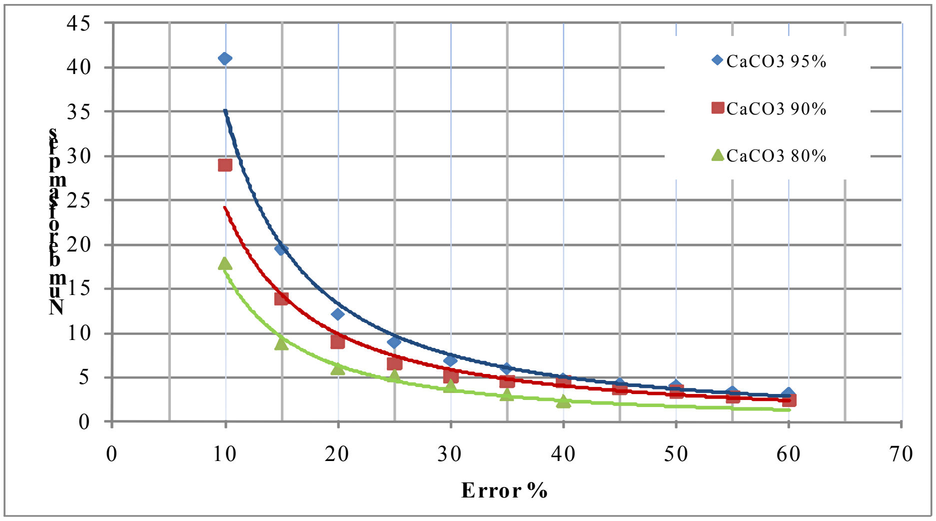
MB05SBS126					I_a
HCO_3	$CaCO_3$	SO_4	Na	TSS	0.78
HCO_3	SO_4	EC	Na	H.$CaCO_3$	0.79
HCO_3	TDS	$CaCO_3$	H.$CaCO_3$	Turb.	0.79
HCO_3	$CaCO_3$	SO_4	EC	TSS	0.79
HCO_3	TDS	SO_4	EC	Turb.	0.79
HCO_3	TDS	SO_4	H.$CaCO_3$	TSS	0.79
TDS	$CaCO_3$	SO_4	EC	TSS	0.79
HCO_3	$CaCO_3$	SO_4	Na	Turb.	0.79
HCO_3	$CaCO_3$	EC	Na	TSS	0.8
$CaCO_3$	SO_4	EC	Na	H.$CaCO_3$	0.8
HCO_3	$CaCO_3$	SO_4	EC	Turb.	0.8
HCO_3	$CaCO_3$	SO_4	H.$CaCO_3$	TSS	0.8
HCO_3	TDS	SO_4	H.$CaCO_3$	Turb.	0.8
TDS	$CaCO_3$	SO_4	EC	Turb.	0.8
TDS	$CaCO_3$	SO_4	H.$CaCO_3$	TSS	0.8
HCO_3	TDS	EC	H.$CaCO_3$	TSS	0.8
HCO_3	$CaCO_3$	EC	Na	Turb.	0.81
HCO_3	$CaCO_3$	Na	H.$CaCO_3$	TSS	0.81
HCO_3	$CaCO_3$	SO_4	H.$CaCO_3$	Turb.	0.81
HCO_3	SO_4	EC	Na	TSS	0.81
HCO_3	$CaCO_3$	EC	H.$CaCO_3$	TSS	0.81
TDS	$CaCO_3$	SO_4	H.$CaCO_3$	Turb.	0.81
HCO_3	TDS	EC	H.$CaCO_3$	Turb.	0.82
TDS	$CaCO_3$	EC	H.$CaCO_3$	TSS	0.82

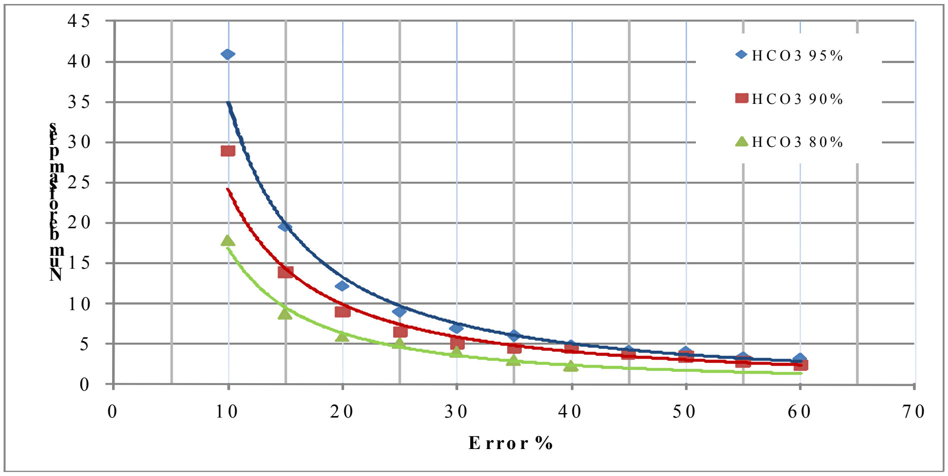
Table 46 (cont). Combinations of five variables to discontinue

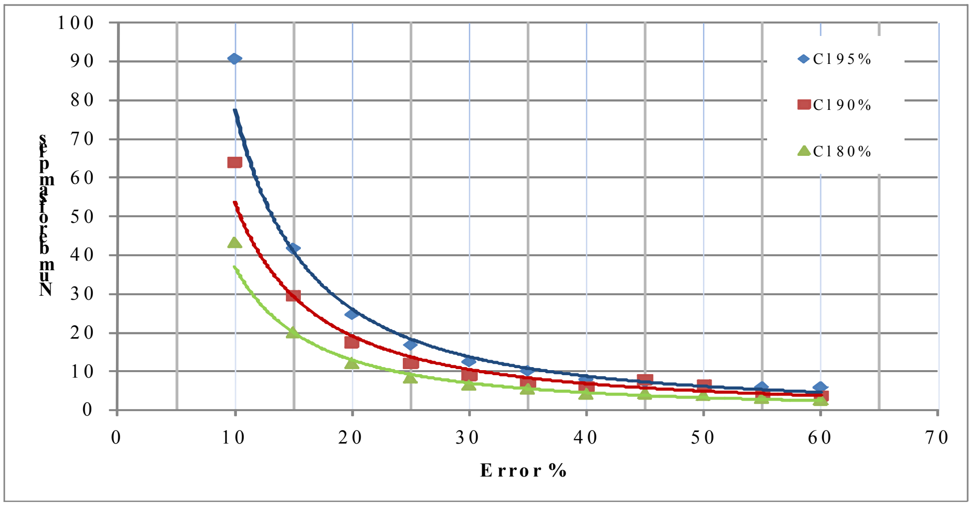
MB05SBS126					I_a
HCO_3	$CaCO_3$	Na	$H.CaCO_3$	Turb.	0.82
HCO_3	SO_4	EC	Na	Turb.	0.82
HCO_3	SO_4	Na	$H.CaCO_3$	TSS	0.82
HCO_3	$CaCO_3$	EC	$H.CaCO_3$	Turb.	0.82
$CaCO_3$	SO_4	EC	Na	TSS	0.82
TDS	$CaCO_3$	EC	$H.CaCO_3$	Turb.	0.83
HCO_3	SO_4	EC	$H.CaCO_3$	TSS	0.83
TDS	SO_4	EC	$H.CaCO_3$	TSS	0.83
HCO_3	SO_4	Na	$H.CaCO_3$	Turb.	0.83
$CaCO_3$	SO_4	EC	Na	Turb.	0.84
$CaCO_3$	SO_4	Na	$H.CaCO_3$	TSS	0.84
HCO_3	EC	Na	$H.CaCO_3$	TSS	0.84
HCO_3	SO_4	EC	$H.CaCO_3$	Turb.	0.84
$CaCO_3$	SO_4	EC	$H.CaCO_3$	TSS	0.84
TDS	SO_4	EC	$H.CaCO_3$	Turb.	0.84
$CaCO_3$	SO_4	Na	$H.CaCO_3$	Turb.	0.85
HCO_3	EC	Na	$H.CaCO_3$	Turb.	0.85
$CaCO_3$	EC	Na	$H.CaCO_3$	TSS	0.85
$CaCO_3$	SO_4	EC	$H.CaCO_3$	Turb.	0.85
$CaCO_3$	EC	Na	$H.CaCO_3$	Turb.	0.86
SO_4	EC	Na	$H.CaCO_3$	TSS	0.86
SO_4	EC	Na	$H.CaCO_3$	Turb.	0.87

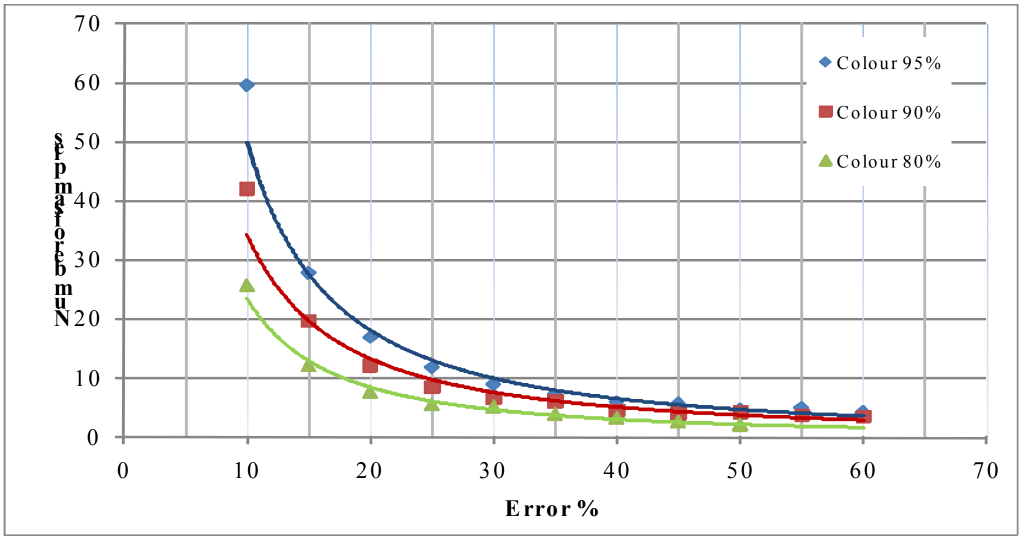
Appendix 6: Detailed results for the assessment of Sampling Frequency Confidence Interval approach

MB05RAS078

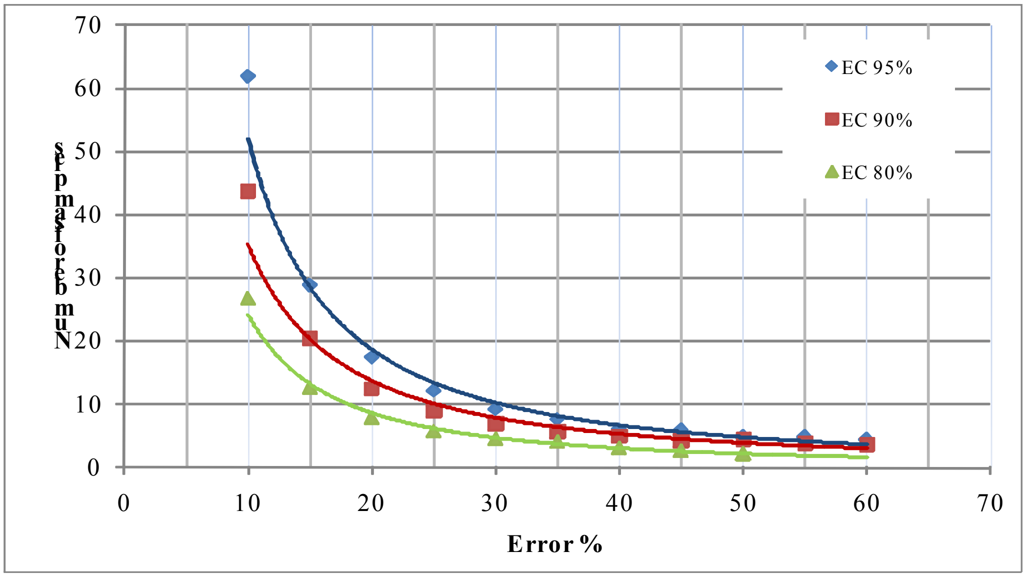




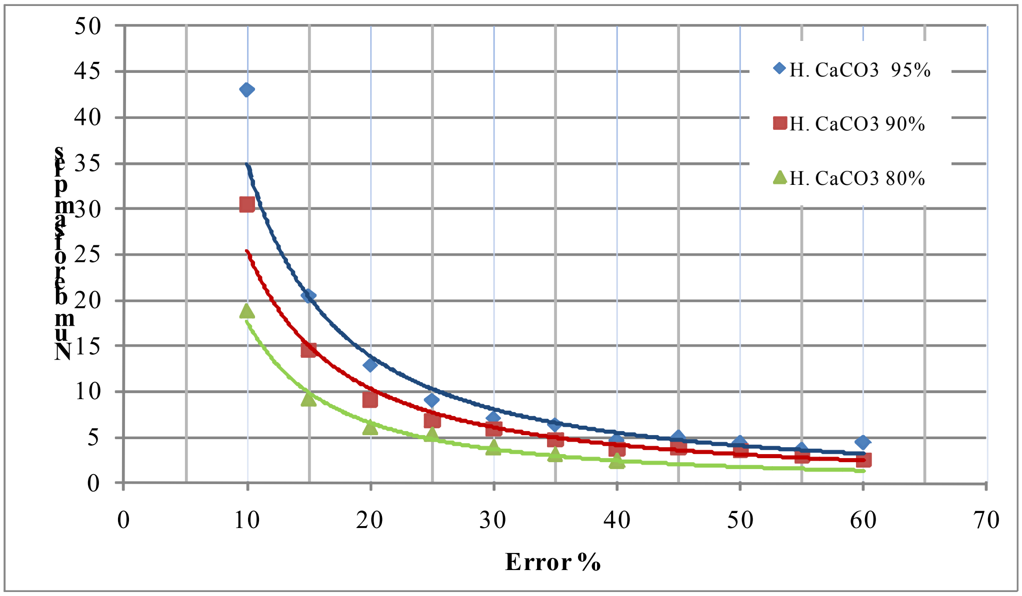




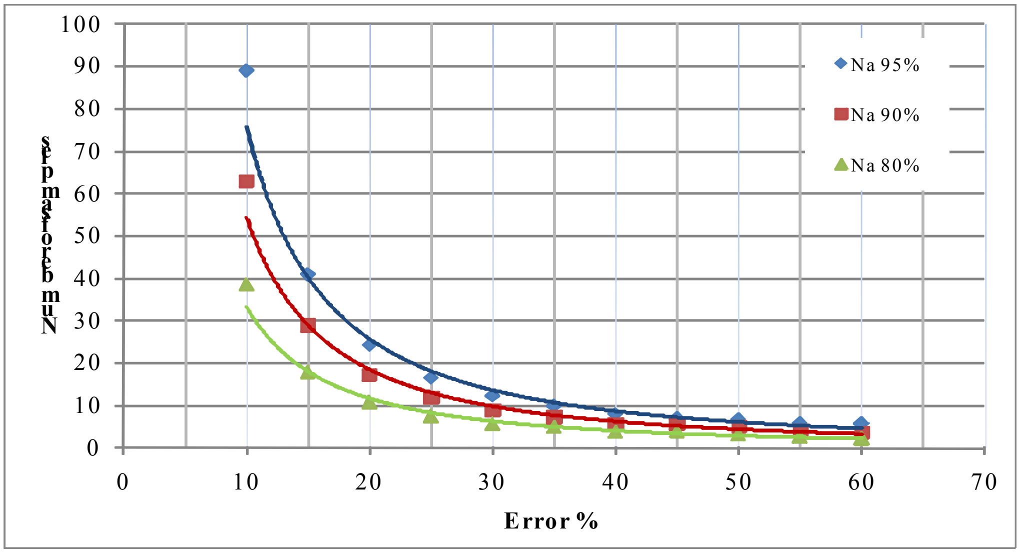
70
60
50
40
30
20
10
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ Colour 95%
■ Colour 90%
▲ Colour 80%

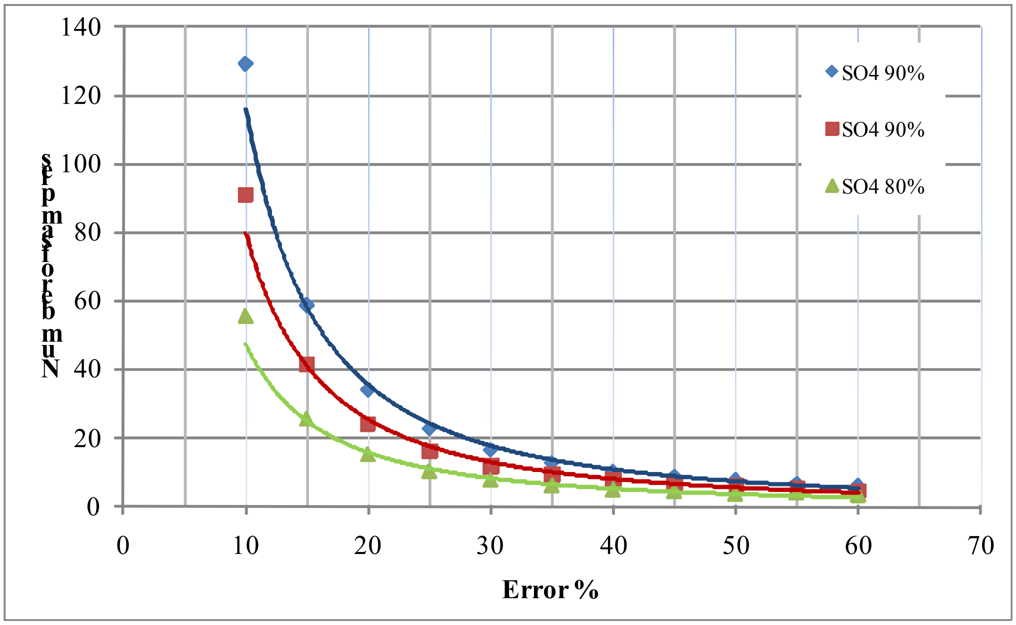


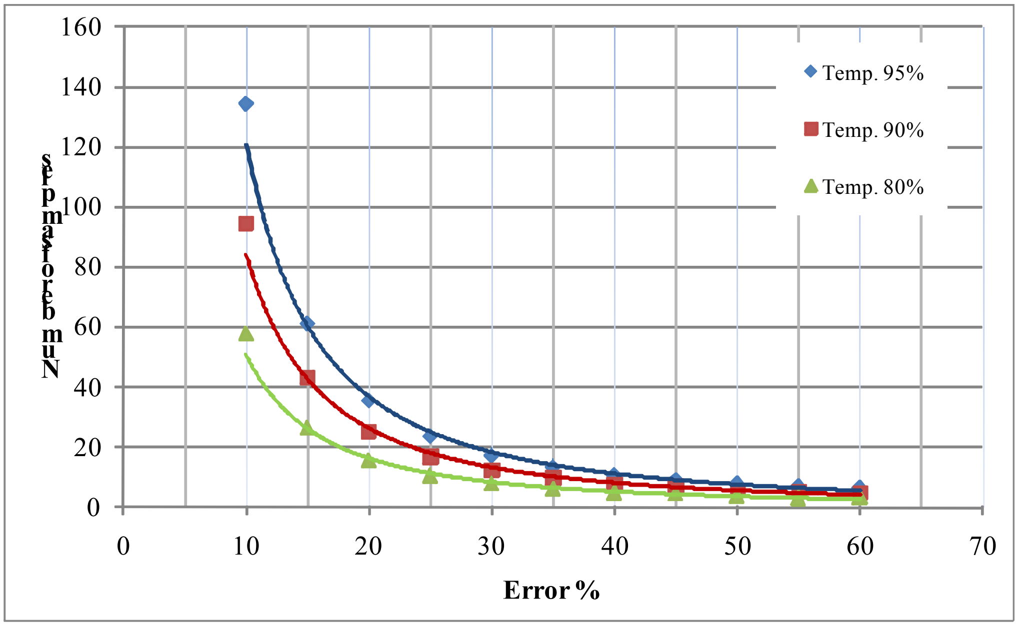
70
60
50
40
30
20
10
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ EC 95%
■ EC 90%
▲ EC 80%

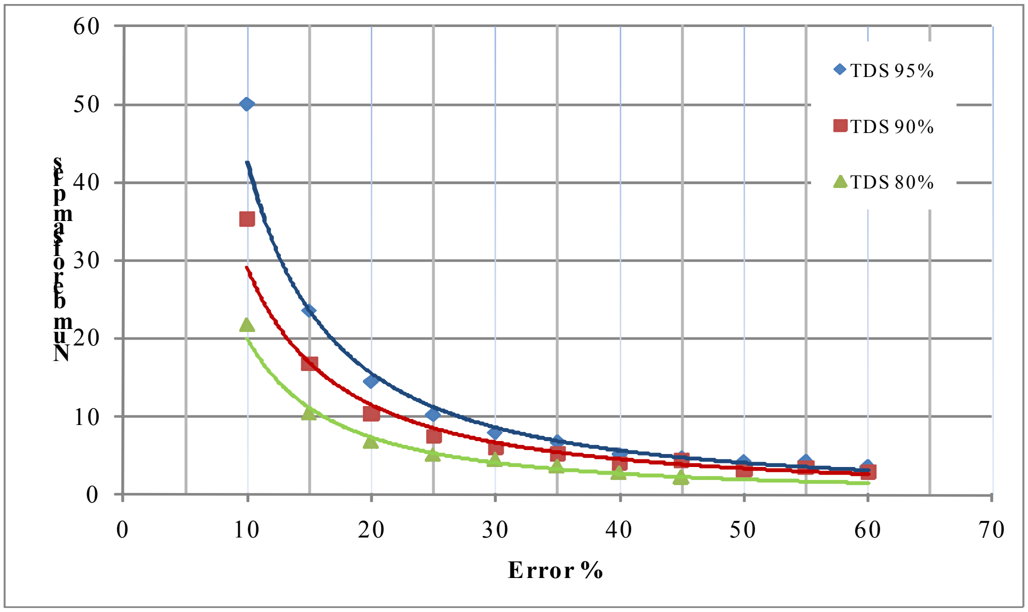


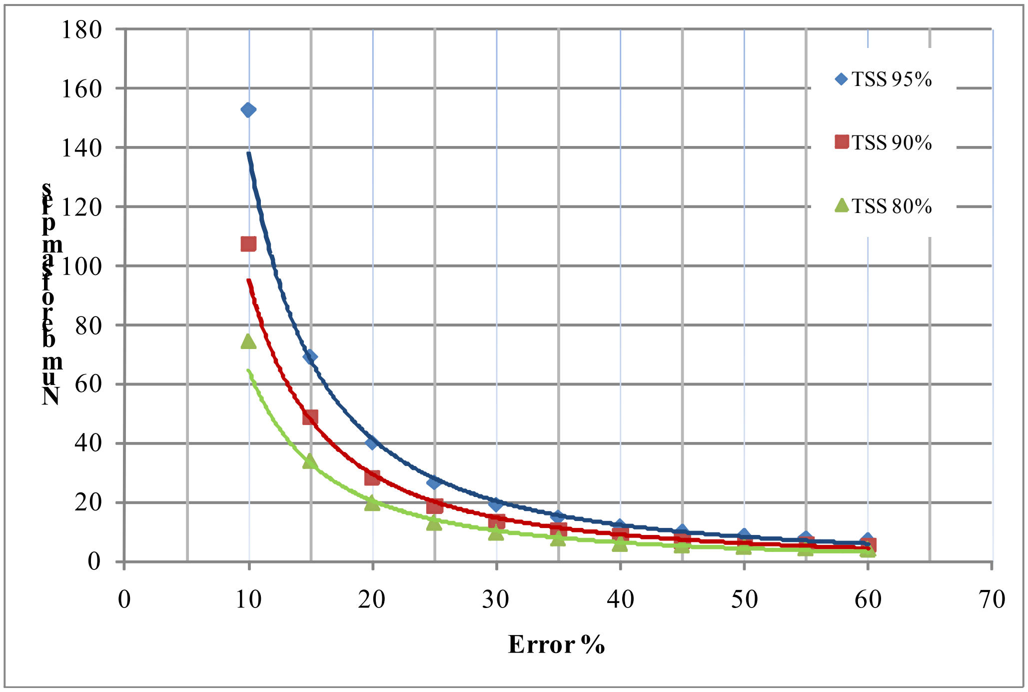
50
45
40
35
30
25
20
15
10
5
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ H. CaCO3 95%
■ H. CaCO3 90%
▲ H. CaCO3 80%

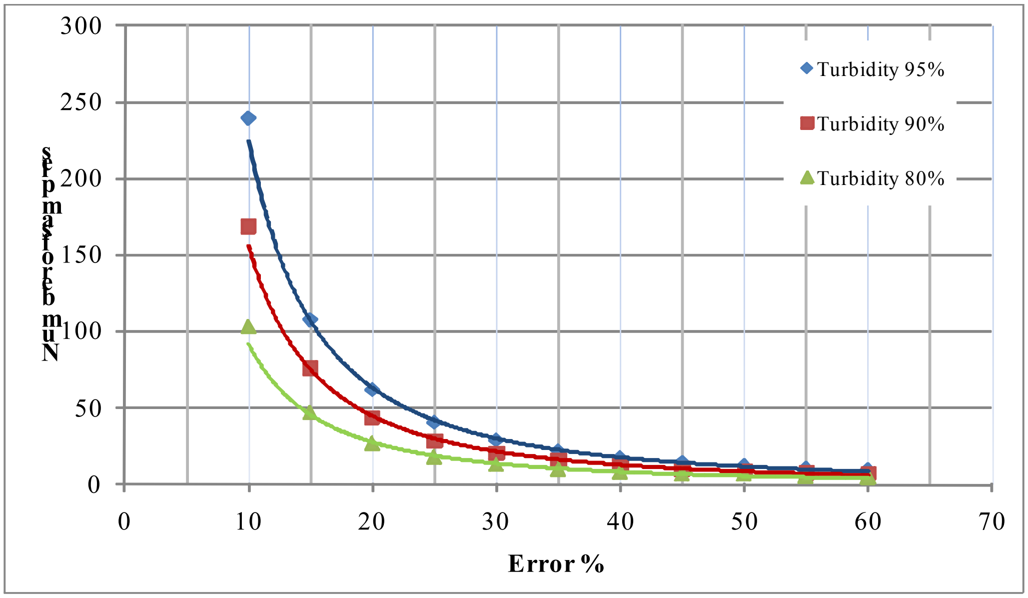

100
90
80
70
60
50
40
30
20
10
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Na 95%
Na 90%
Na 80%


140
120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
SO4 90%
SO4 90%
SO4 80%

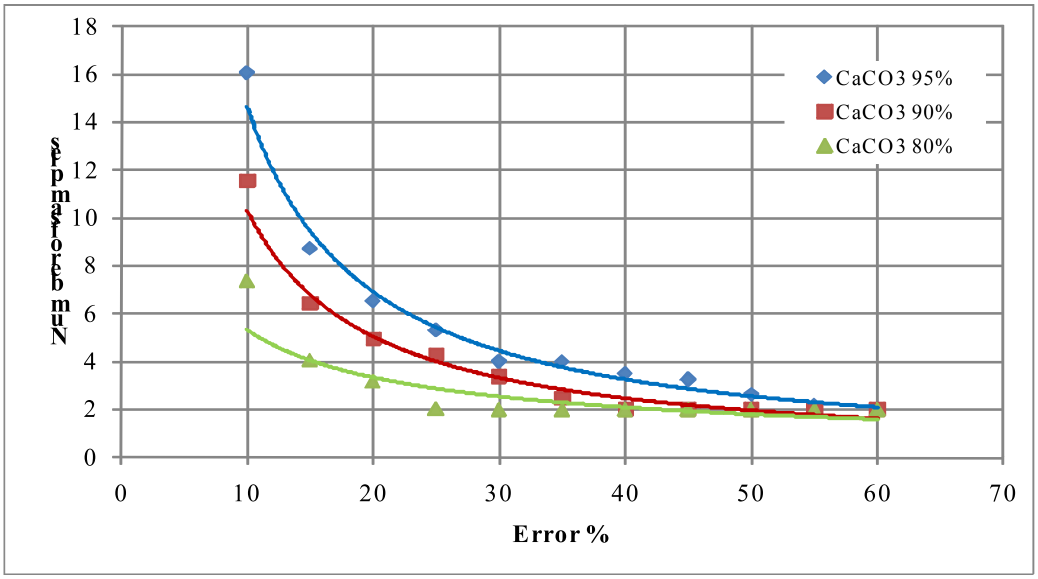

160
140
120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Temp. 95%
Temp. 90%
Temp. 80%


0
10
20
30
40
50
60
0
10
20
30
40
50
60
70
Number of samples
Error %
TDS 95%
TDS 90%
TDS 80%

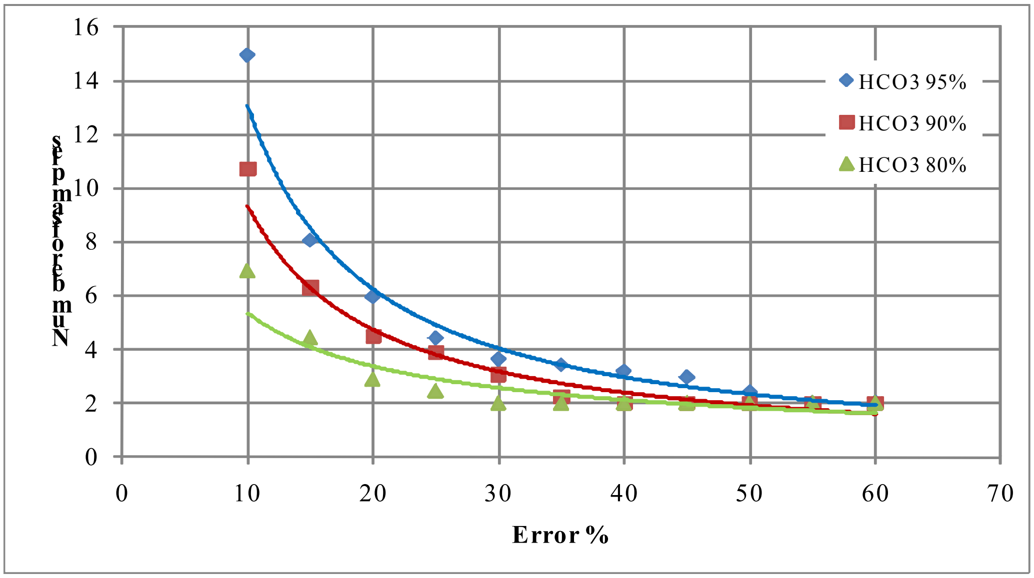

0
20
40
60
80
100
120
140
160
180
0
10
20
30
40
50
60
70
Number of samples
Error %
TSS 95%
TSS 90%
TSS 80%


0
50
100
150
200
250
300
0
10
20
30
40
50
60
70
Number of samples
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%

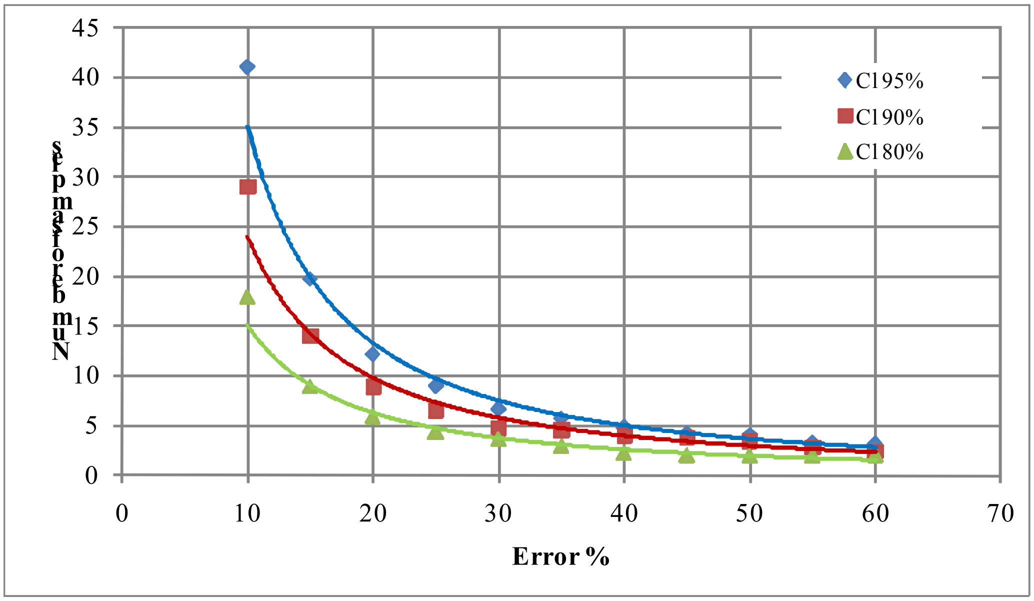
MB05RBS003



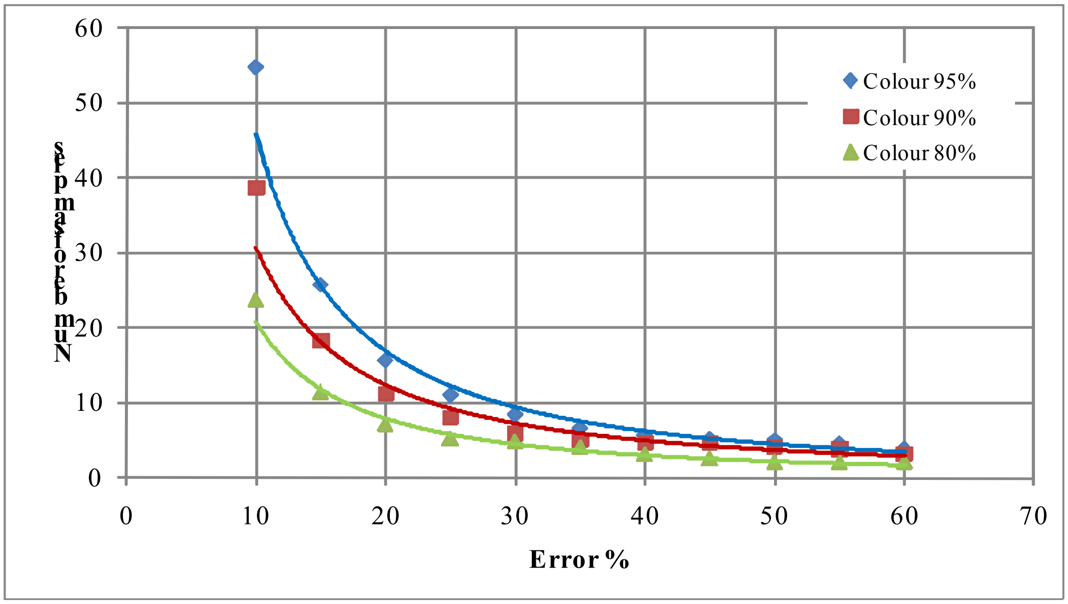
Number of samples
0
2
4
6
8
10
12
14
16
18
0
10
20
30
40
50
60
70
Error %
CaCO3 95%
CaCO3 90%
CaCO3 80%

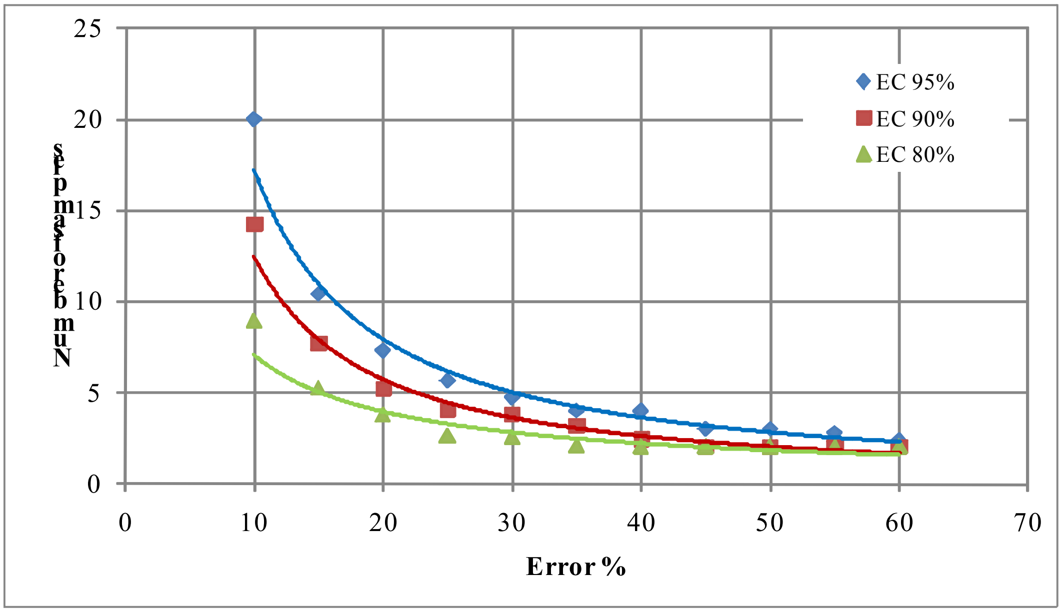


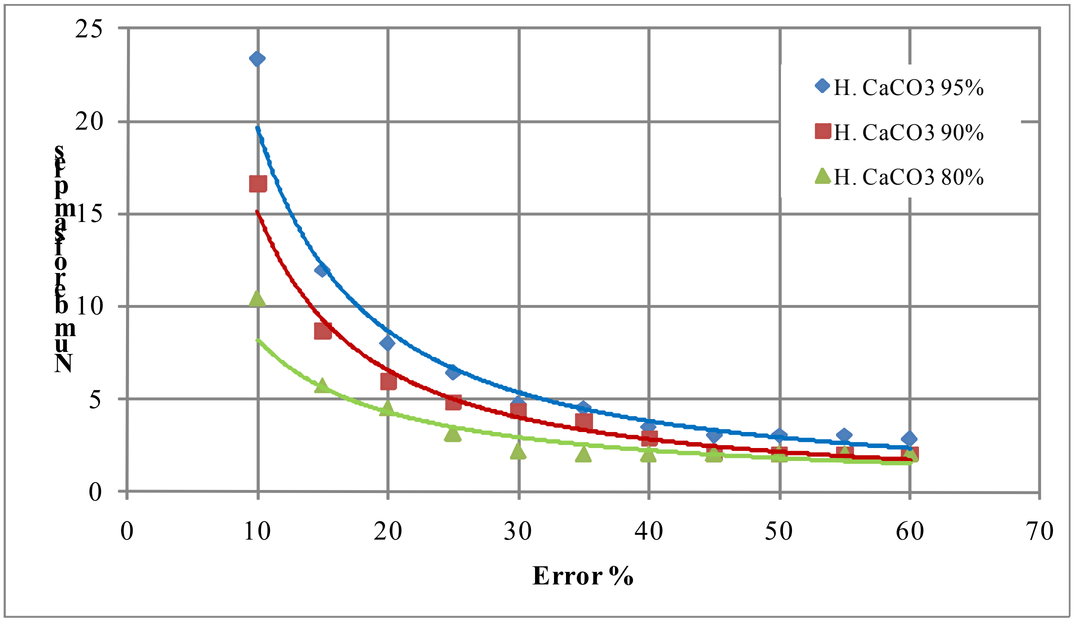
Number of samples
0
2
4
6
8
10
12
14
16
0
10
20
30
40
50
60
70
Error %
HCO3 95%
HCO3 90%
HCO3 80%

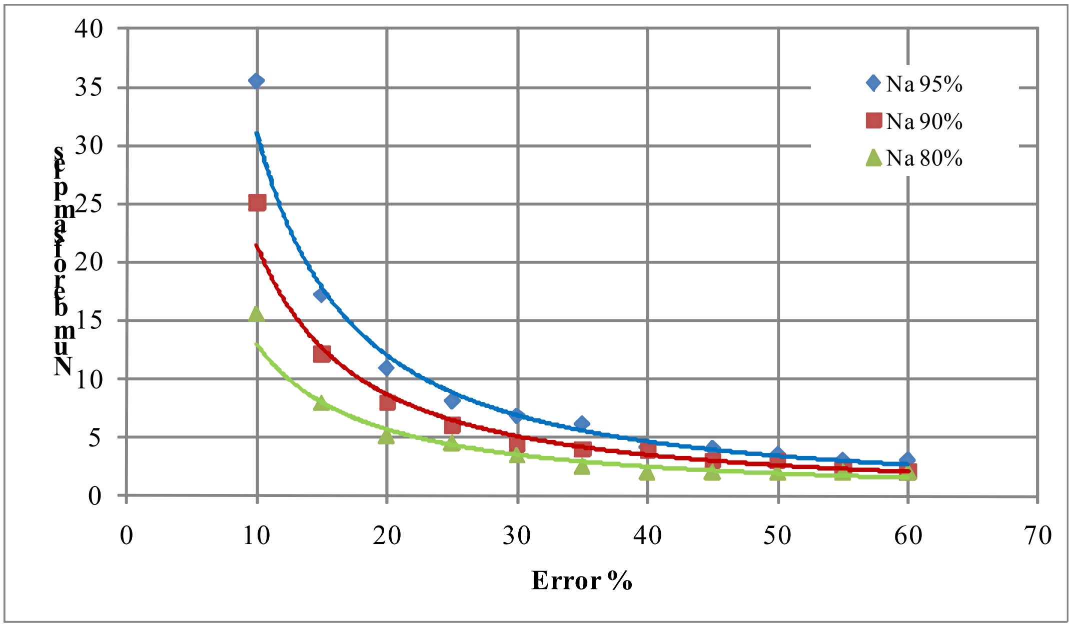


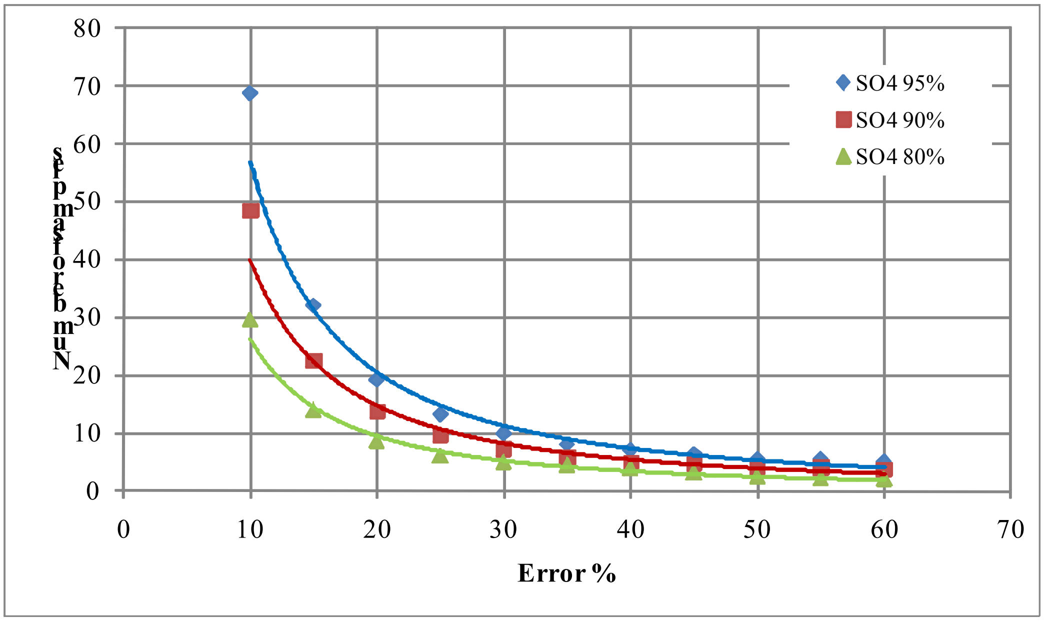
Number of samples
0
5
10
15
20
25
30
35
40
45
0
10
20
30
40
50
60
70
Error %
Cl 95%
Cl 90%
Cl 80%

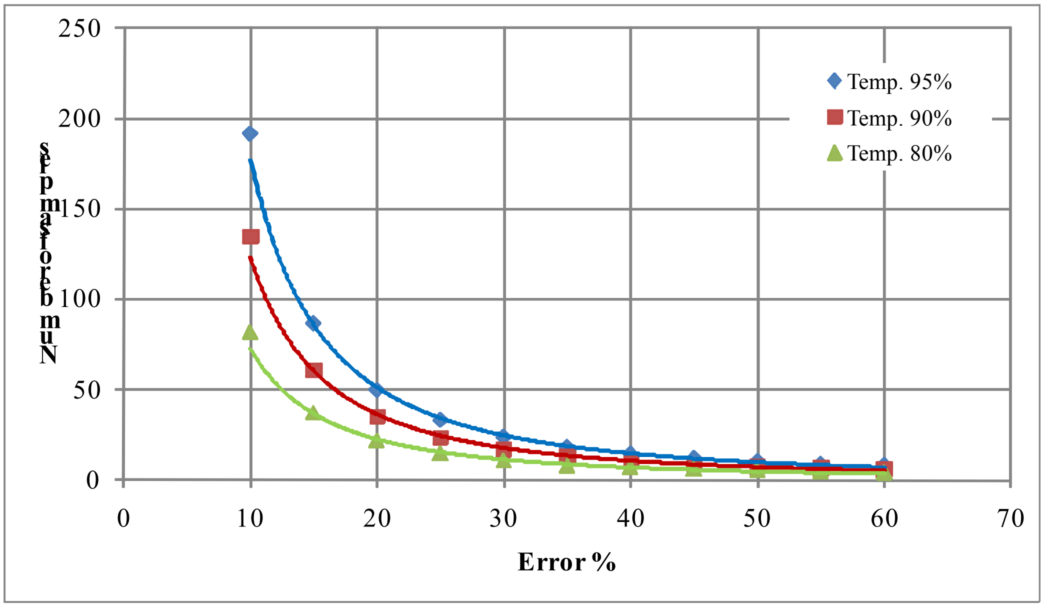

60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
Number of samples
Error %
◆ Colour 95%
■ Colour 90%
▲ Colour 80%

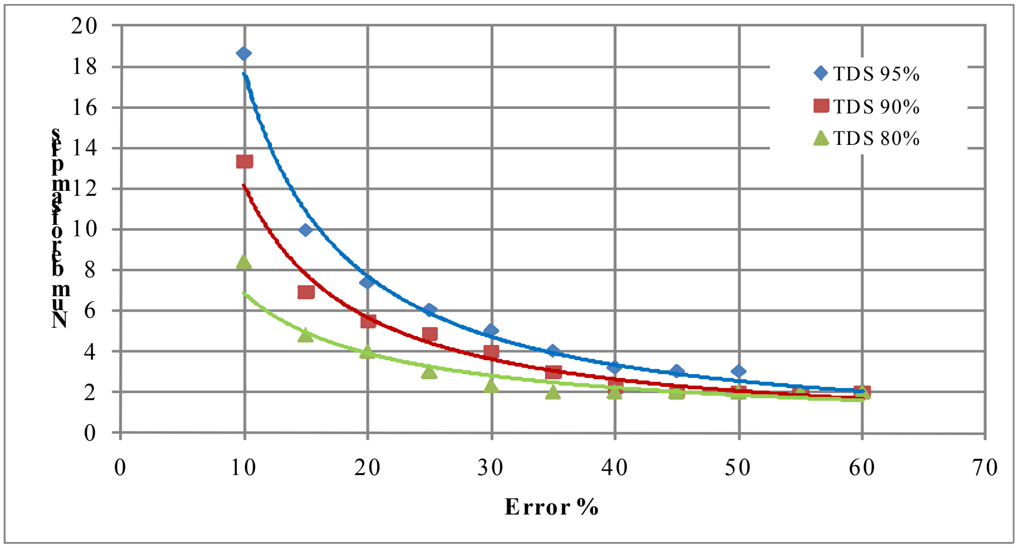

25
20
15
10
5
0
0
10
20
30
40
50
60
70
Number of samples
Error %
◆ EC 95%
■ EC 90%
▲ EC 80%

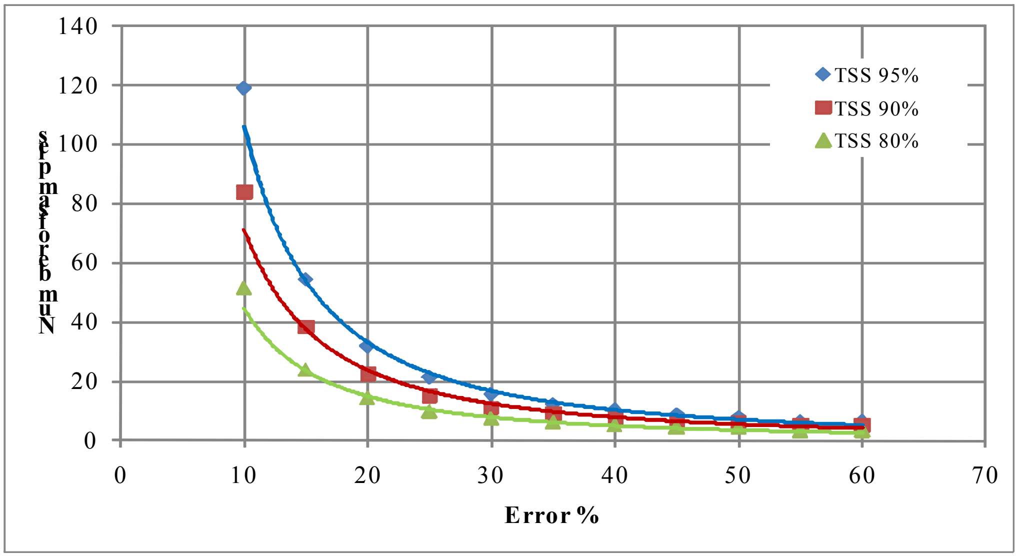

25
20
15
10
5
0
0
10
20
30
40
50
60
70
Number of samples
Error %
◆ H. CaCO3 95%
■ H. CaCO3 90%
▲ H. CaCO3 80%

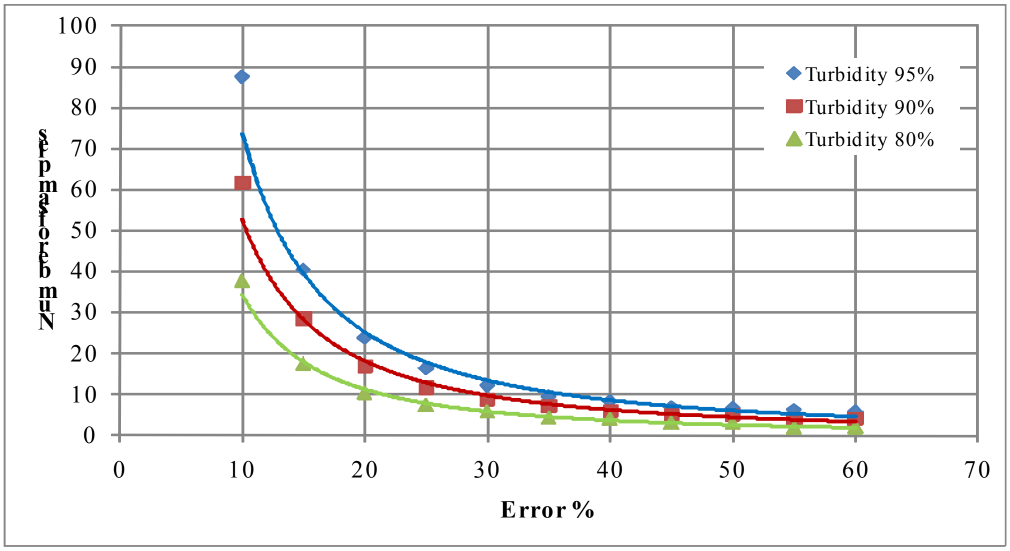

Number of samples
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
Error %
Na 95%
Na 90%
Na 80%

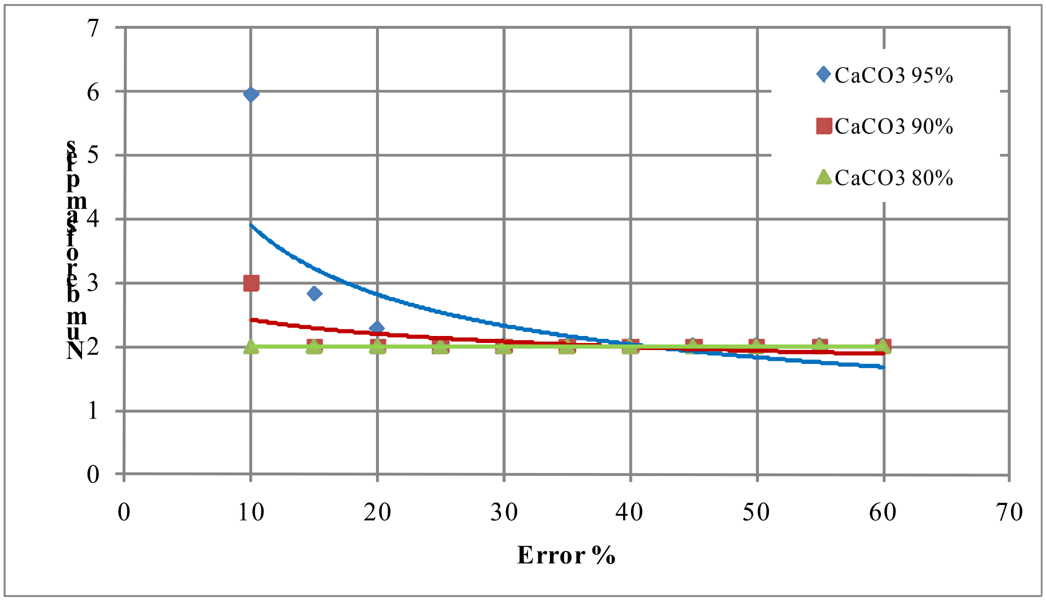

Number of samples
0
10
20
30
40
50
60
70
80
0
10
20
30
40
50
60
70
Error %
SO4 95%
SO4 90%
SO4 80%

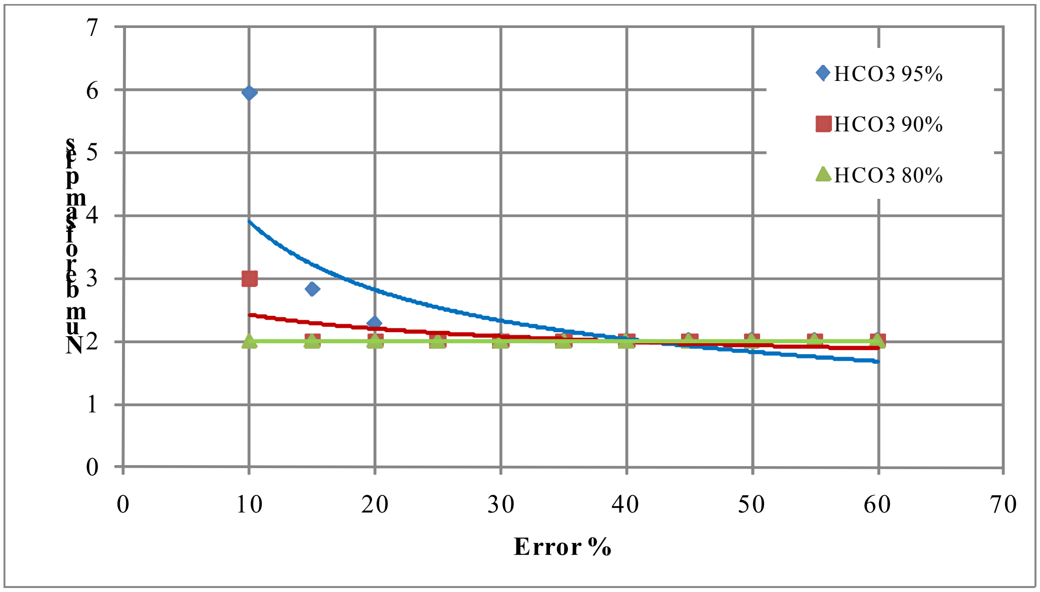

Number of samples
0
50
100
150
200
250
0
10
20
30
40
50
60
70
Error %
Temp. 95%
Temp. 90%
Temp. 80%

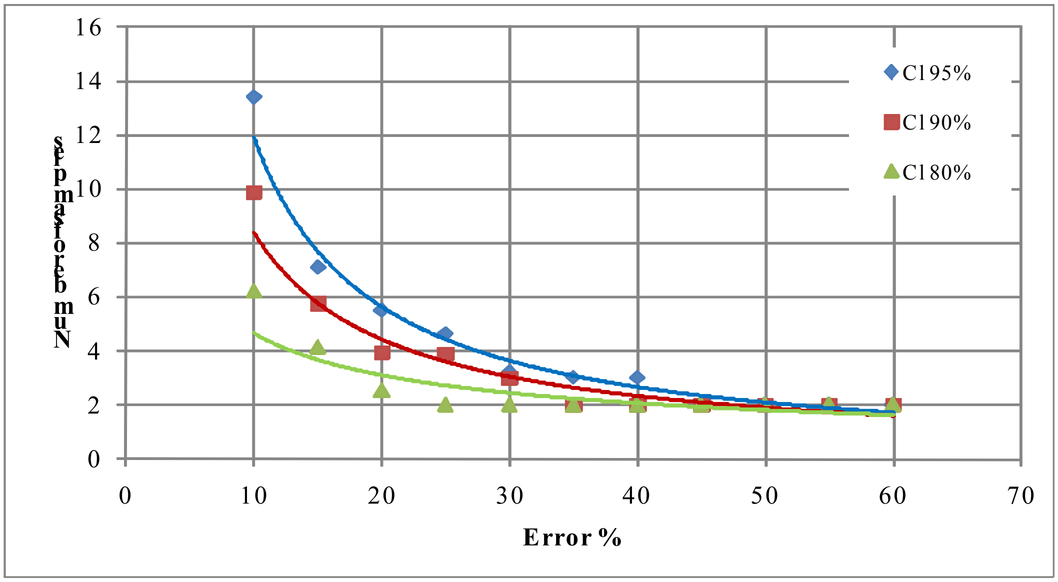

Number of samples
0
2
4
6
8
10
12
14
16
18
20
0
10
20
30
40
50
60
70
Error %
TDS 95%
TDS 90%
TDS 80%

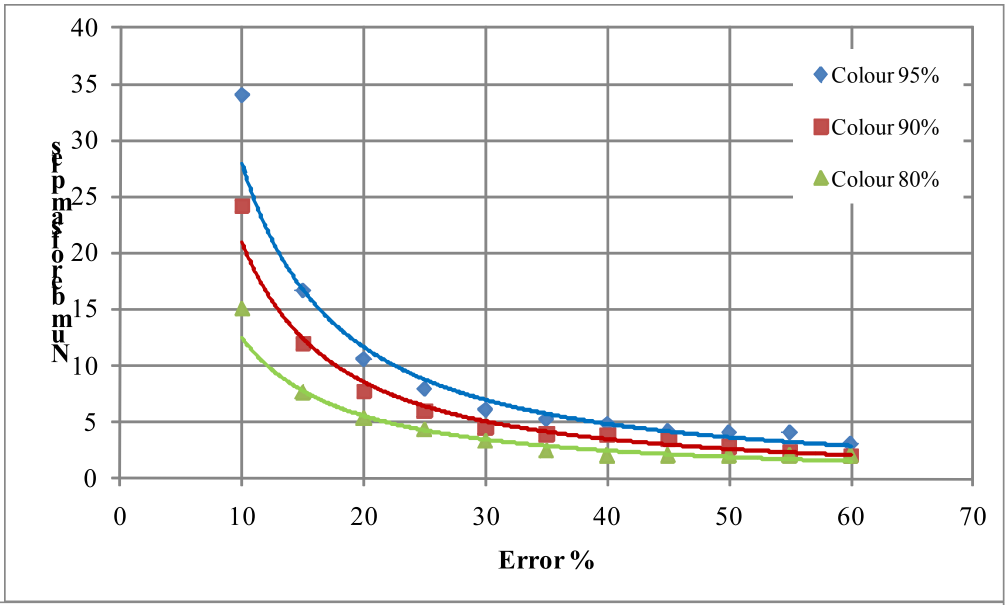

Number of samples
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
Error %
TSS 95%
TSS 90%
TSS 80%

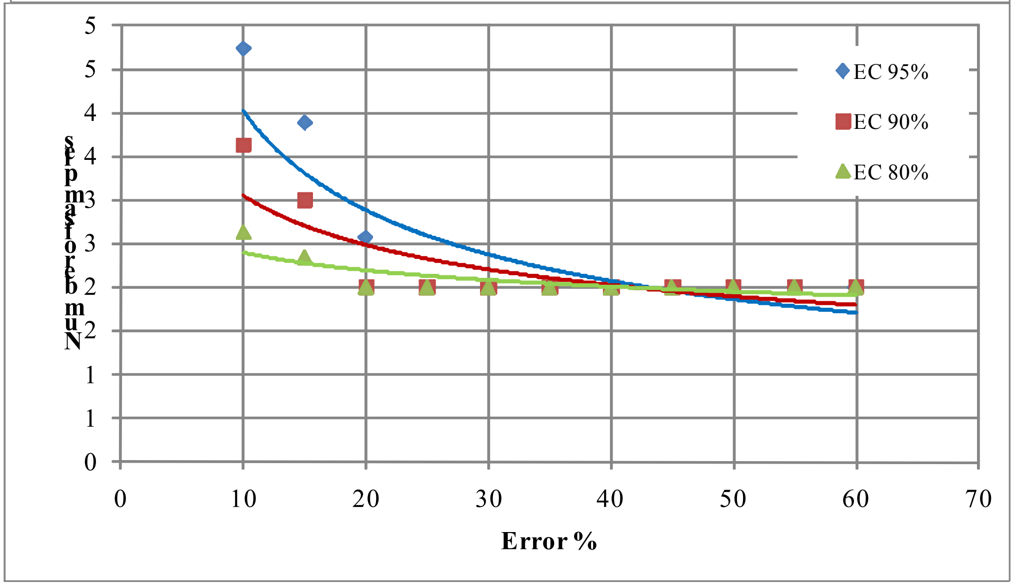

Number of samples
0
10
20
30
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%

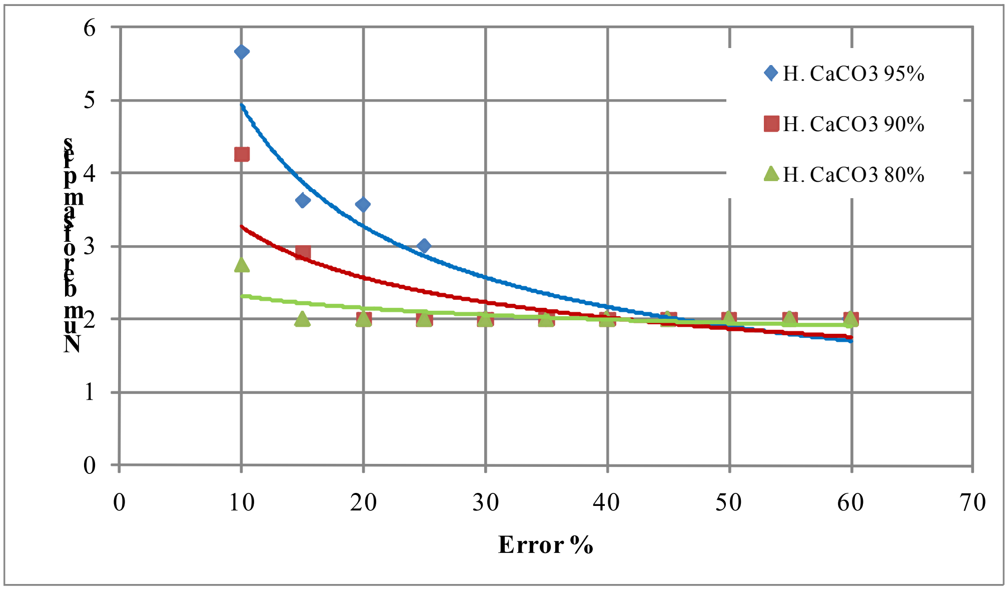

7
6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
CaCO3 95%
CaCO3 90%
CaCO3 80%

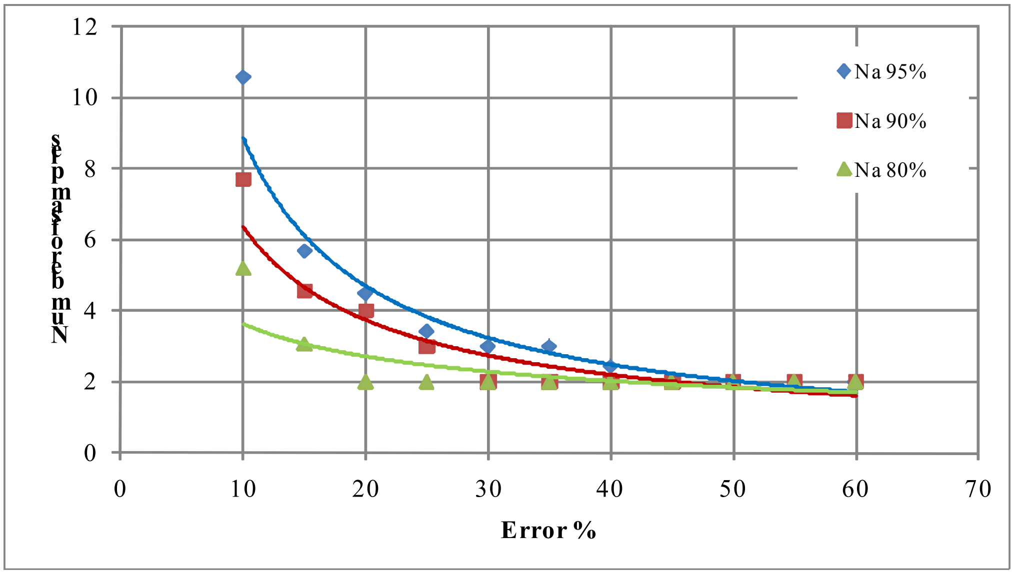

7
6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
HCO3 95%
HCO3 90%
HCO3 80%

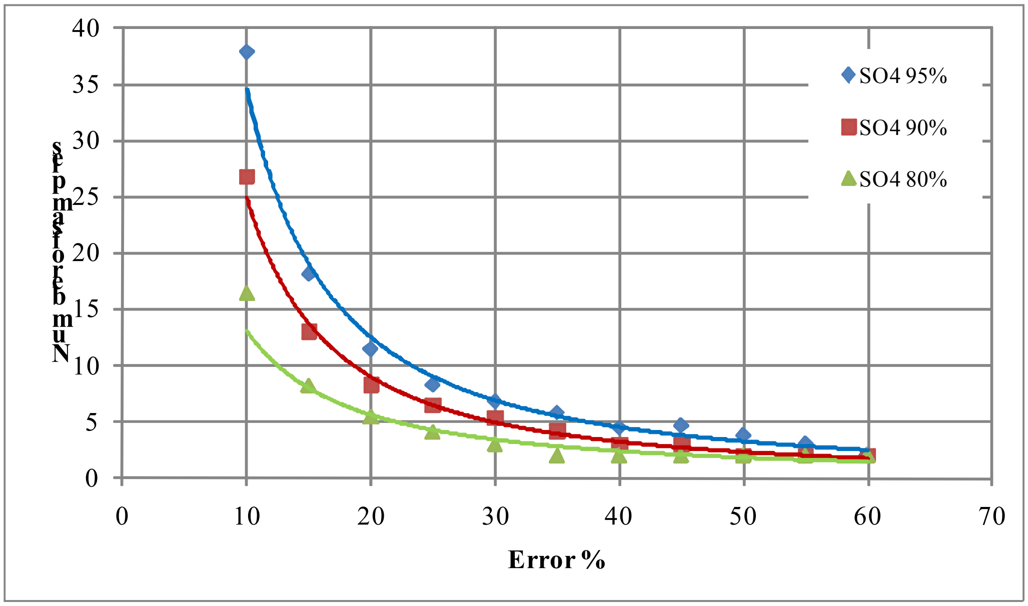

16
14
12
10
8
6
4
2
0
Number of samples
0
10
20
30
40
50
60
70
Error %
C195%
C190%
C180%

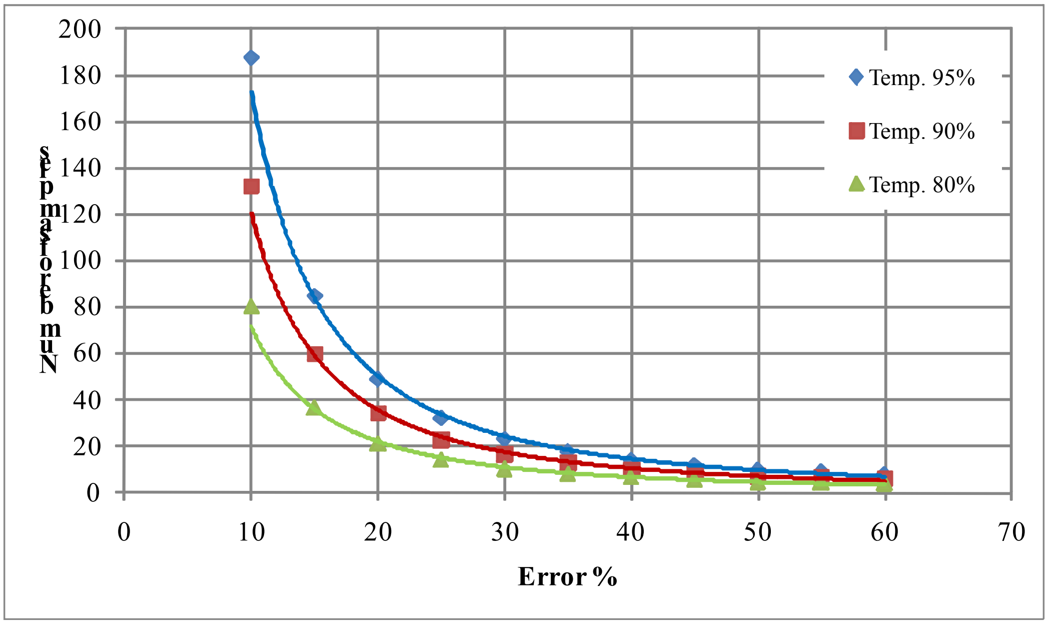

Number of samples
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %
Colour 95%
Colour 90%
Colour 80%

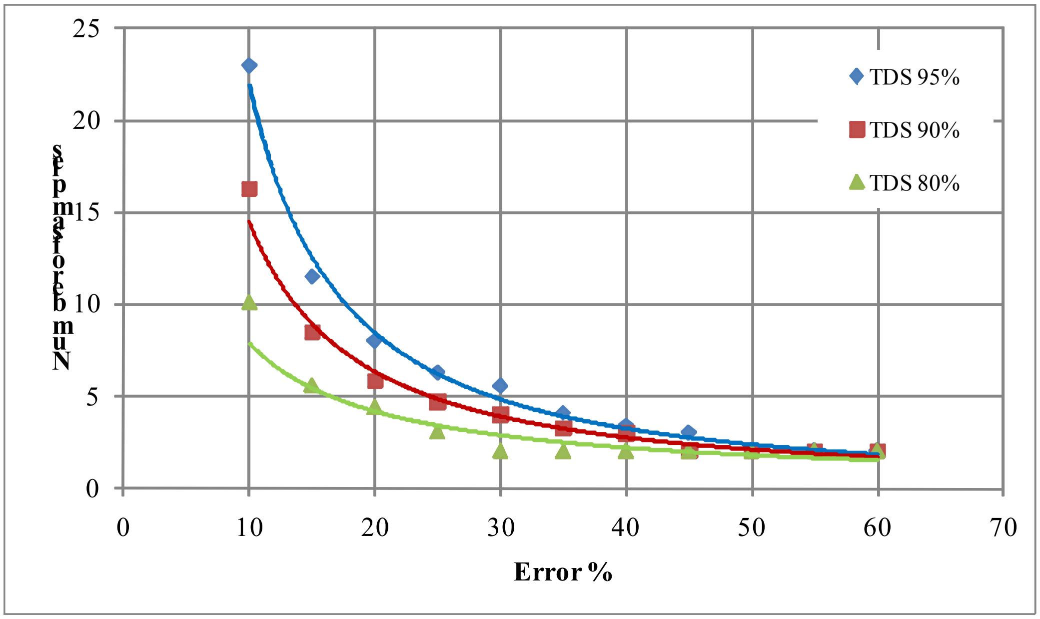

Number of samples
5
5
4
4
3
3
2
2
1
1
0
0
10
20
30
40
50
60
70
Error %
EC 95%
EC 90%
EC 80%


Number of samples
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
Error %
H. CaCO3 95%
H. CaCO3 90%
H. CaCO3 80%

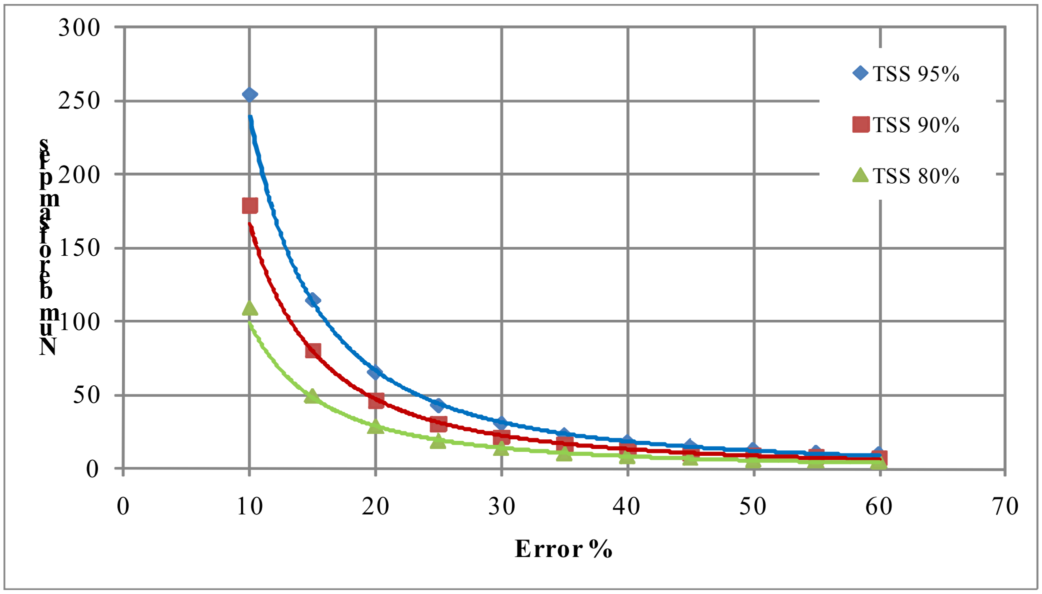

12
10
8
6
4
2
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Na 95%
Na 90%
Na 80%


40
35
30
25
20
15
10
5
0
Number of samples
0
10
20
30
40
50
60
70
Error %
SO4 95%
SO4 90%
SO4 80%

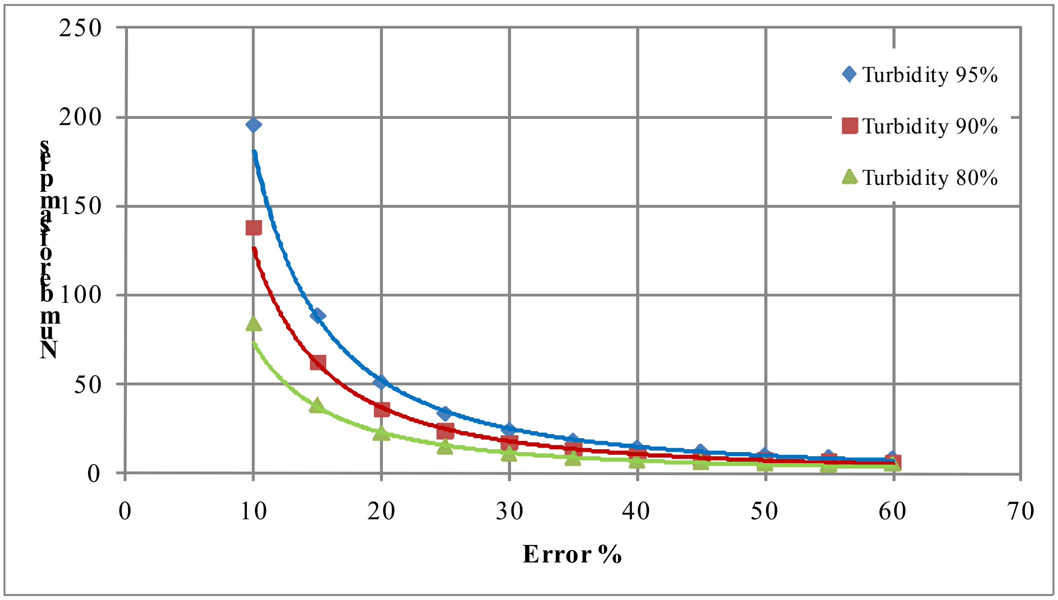

200
180
160
140
120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Temp. 95%
Temp. 90%
Temp. 80%



Number of samples
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %
TDS 95%
TDS 90%
TDS 80%

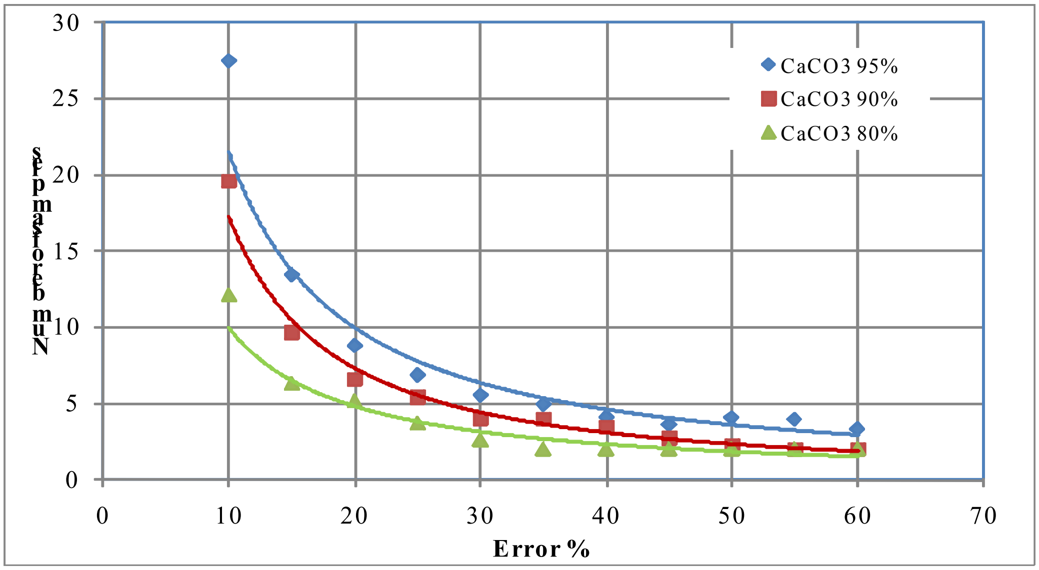


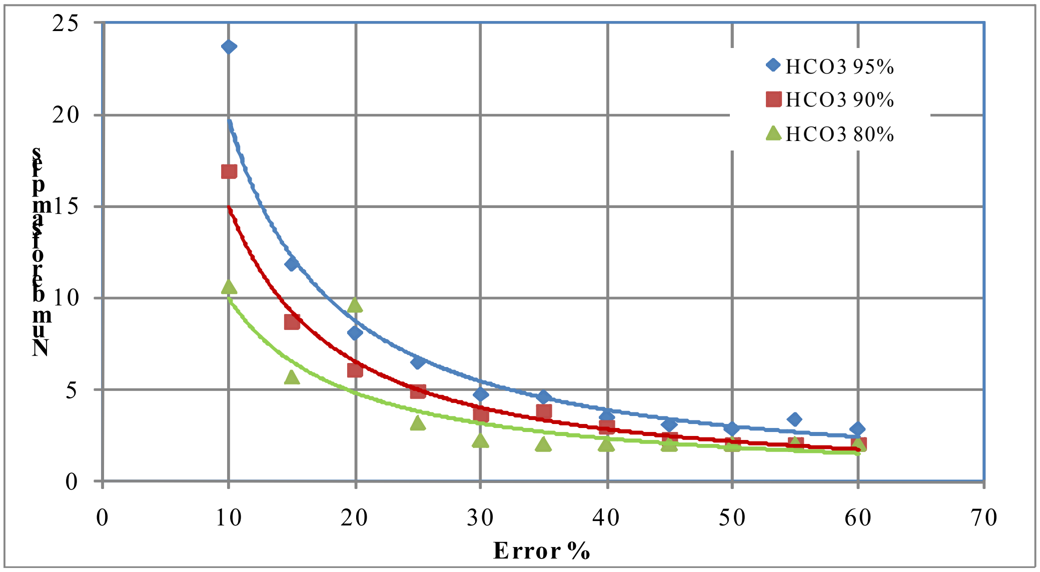
Number of samples
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Error %
TSS 95%
TSS 90%
TSS 80%

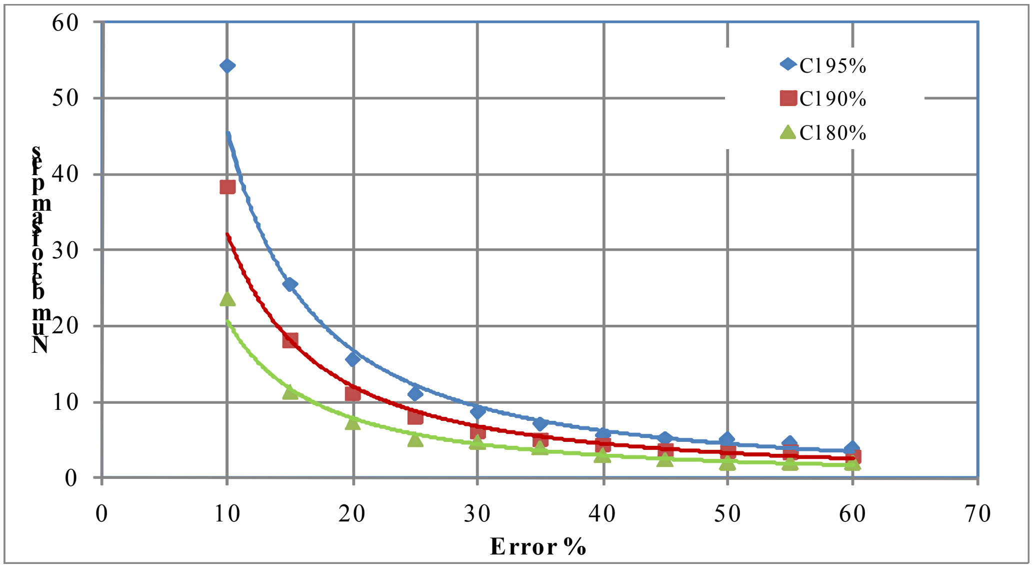


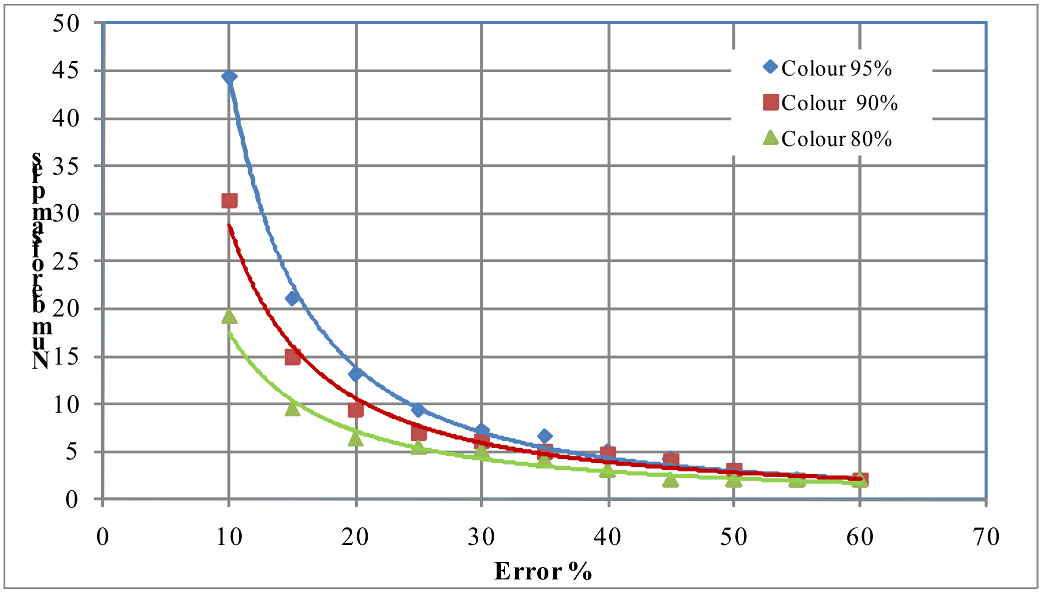
Number of samples
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%

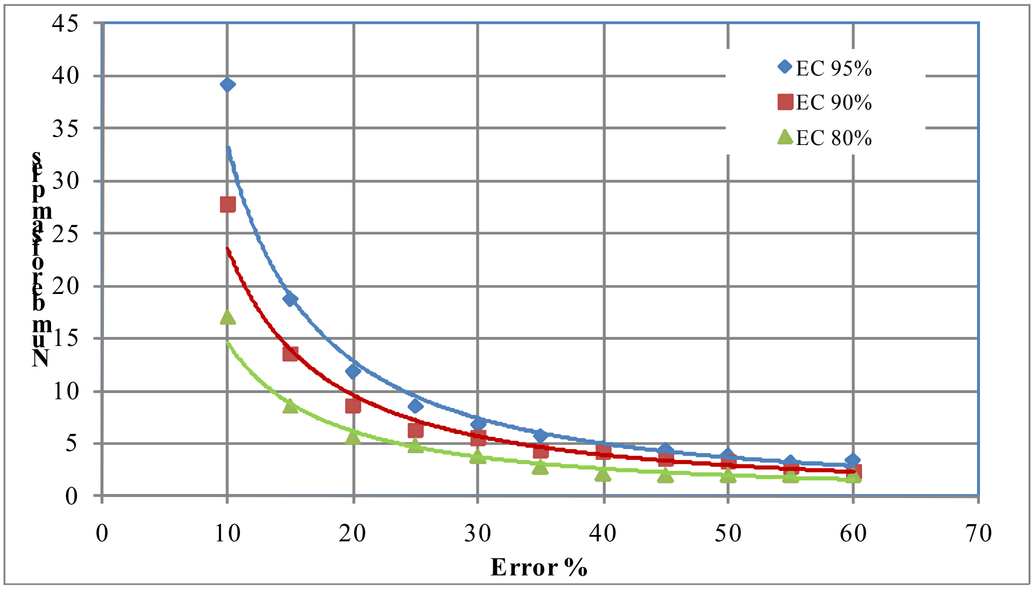
MB05SBS126

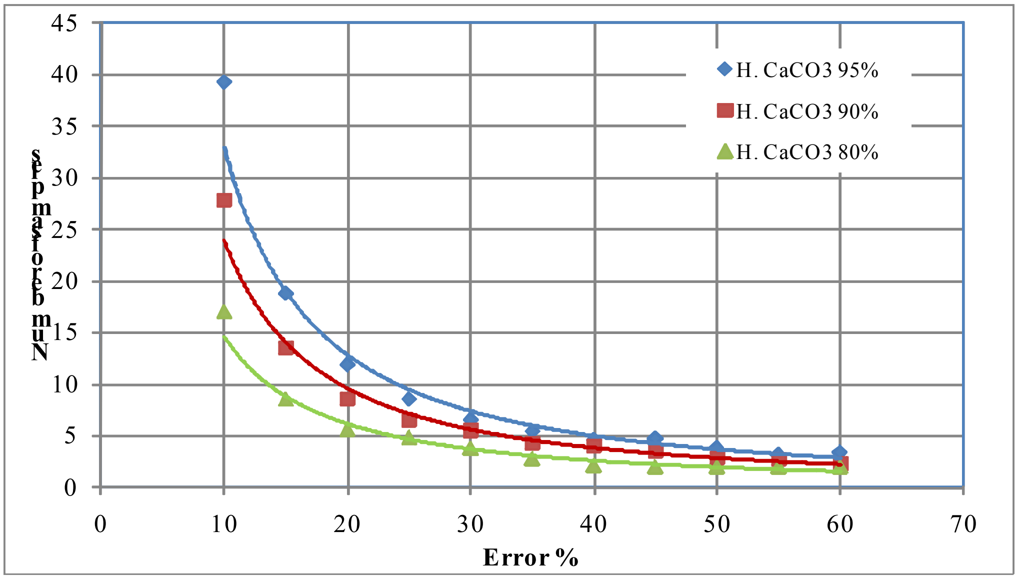

CaCO3 95%
CaCO3 90%
CaCO3 80%
Number of samples
0
5
10
15
20
25
30
0
10
20
30
40
50
60
70
Error %

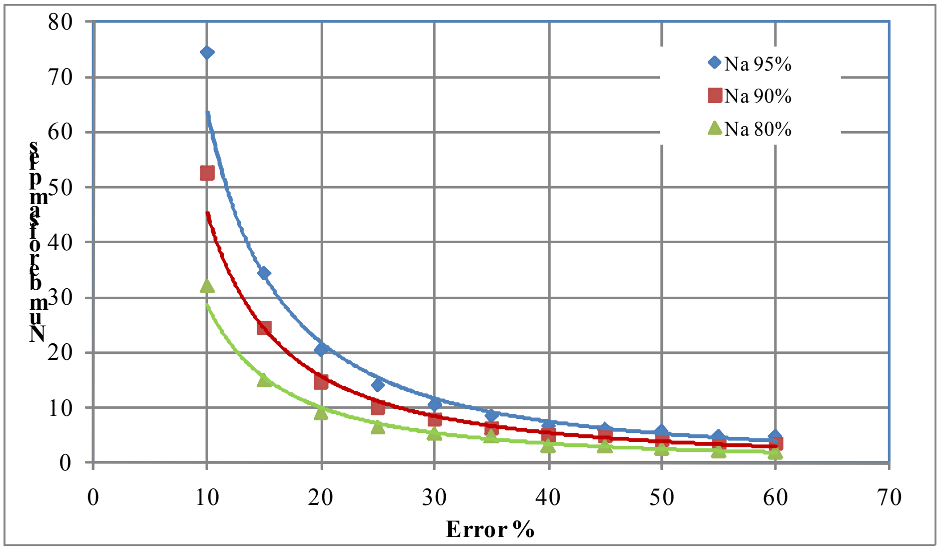

HCO3 95%
HCO3 90%
HCO3 80%
Number of samples
0
5
10
15
20
25
0
10
20
30
40
50
60
70
Error %

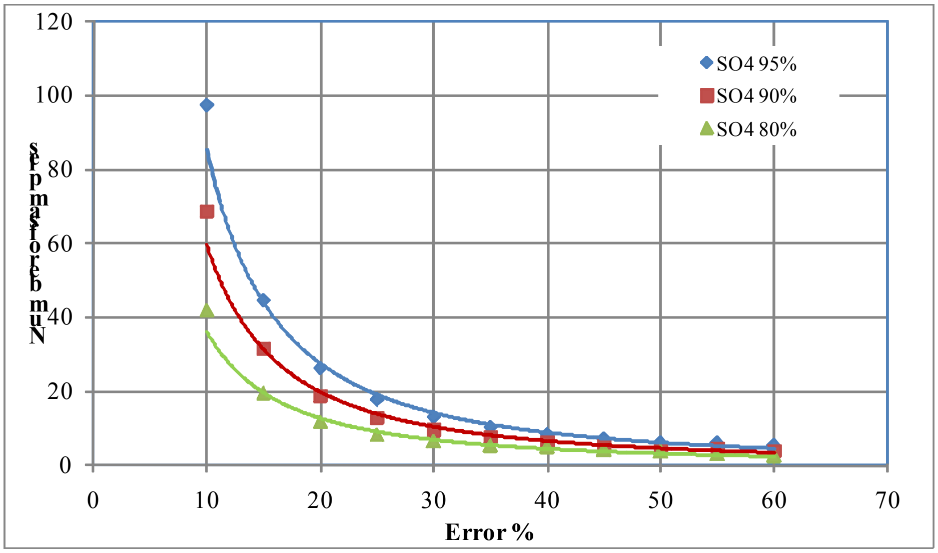

C195%
C190%
C180%
Number of samples
0
10
20
30
40
50
60
0
10
20
30
40
50
60
70
Error %

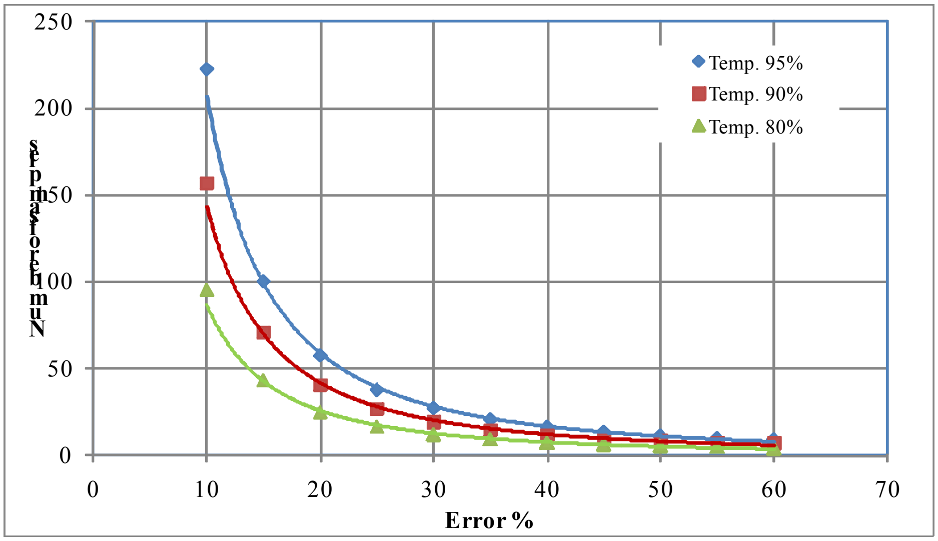

Colour 95%
Colour 90%
Colour 80%
Number of samples
0
5
10
15
20
25
30
35
40
45
50
0
10
20
30
40
50
60
70
Error %

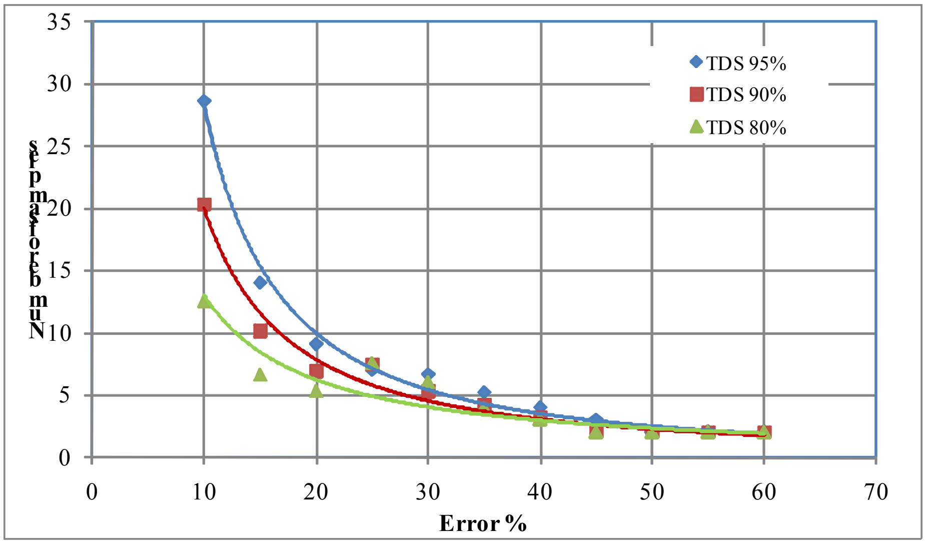

EC 95%
EC 90%
EC 80%
Number of samples
0
5
10
15
20
25
30
35
40
45
0
10
20
30
40
50
60
70
Error %

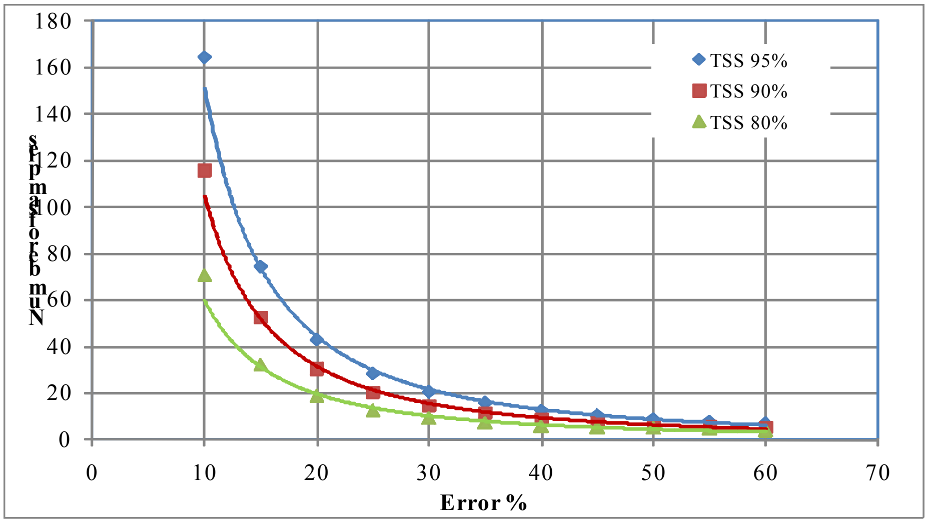

H. CaCO3 95%
H. CaCO3 90%
H. CaCO3 80%
Number of samples
0
5
10
15
20
25
30
35
40
45
0
10
20
30
40
50
60
70
Error %

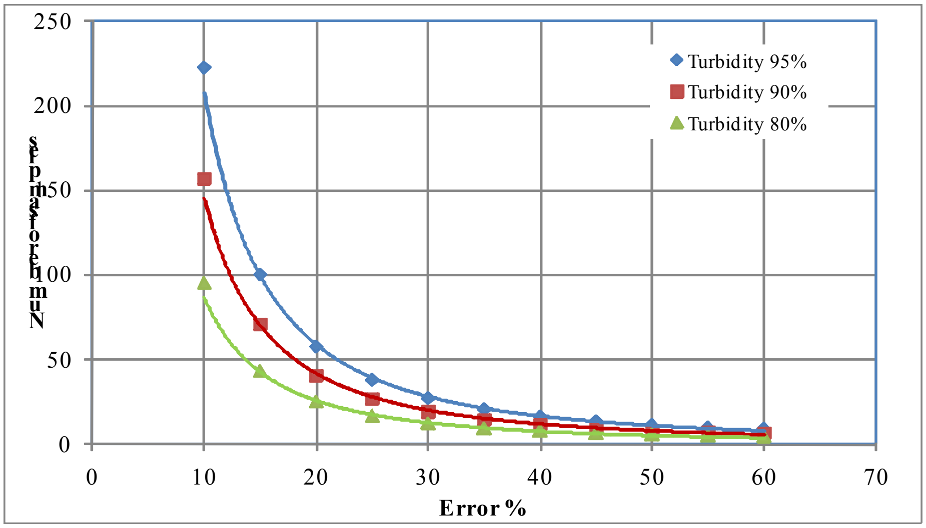

Na 95%
Na 90%
Na 80%
Number of samples
0
10
20
30
40
50
60
70
80
0
10
20
30
40
50
60
70
Error %

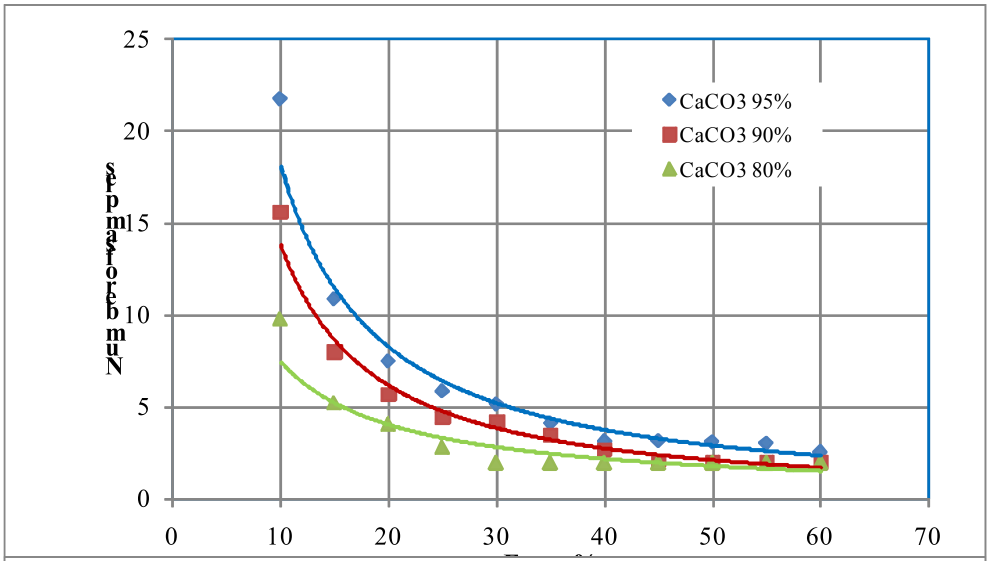

SO4 95%
SO4 90%
SO4 80%
Number of samples
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
Error %

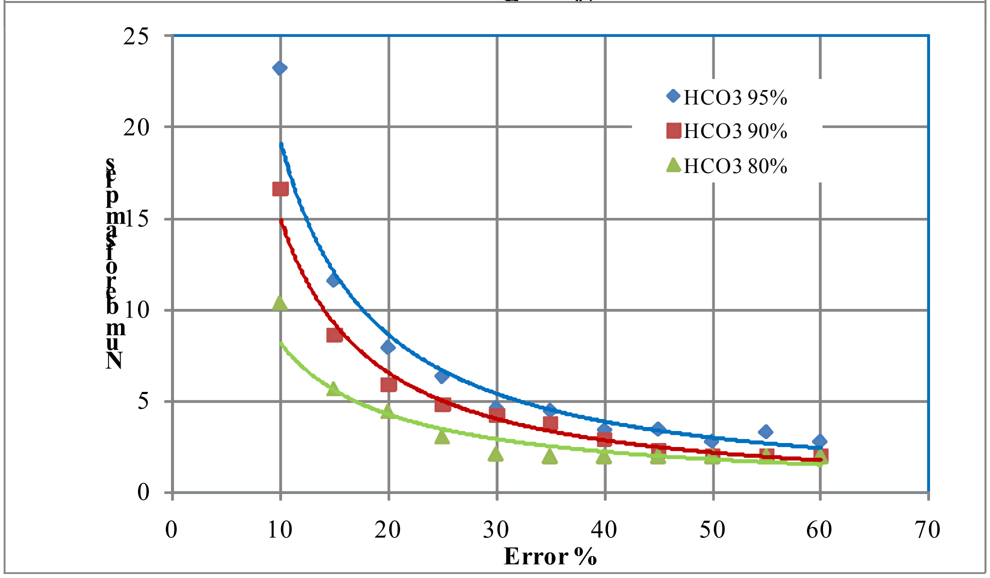

Temp. 95%
Temp. 90%
Temp. 80%
Number of samples
0
50
100
150
200
250
0
10
20
30
40
50
60
70
Error %

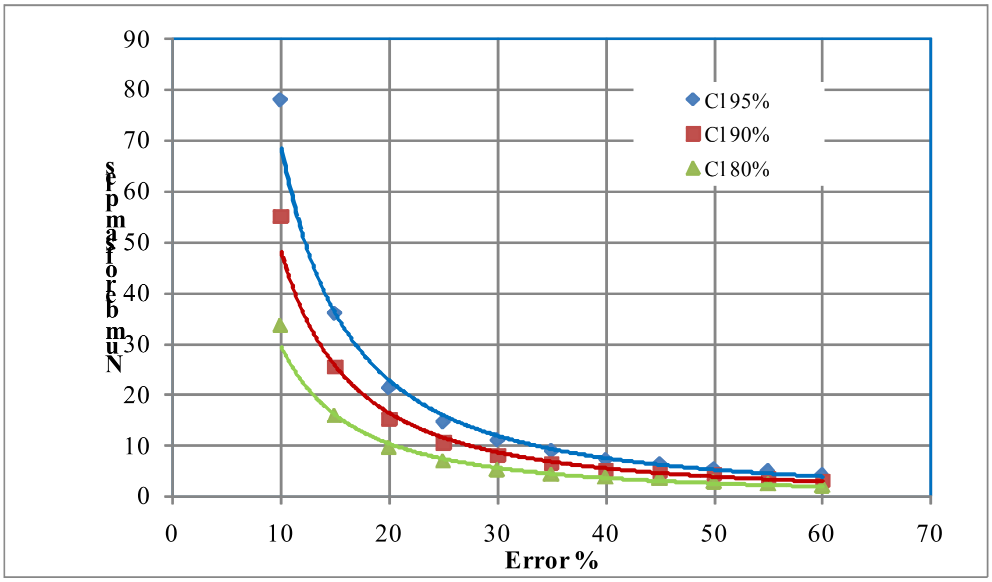

TDS 95%
TDS 90%
TDS 80%
Number of samples
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %

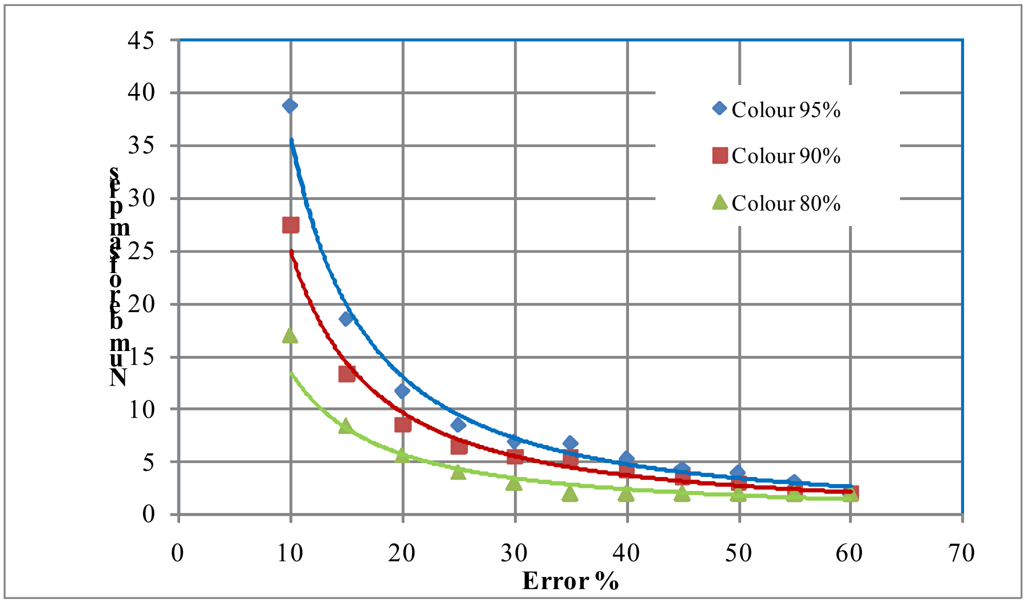

TSS 95%
TSS 90%
TSS 80%
Number of samples
180
160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Error %

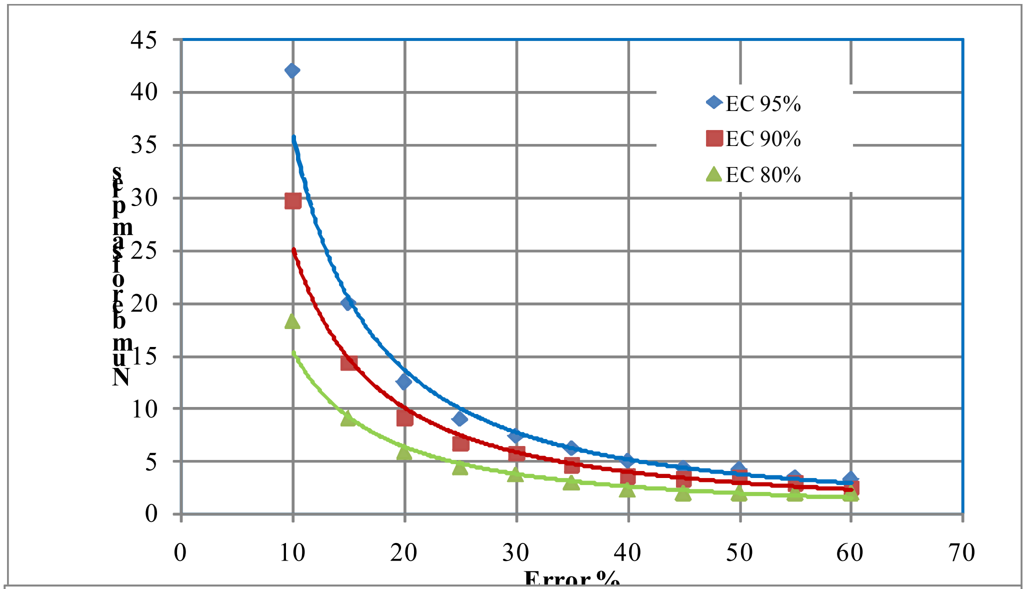

Turbidity 95%
Turbidity 90%
Turbidity 80%
Number of samples
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Error %

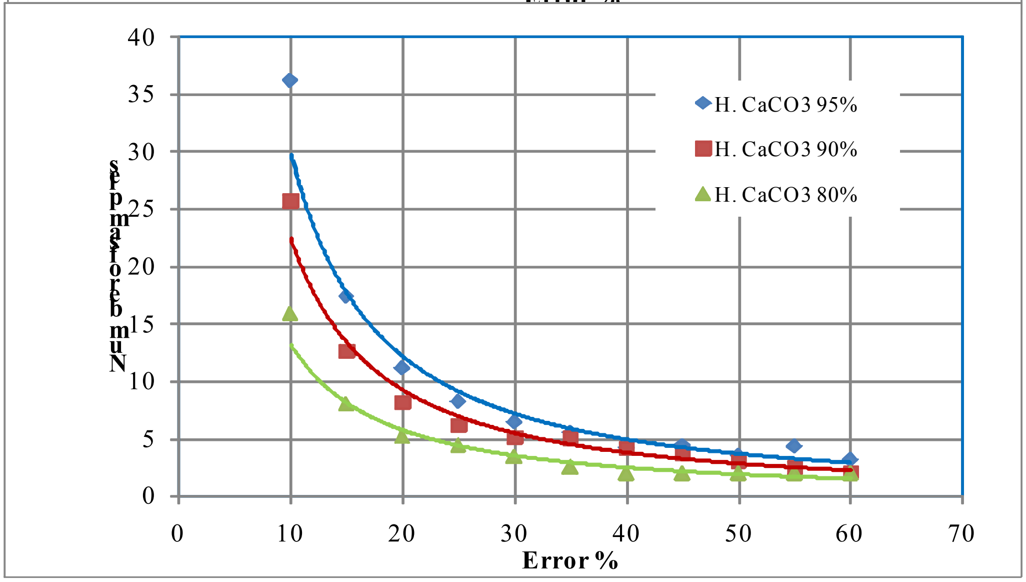

25
20
15
10
5
0
Number of samples
0
10
20
30
40
50
60
70
◆ CaCO3 95%
■ CaCO3 90%
▲ CaCO3 80%

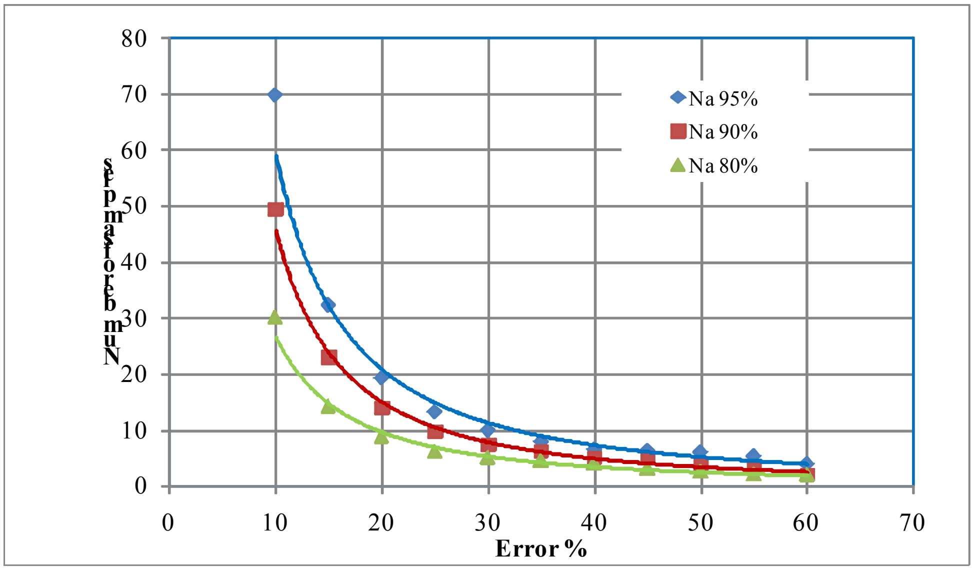

25
20
15
10
5
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ HCO3 95%
■ HCO3 90%
▲ HCO3 80%

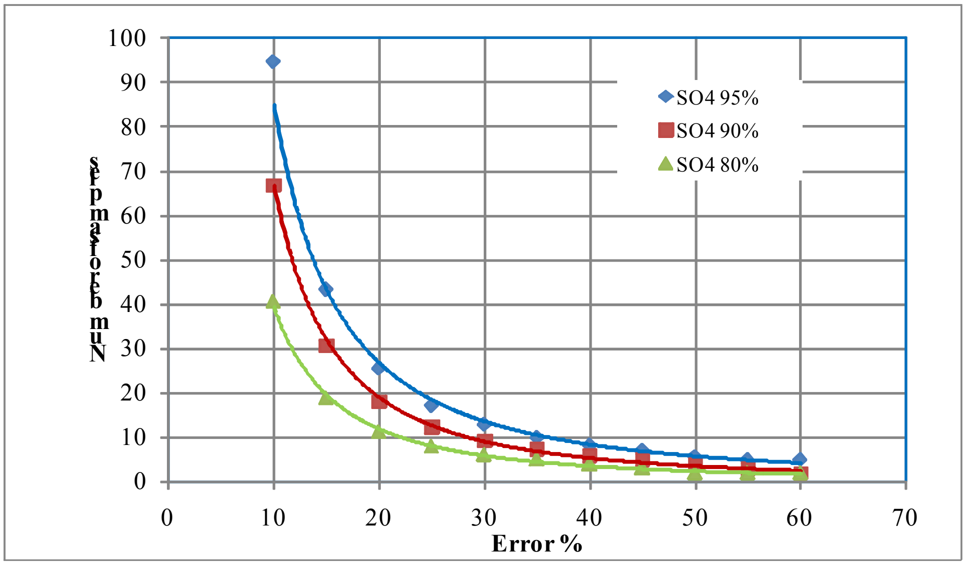

90
80
70
60
50
40
30
20
10
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ Cl 95%
■ Cl 90%
▲ Cl 80%

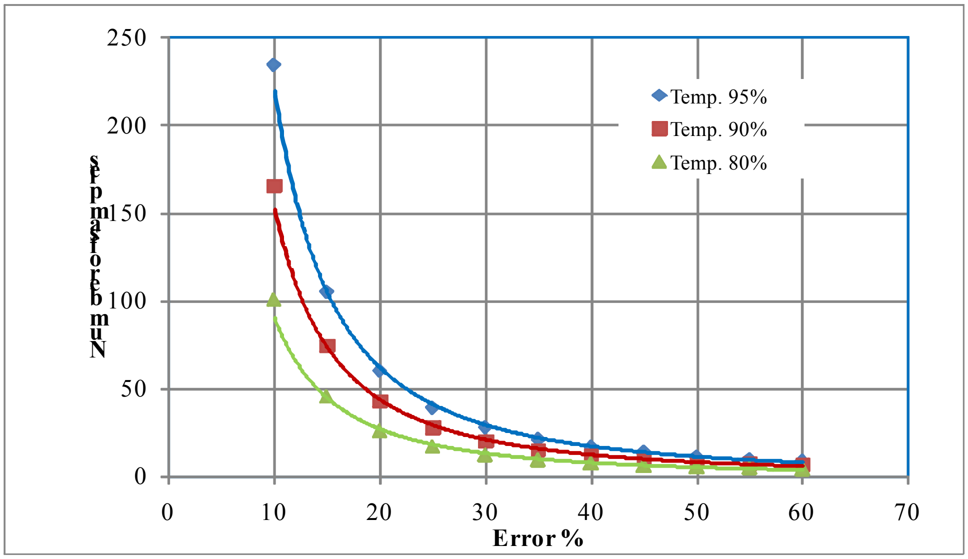

Number of samples
45
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %
Colour 95%
Colour 90%
Colour 80%

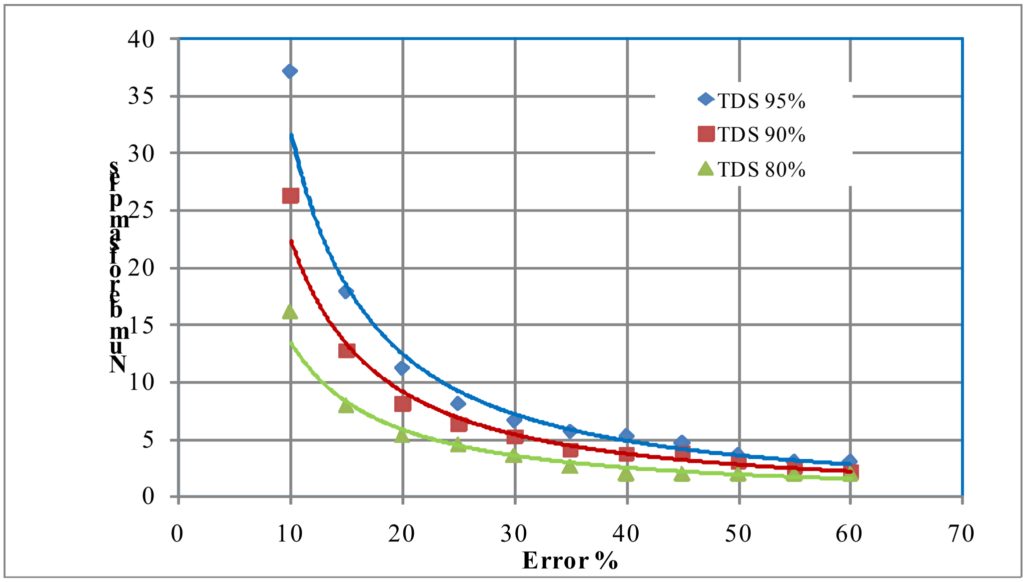

Number of samples
45
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %
EC 95%
EC 90%
EC 80%

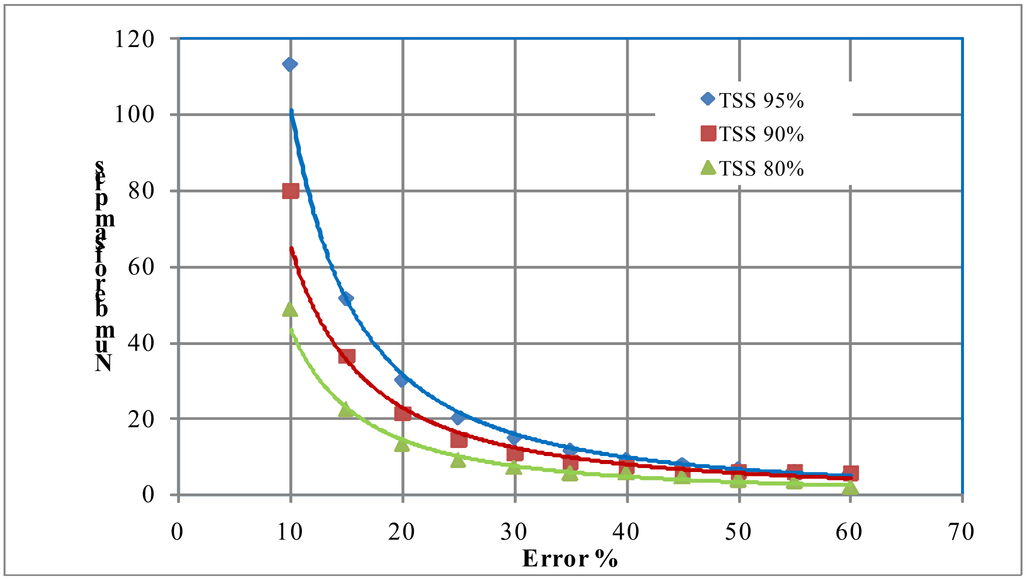

Number of samples
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Error %
H. CaCO3 95%
H. CaCO3 90%
H. CaCO3 80%

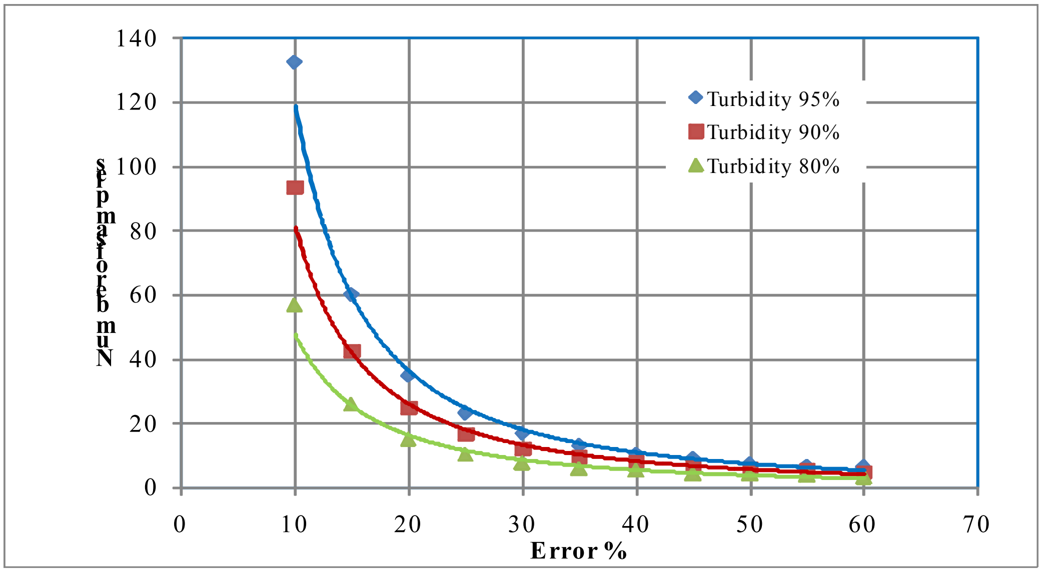

Number of samples
0
10
20
30
40
50
60
70
80
0
10
20
30
40
50
60
70
Error %
Na 95%
Na 90%
Na 80%

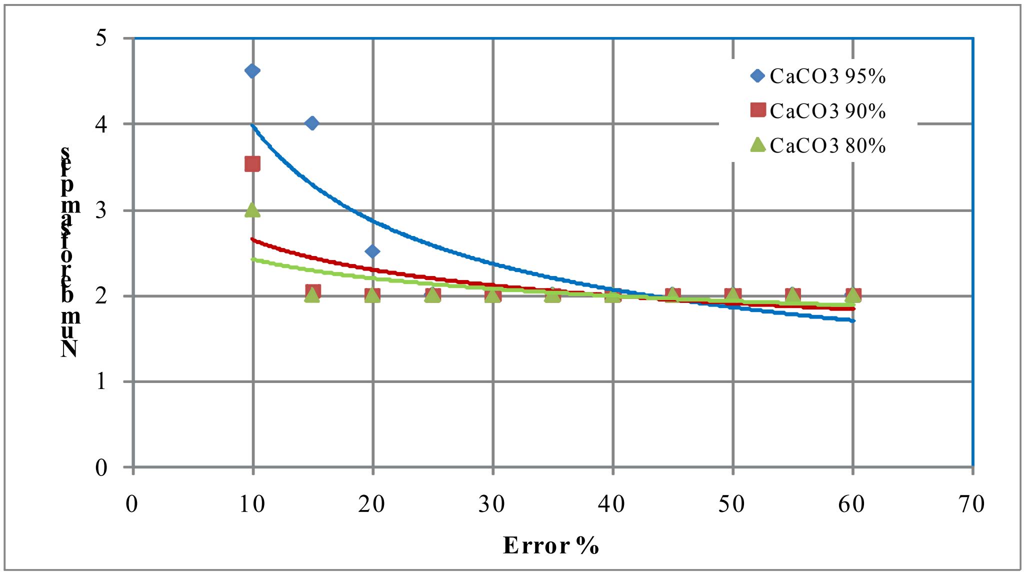

Number of samples
0
10
20
30
40
50
60
70
80
90
100
0
10
20
30
40
50
60
70
Error %
SO4 95%
SO4 90%
SO4 80%

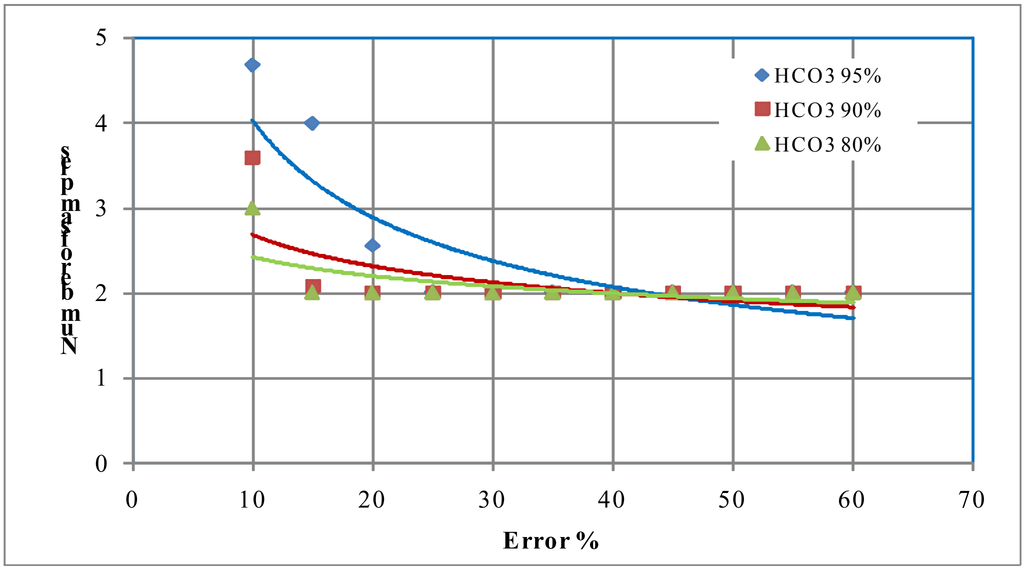

Number of samples
0
50
100
150
200
250
0
10
20
30
40
50
60
70
Error %
Temp. 95%
Temp. 90%
Temp. 80%

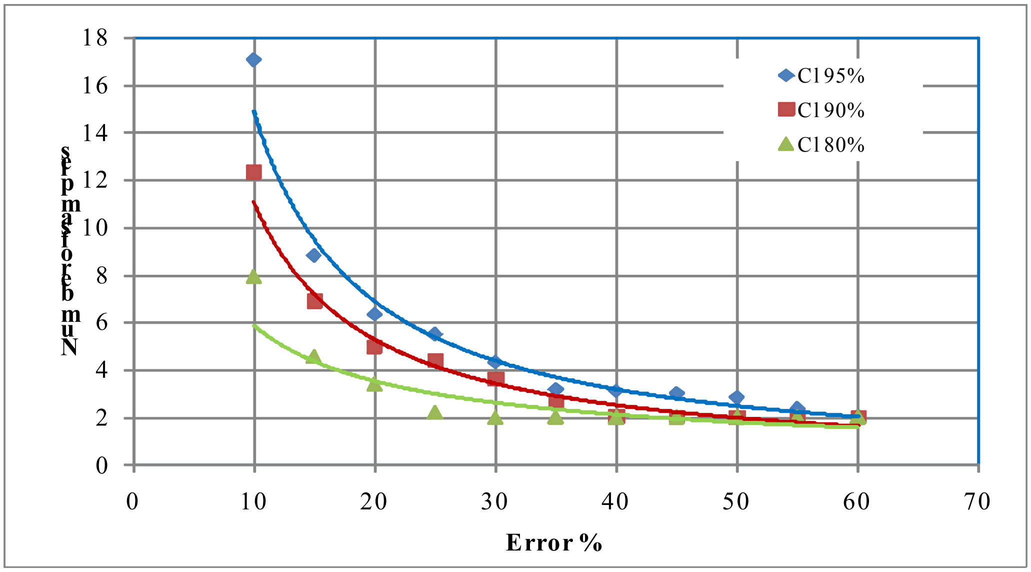

40
35
30
25
20
15
10
5
0
Number of samples
0
10
20
30
40
50
60
70
Error %
TDS 95%
TDS 90%
TDS 80%

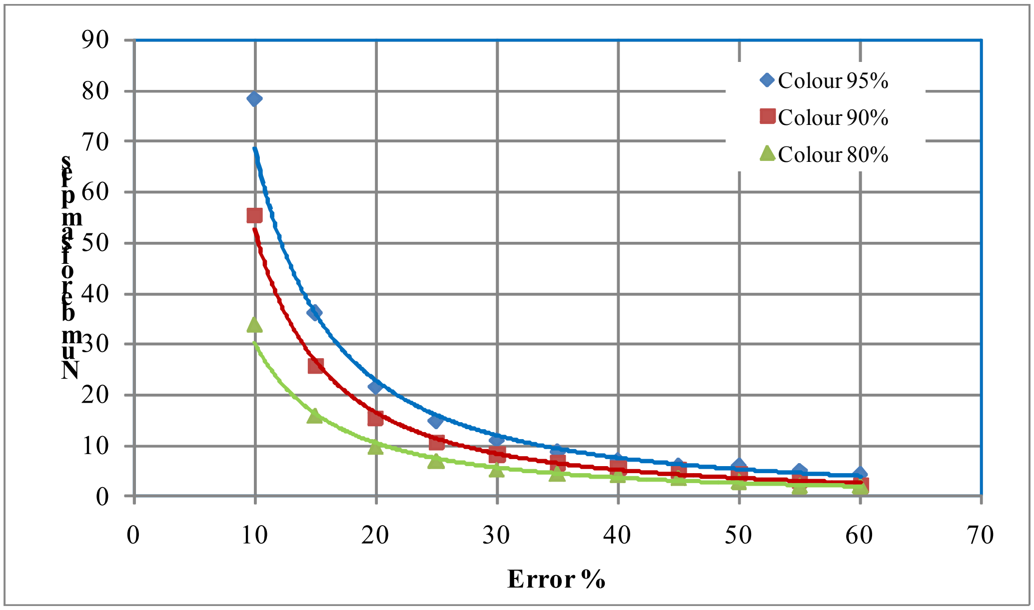

120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
TSS 95%
TSS 90%
TSS 80%

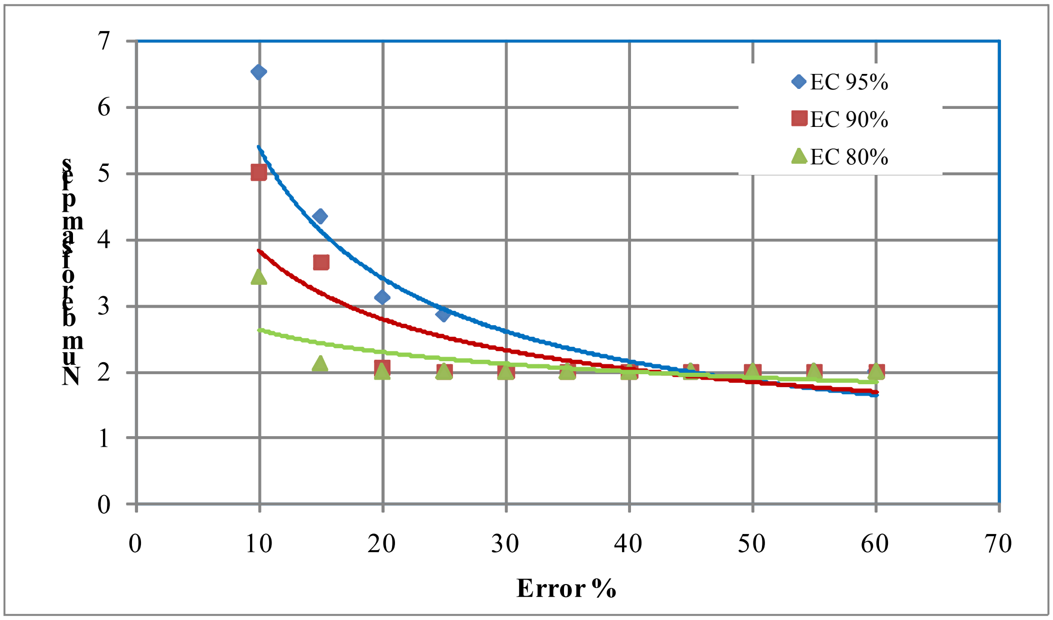

140
120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%

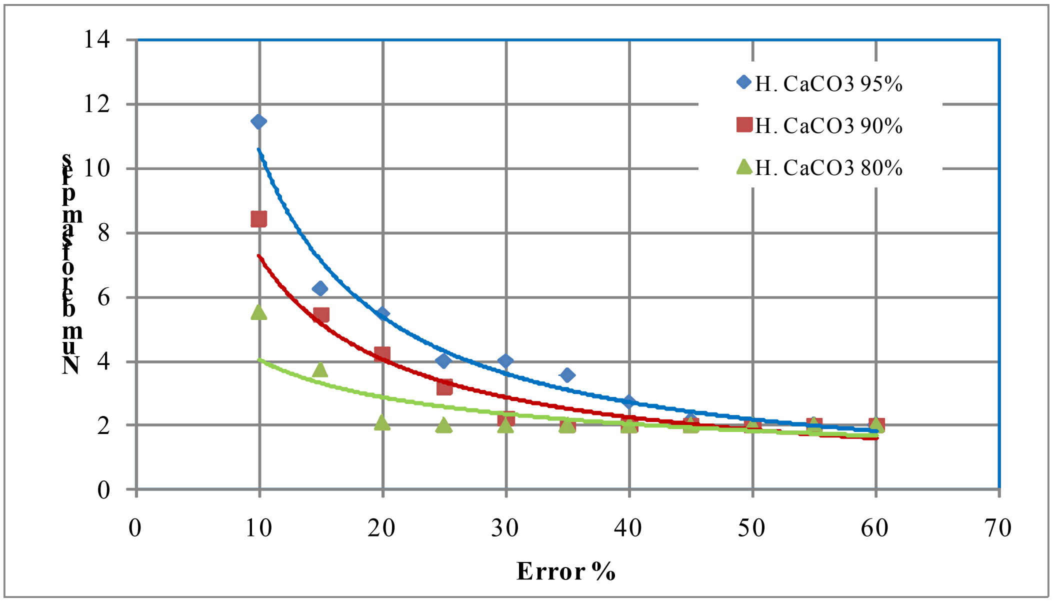

Number of samples
0
1
2
3
4
5
0
10
20
30
40
50
60
70
Error %
◆ CaCO3 95%
■ CaCO3 90%
▲ CaCO3 80%

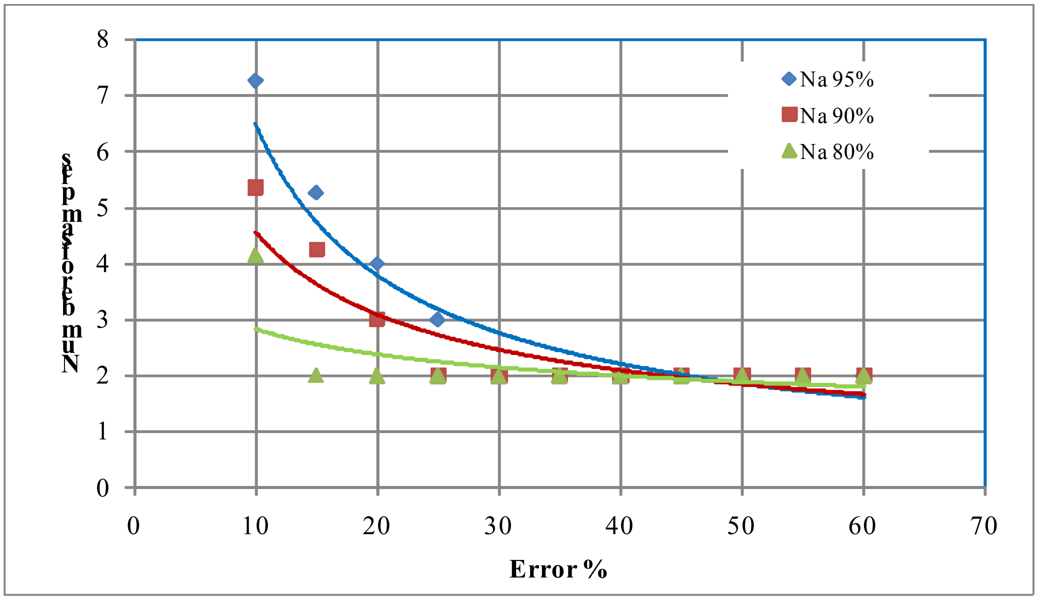

Number of samples
0
1
2
3
4
5
0
10
20
30
40
50
60
70
Error %
◆ HCO3 95%
■ HCO3 90%
▲ HCO3 80%

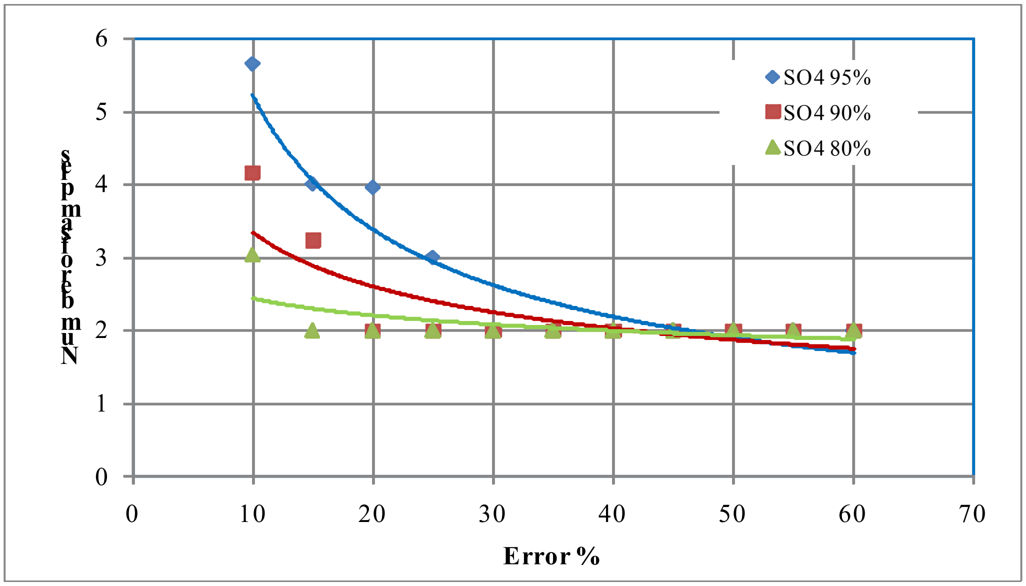

Number of samples
0
2
4
6
8
10
12
14
16
18
0
10
20
30
40
50
60
70
Error %
◆ Cl 95%
■ Cl 90%
▲ Cl 80%

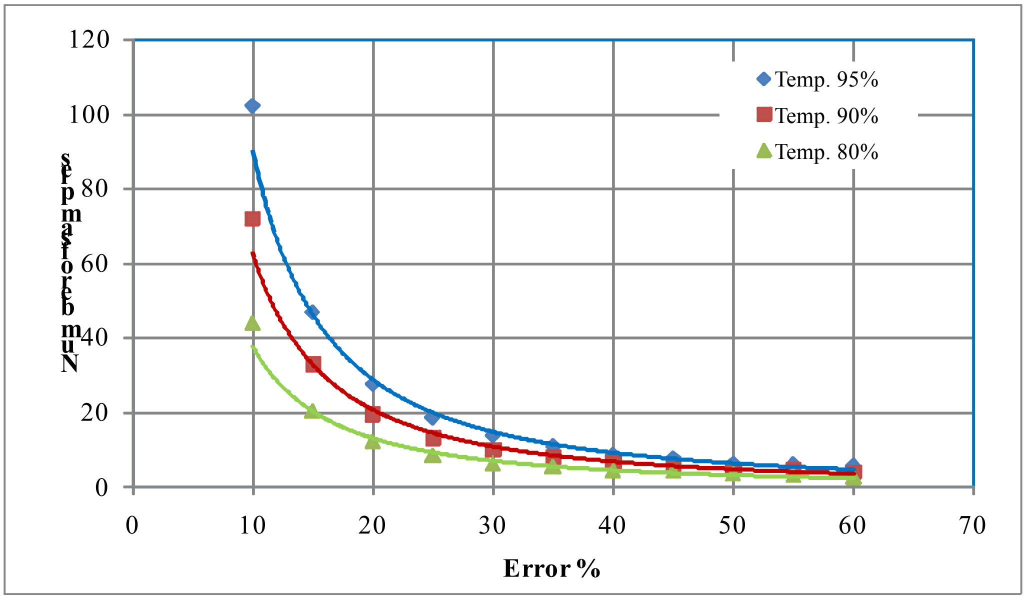

Number of samples
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
Error %
◆ Colour 95%
■ Colour 90%
▲ Colour 80%

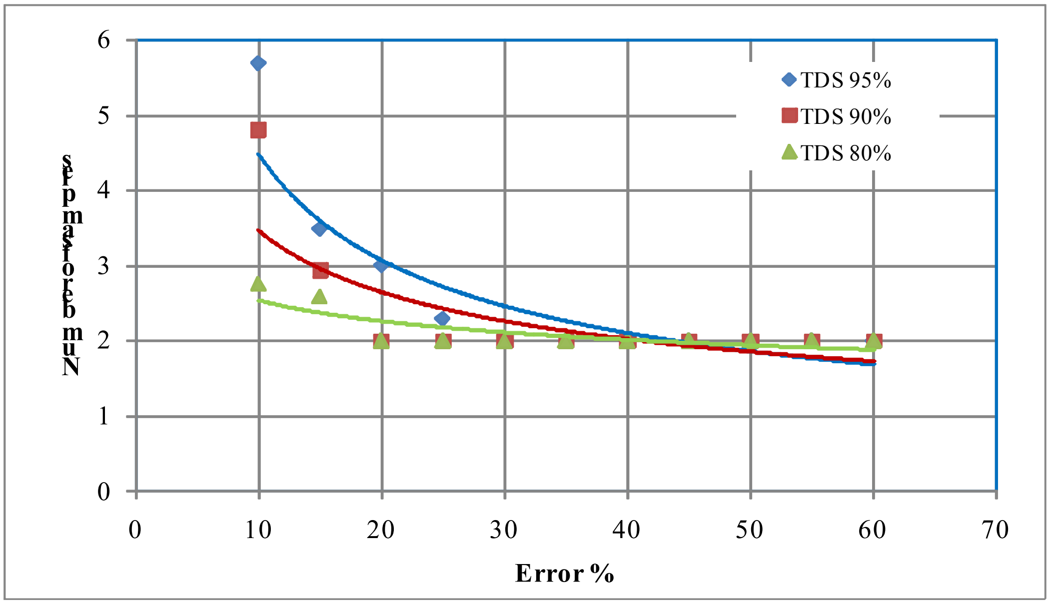

Number of samples
7
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
Error %
◆ EC 95%
■ EC 90%
▲ EC 80%

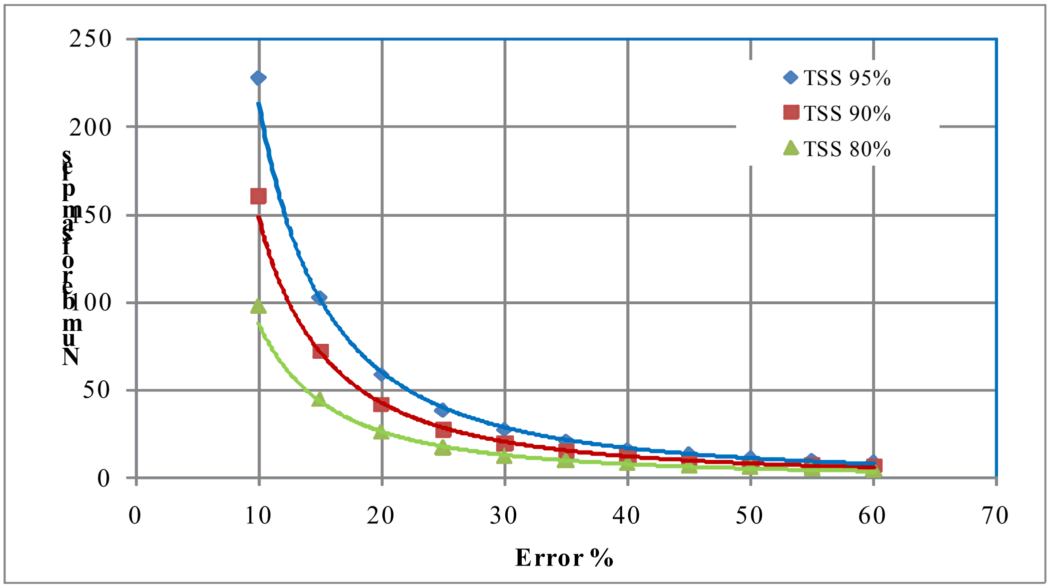

Number of samples
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
70
Error %
◆ H. CaCO3 95%
■ H. CaCO3 90%
▲ H. CaCO3 80%

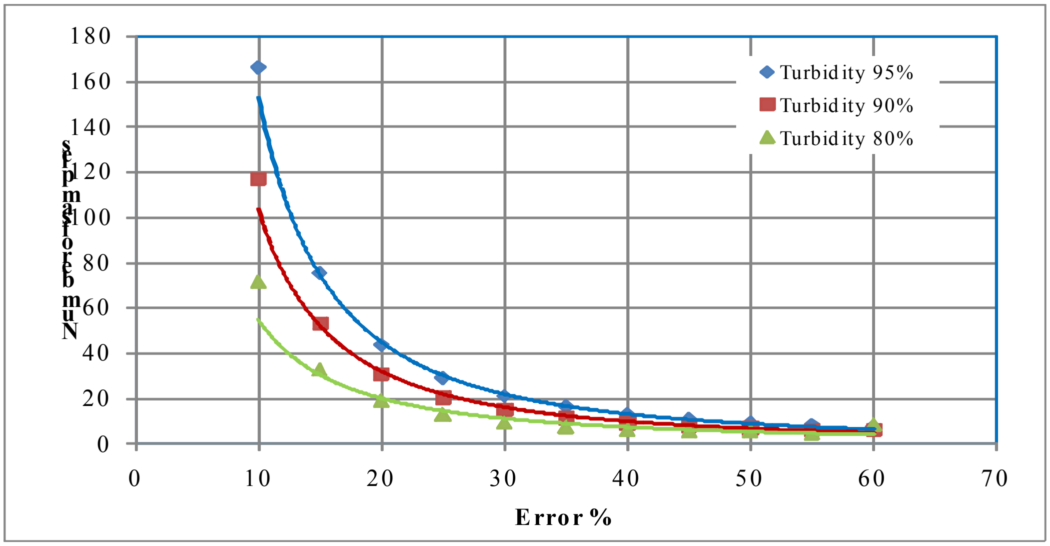

Number of samples
Error %
0
1
2
3
4
5
6
7
8
0
10
20
30
40
50
60
70
Na 95%
Na 90%
Na 80%

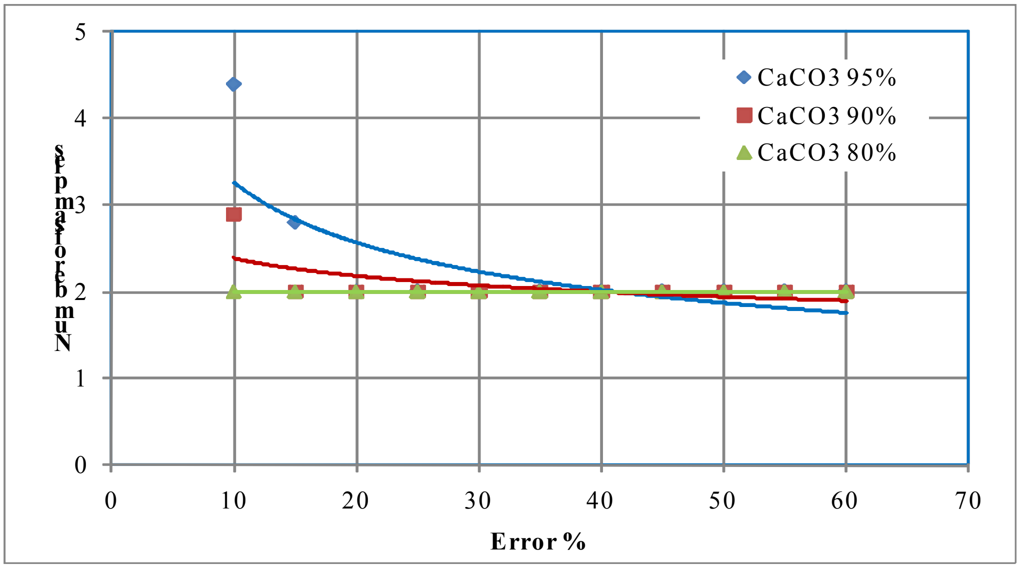

Number of samples
Error %
0
1
2
3
4
5
6
0
10
20
30
40
50
60
70
SO4 95%
SO4 90%
SO4 80%

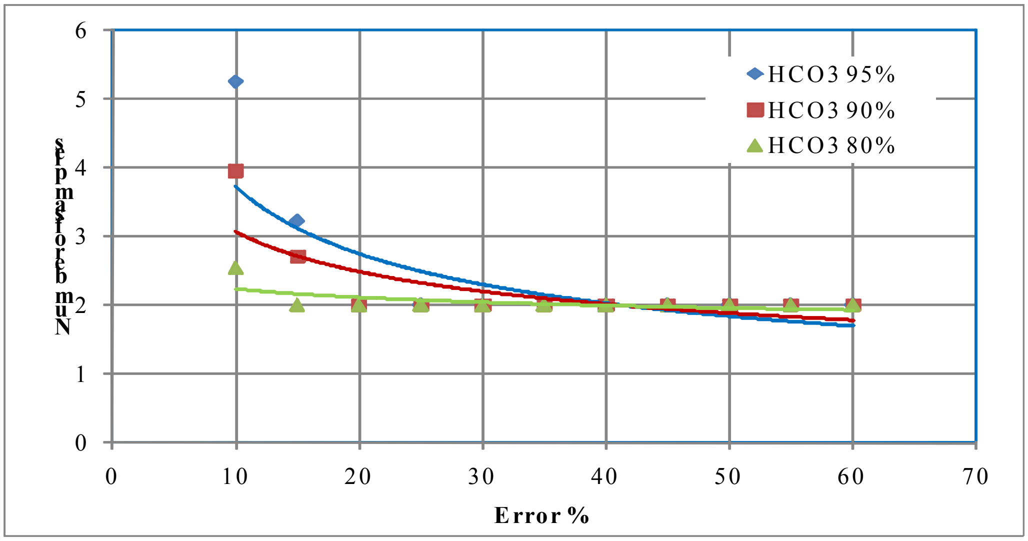

Number of samples
Error %
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
Temp. 95%
Temp. 90%
Temp. 80%

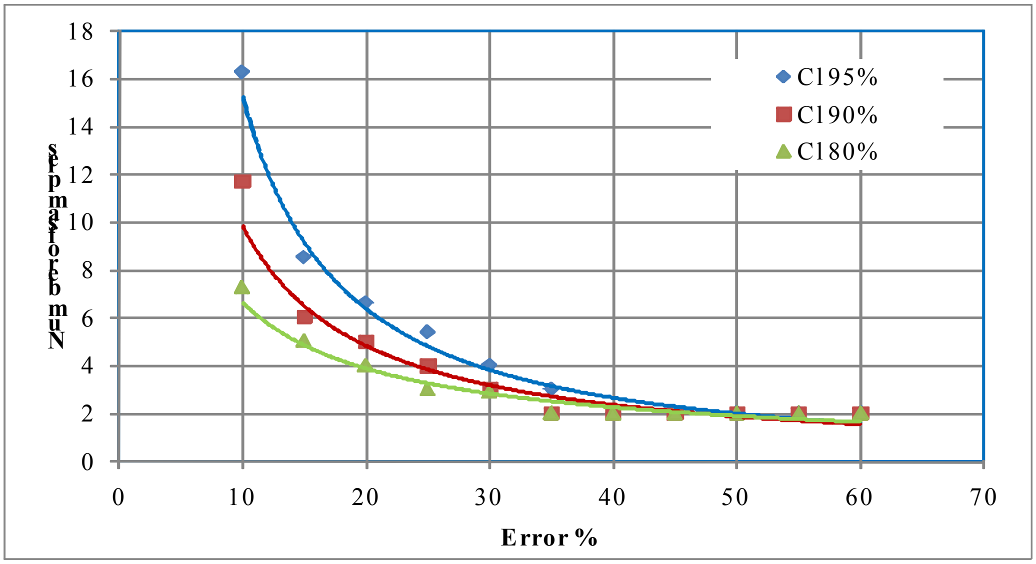

6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
TDS 95%
TDS 90%
TDS 80%

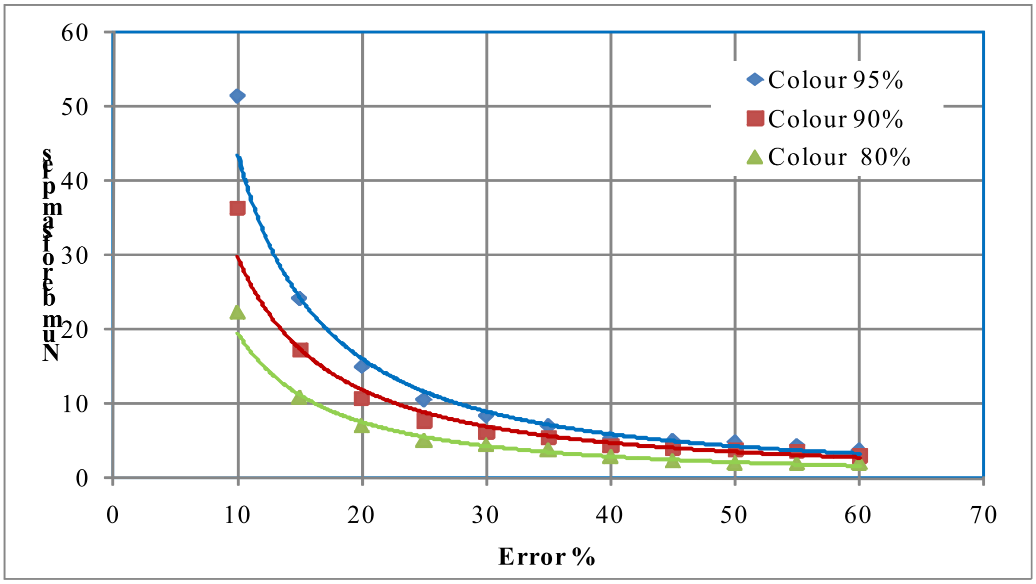

250
200
150
100
50
0
Number of samples
0
10
20
30
40
50
60
70
Error %
TSS 95%
TSS 90%
TSS 80%


180
160
140
120
100
80
60
40
20
0
Number of samples
0
10
20
30
40
50
60
70
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%

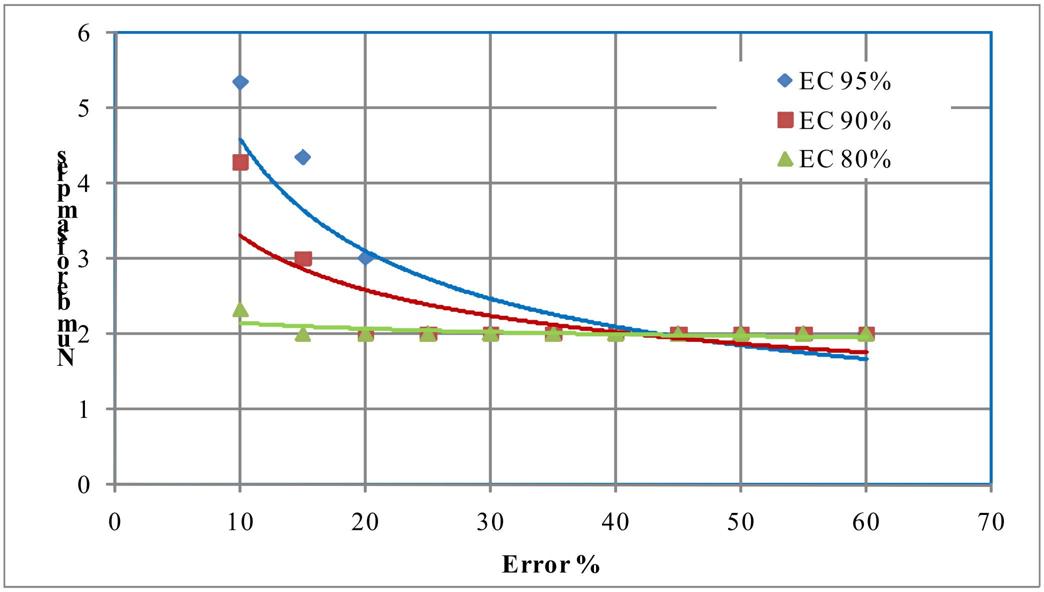

5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ CaCO3 95%
■ CaCO3 90%
▲ CaCO3 80%


6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ HCO3 95%
■ HCO3 90%
▲ HCO3 80%

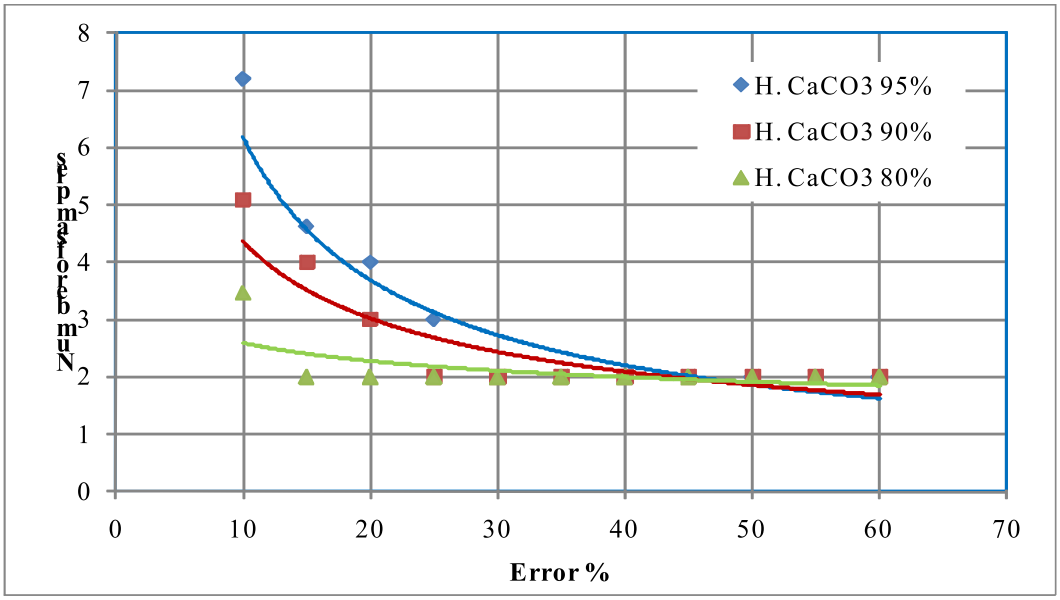

18
16
14
12
10
8
6
4
2
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ Cl 95%
■ Cl 90%
▲ Cl 80%



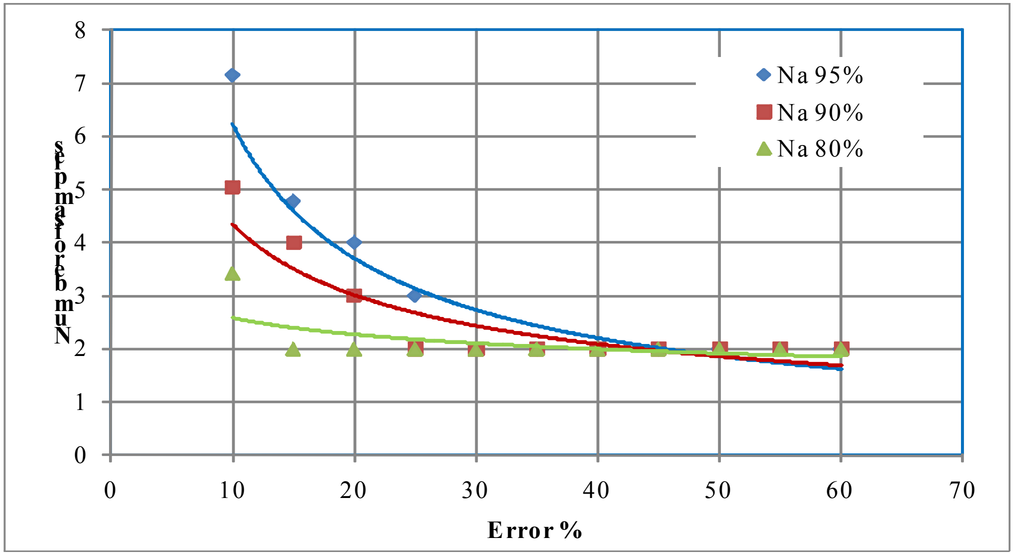
60
50
40
30
20
10
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ Colour 95%
■ Colour 90%
▲ Colour 80%

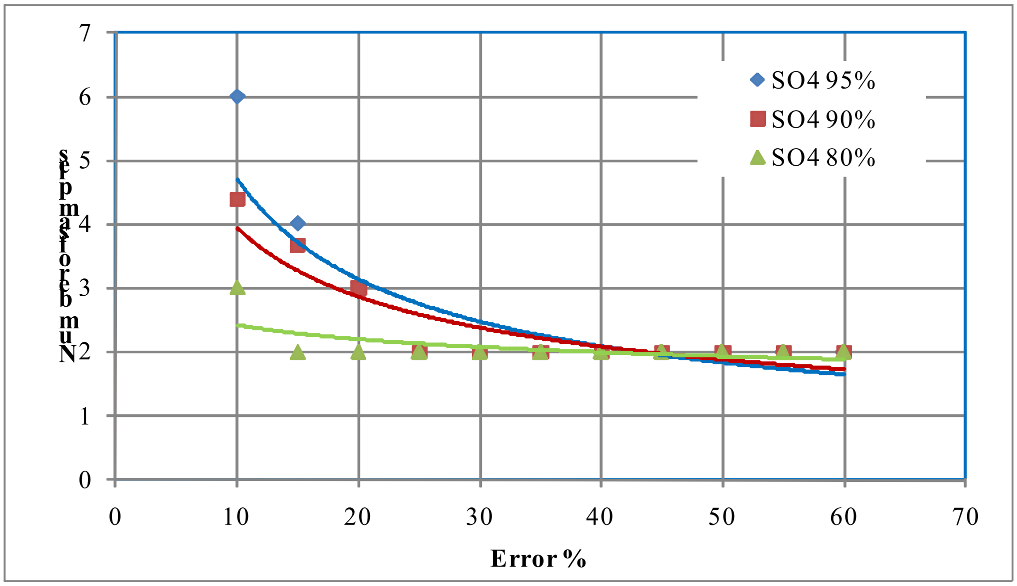


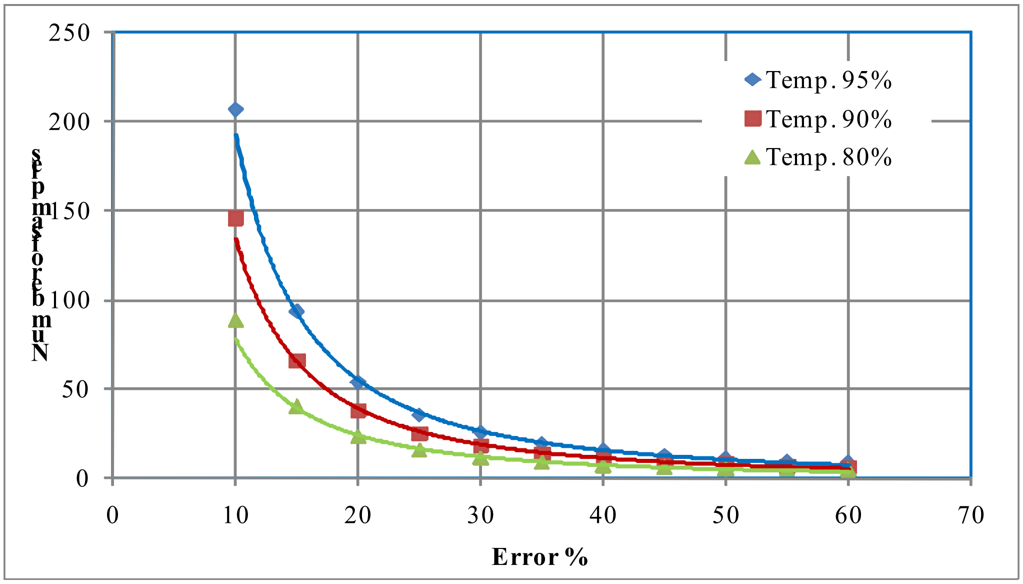
6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ EC 95%
■ EC 90%
▲ EC 80%

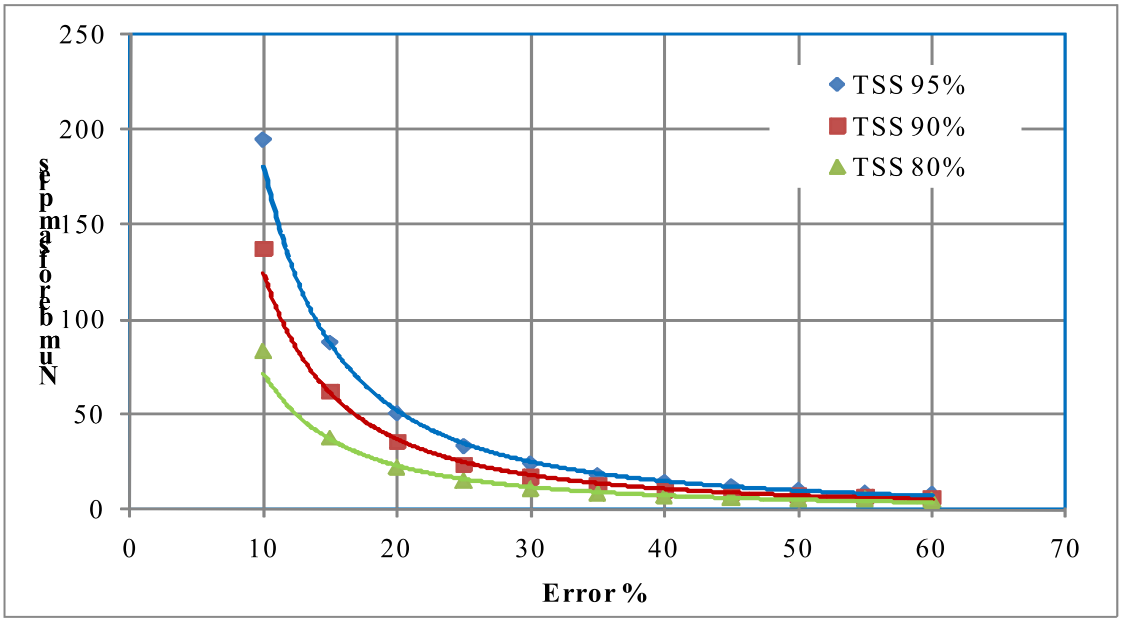


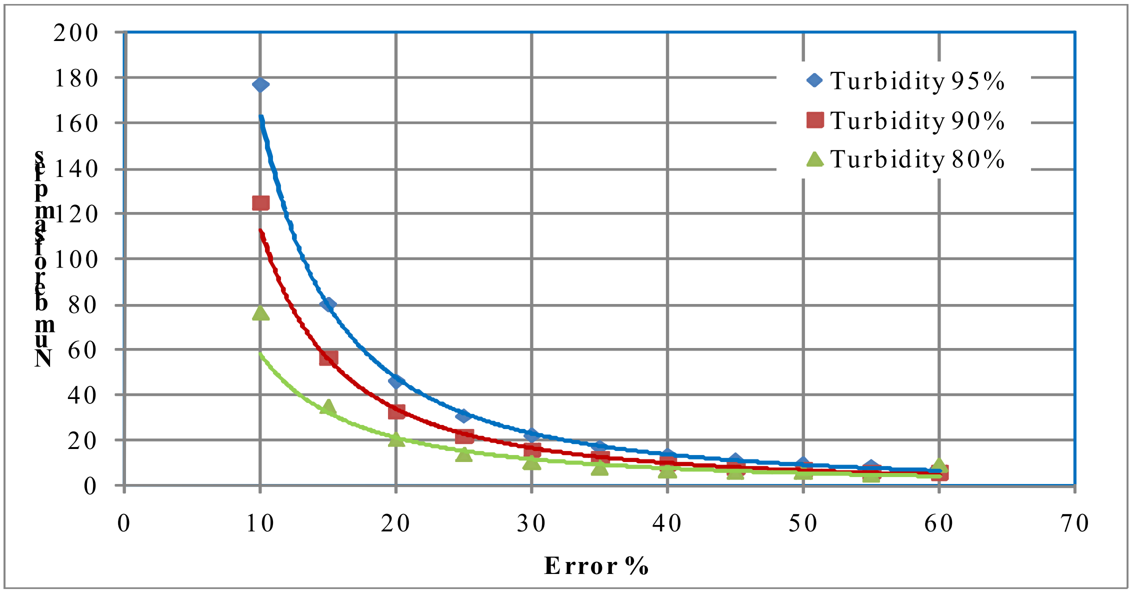
8
7
6
5
4
3
2
1
0
Number of samples
0
10
20
30
40
50
60
70
Error %
◆ H. CaCO3 95%
■ H. CaCO3 90%
▲ H. CaCO3 80%

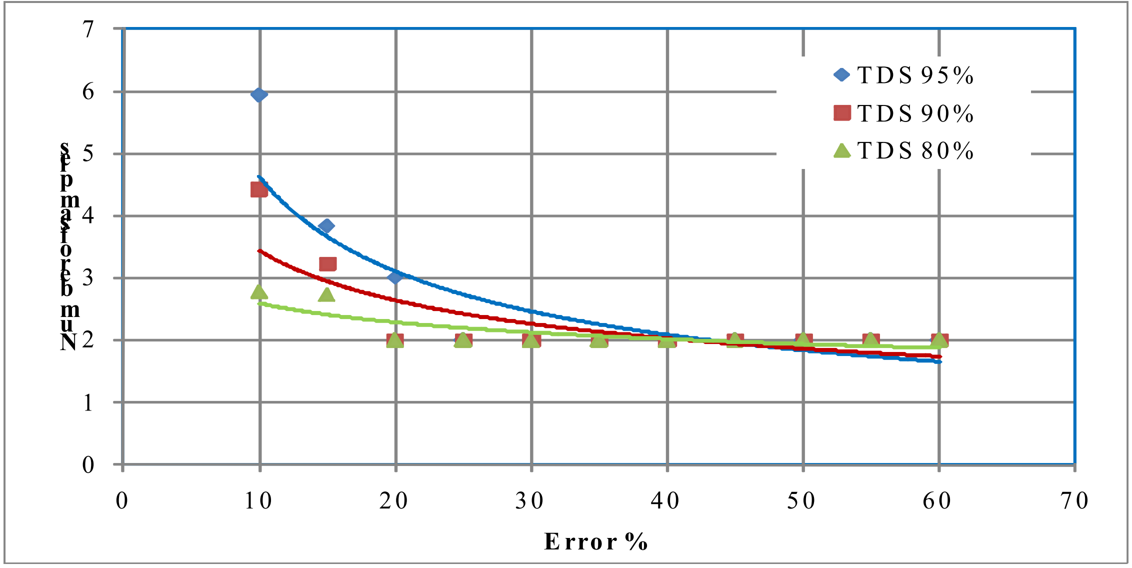

Number of samples
8
7
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
Error %
Na 95%
Na 90%
Na 80%


Number of samples
7
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
Error %
SO4 95%
SO4 90%
SO4 80%


Number of samples
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Error %
Temp. 95%
Temp. 90%
Temp. 80%


250
200
150
100
50
0
0
10
20
30
40
50
60
70
Number of samples
Error %
TSS 95%
TSS 90%
TSS 80%


200
180
160
140
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Number of samples
Error %
Turbidity 95%
Turbidity 90%
Turbidity 80%


7
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
Number of samples
Error %
TDS 95%
TDS 90%
TDS 80%