ELSEVIER

Chemometrics and Intelligent Laboratory Systems 60 (2002) 265–281

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Treatment of nondetects in multivariate analysis of groundwater geochemistry data

Irene M. Farnham [a,*], Ashok K. Singh [b], Klaus J. Stetzenbach [a], Kevin H. Johannesson [c]

[a] *Harry Reid Center for Environmental Studies, University of Nevada, Las Vegas, NV 89154-4009, USA*
[b] *Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154-4020, USA*
[c] *Department of Ocean, Earth, and Atmospheric Sciences, Old Dominion University, Norfolk, VA 23529-0276, USA*

Abstract

Principal components analysis (PCA) is used to evaluate similarities in the trace element chemistry of groundwaters. Many of the trace elements, however, occur at concentrations below the detection limits (DL), which presents problems for statistical analyses. Since the optimal methods for dealing with the '<DL' values require the assumption of normality or lognormality, which is not the case for these data, simple substitution methods may be preferred.

In this study, a new approach was developed to determine the best substitution methods when dealing with the 'DL' values for a given data set. Monte Carlo simulation experiments, using a mixture multivariate model, were performed to test the effects of substitution of the '<DL' values with 0, DL, and DL/2 on the results of PCA. In general, the results showed that substitution with DL/2 gave superior results compared to substitution with DL or 0. Deterioration of the performance of all substitution methods was observed when the number of '<DL' values exceeded approximately 25%.

When '<DL' values are present within a data set, use of uncensored data (if available) may be preferred. Use of uncensored data requires statistical methodology to take into account varying uncertainties in the measurements. A new approach for the selection of variables for multivariate analysis of the trace element chemistry data was therefore developed. Selection of variables was based on the reproducibility between measurements made on duplicate samples. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Multivariate analysis; Groundwater; Nondetects

1. Introduction

The trace element chemistry of groundwater may be used to assign a fingerprint with which water from similar flow paths can be identified [1–6]. The concentrations of over 50 trace elements in groundwater samples are measured and principal components analysis (PCA) is applied for data reduction. Informal clustering of springs and wells, based on the PCA results, is then used to evaluate chemical similarities between the groundwaters. Similarities are also examined using the clustering techniques, *k*-means clustering, and hierarchical clustering [6].

Difficulties arise in the application of PCA and cluster analysis when the concentration of an element is below the detection limit (DL). It is often standard laboratory practice to report these data simply as 'DL'

* Corresponding author.
E-mail address: Farnham@Nevada.edu (I.M. Farnham).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(01)00201-5

14536

values. Such 'censoring' of data, however, can complicate all subsequent statistical analyses.

The procedures for handling censored data depends on the technical application involved. The best method to use generally depends on the amount of data below the detection limit, the size of the data set, and the probability distribution of the measurements. When the number of '<DL' observations is small, replacing them with a constant (i.e., DL/2) is generally satisfactory [7]. The values that are commonly used to replace the '<DL' values are 0, DL, or DL/2. Distributional methods such as the marginal maximum likelihood estimation [8] or more robust techniques [9] are often required when a large number of '<DL' observations are present. Substitution of a constant may still be preferred over the more complicated methods such as maximum likelihood techniques and regression order statistics when the data sets are very small ($n<10$). This is due to the inability to accurately infer the distributional properties from small data sets that are required for these methods [10]. Substitution of censored data with a constant value has also been shown to give preferred results over substitution of missing data using PCA [11].

Another approach to deal with data below the detection limit is to use the measured value (i.e., uncensored data). Numerical values are often produced from the analytical instrument but are not reported by the chemist because the value is judged "unreliable" [12]. Although such censoring of data prevents the misuse of low quality data, it commonly leads to a loss of important information [13]. Samples taken over time may be censored at different levels as changes in the precision of a method occur. This is often the case owing to changes in sensitivity of the analytical instrumentation over time or when comparing measurements made in different laboratories. This can lead to very complicated patterns of censoring [14]. Recording only the DL can create an uncorrectable bias in determining long-term trends [15]. In fact, trends are often more effectively detected in uncensored data than censored data even when the censored data are highly unreliable [16]. Censoring data is therefore not recommended in most cases. Instead, it is generally preferred to report the measured value along with an estimate of measurement precision [14,15,17].

Although censoring data is not the recommended practice in most data reporting situations, it is still the most common. Techniques are therefore required to deal with the '<DL' values when uncensored data is not available. As is often the case, the distribution of the geochemistry data evaluated in this study is neither normal nor lognormal, thus eliminating the applicability of the distributional methods. This leaves simple substitution with a constant as one viable alternative.

The goal of this study was to develop an approach to determine the best constant (0, DL, or DL/2) to substitute for values that are below the detection limit and also to evaluate the effect of increasing the percentage of '<DL' values on the results of PCA. In our approach, which is similar to the parametric bootstrap method, the structure of the parent data matrix containing the concentrations of trace elements in groundwater samples was first simulated. A detection limit was then applied that resulted in a specified number of values below the detection limit (5%, 10%, 15%, 20%, 30%, 40%, 50%, and 75%). All values below this DL were substituted with 0, DL, and DL/2. PCA results for these substituted data sets were compared to those of the unsubstituted "true" data set. The use of constants for substituting values below the detection limit was further examined by comparing the PCA results when different detection limits are associated with each set of samples. Although, in this study, the censoring limit is referred to as the detection limit, this approach equally applies to data that is censored at the quantitation limit or any other specified threshold limit.

In addition to the evaluation of substitution methods, an approach for the selection of variables was developed. It is important that the variables (i.e., trace elements) included in the PCA are good descriptors of the groundwater systems that are being compared and not simply describing noise. Although a high uncertainty may be associated with the measurement of a given trace element in a sample, this uncertainty may or may not be significant when compared to the differences in concentrations between samples from different locations. A method was therefore developed to select the variables based on the variance between duplicate measurements relative to the variance observed over all samples within the study area. Selection of variables based on this relative variance

may become crucial when standardized data are used for the PCA. Geochemical data are typically standardized to unit variance prior to PCA because the concentrations of the various chemicals often vary by several orders of magnitude. This often scales up the influence of the variables that are relatively constant [18] and that may be describing primarily the measurement variability on the PCA. This can severely impact the PCA when a significant number of these variables are present. This approach for the selection of variables also allows for the use of uncensored data when available. Although use of uncensored data is often desirable, it is important that some measure of uncertainty be associated with each of the measurements. Variance between duplicates is generally the best measure of uncertainty that is available.

2. Methods

2.1. Sample collection and analysis

Groundwater samples were collected following ultra-clean sampling techniques [19]. For sampling of spring water, precleaned (acid-washed) Teflon® tubing was placed near the spring's source and groundwater was then pumped, using a peristaltic pump, through Gelman Sciences (0.45 μm) groundwater filter capsules. Before filling the sample bottles (precleaned, acid-washed high-density linear polyethylene) with a particular groundwater sample, the bottles were rinsed three times with filtered groundwater. The samples were immediately acidified to pH < 2 with ultra-pure nitric acid (SeaStar). Sampling for well waters was essentially identical to spring water sampling, except that prior to sample collection, each well was pumped for a number of well volumes in order to ensure that the collected sample was representative of the actual groundwater and not water that had been sitting in the well bore for some time. In addition, the well waters were collected only after the pH, temperature, and conductivity of the well water had stabilized. The sample bottles containing the groundwater samples were double bagged in clean plastic bags, placed into clean plastic chests on ice, and transported to the laboratory where they were stored at approximately 4 °C. Samples were then analyzed for trace elements using an inductively coupled plasma mass spectrometer (ICP-MS) (Perkin Elmer Elan 5000), equipped with ultrasonic nebulization (Cetac Technologies, Model U-5000AT), which increased the sensitivity of the instrument by a factor of approximately 30, and greatly reduced oxide formations in the plasma stream that can lead to mass interferences during analysis [19]. Laboratory blanks and field blanks were used to evaluate potential contamination of the samples. Many samples were collected in duplicate and at multiple times on different dates for the evaluation of reproducibility of the trace element measurements.

2.2. Principal components analysis

There are many reviews of PCA (e.g., Refs. [18,20,21]) and the reader is referred to these for more detailed discussions. PCA is a multivariate statistical technique used to reduce the dimensionality of a data matrix or to state this more accurately: PCA is used to find the "true" dimensionality within a set of data. With PCA, a set of correlated measured variables is transformed into a set of uncorrelated principal components. The principal components are described using two lower dimensional data matrices (scores and loadings) that capture the underlying patterns within the original data [18]. The dominant patterns present within the samples and variables are illustrated by plotting the columns of the score matrix and the loadings matrix, respectively.

The principal components are obtained through eigenanalysis of the correlation or covariance matrix. The correlation matrix is generally used, as is the case in this study, so that each variable is mean centered with unit variance, thus providing equal weight of each of the variables. Eigenanalysis of a $p \times p$ correlation matrix produces p pairs of eigenvalues and eigenvectors. Each eigenvalue/eigenvector pair describes a principal component. The eigenvalues describe the amount of variance explained by each principal component and the loadings are the coordinates of the eigenvector. The principal component scores are then given as linear combinations of the original standardized (autoscaled) data with the loadings as the coefficients. Principal components are extracted so that the maximum amount of variance is explained in (has the largest eigenvalue associated with) the first principal component and progressively

Table 1

Sample location	Date	Li (ppb)	Ti (ppb)	V (ppb)	Cr (ppb)	Mn (ppb)	Ni (ppb)	Ge (ppb)	As (ppb)	Se (ppb)	Rb (ppb)	Sr (ppb)	Mo (ppb)	Sb (ppb)	Ba (ppb)	W (ppb)	U (ppb)	Ga (ppt)	Ru (ppt)	Tl (ppt)
Wells																				
Beatty-WD#1	2/97	93	1.15	10.3	1.42	0.15	0.12	0.32	13.9	1.78	22.5	148	10.5	0.21	5.15	0.70	8.50	3.1	1.5	108
Beatty-WD#2	4/97	103	1.87	6.96	0.67	0.09	0.18	0.26	15.1	0.92	23.1	158	9.0	0.18	5.48	0.76	8.90	3.5	1.5	98.2
Beatty-WD#2D	4/97	104	1.84	7.07	0.67	0.09	0.18	0.27	15.2	0.89	23.6	160	9.0	0.19	5.41	0.76	8.90	5.5	1.4	85.0
CofferRanch#1	9/94	166	0.47	2.67	0.80	0.34	0.10	1.55	6.4	0.57	18.1	163	12.2	0.19	9.80	0.80	15.4	2.7	3.3	74.0
CofferRanch#2	8/96	179	0.34	2.16	0.37	0.49	0.09	2.36	6.4	0.64	18.4	167	12.8	0.17	10.2	0.81	15.5	2.5	3.0	136
ER-20-6-3	6/97	52	2.78	2.20	0.93	26.6	4.76	0.51	5.8	0.45	11.4	23.7	3.4	0.18	2.89	1.27	2.98	227	3.1	11.0
ER-20-6-3D	6/97	51	3.04	2.00	0.87	26.1	4.33	0.50	5.5	0.51	11.1	23.0	3.3	0.18	2.85	1.26	3.08	216	2.0	10.7
ER-30	2/95	63	0.16	4.93	2.43	17.0	0.67	0.19	7.8	0.57	3.7	6.7	3.0	0.34	1.18	0.66	1.80	170	0.0	23.6
ER-OV1	11/97	175	0.42	1.78	1.47	0.54	1.07	1.12	6.3	0.82	25.3	4.7	9.5	0.16	2.04	0.88	8.50	21	3.0	43.3
ER-OV1D	11/97	178	0.45	1.77	1.48	0.51	1.04	1.13	6.5	0.82	25.4	4.6	9.6	0.16	1.97	0.87	8.32	20	3.1	41.3
ER-OV2	11/97	192	0.73	1.54	1.38	0.99	1.91	1.33	5.2	0.79	12.3	44.5	9.7	0.20	3.20	0.64	18.5	7.7	2.0	22.5
ER-OV3a	11/97	146	0.54	2.71	1.64	0.73	1.58	1.01	8.2	0.84	11.8	76.1	7.5	0.21	4.44	1.45	7.76	19	2.3	20.1
ER-OV3aD	11/97	147	0.52	3.20	1.64	0.77	1.21	1.02	9.2	0.92	11.7	74.4	7.6	0.20	5.7	1.64	7.94	12	2.1	26.3
ER-OV3a3	11/97	143	0.45	2.70	1.58	0.70	1.42	1.03	7.9	0.82	11.7	75.5	8.1	0.30	4.38	0.73	7.95	22	2.5	19.2
ER-OV3c	11/97	119	0.46	2.69	0.42	0.47	0.64	1.23	15.3	0.49	5.0	97.3	12.7	0.32	1.63	1.78	4.01	33	3.5	11.9
ER-OV3cD	11/97	123	0.40	2.42	0.39	0.45	0.69	1.24	15.4	0.41	4.8	97.8	12.8	0.32	1.65	1.84	4.10	30	2.7	11.7
ER-OV4a	11/97	127	0.36	3.74	1.70	0.26	0.46	1.19	14.2	0.59	18.9	21.7	8.9	0.31	2.22	1.72	2.69	12	3.0	13.1
ER-OV4aD	11/97	129	0.34	3.72	1.72	0.23	0.50	1.25	14.6	0.52	19.2	21.8	9.1	0.31	2.21	1.71	2.66	11	3.0	13.0
ER-OV5	11/97	72	0.32	14.5	2.59	1.68	5.33	0.58	16.2	1.10	22.2	191	5.9	0.11	9.15	1.70	2.04	2.2	2.0	17.2
ER-OV6a	11/97	167	0.50	7.97	0.99	2.39	0.43	0.83	12.9	0.81	11.6	10.7	10.6	0.27	0.49	0.94	4.80	54	3.0	21.3
Sdale	11/97	97	0.44	9.65	1.41	0.08	0.20	0.78	13.7	0.89	19.4	277	9.3	0.08	21.1	0.95	2.66	4.0	1.3	17
Tolicha Pk.#1	11/96	51	0.36	4.30	1.14	23.9	0.19	0.31	16.3	1.18	17.2	122	4.5	0.12	3.38	0.47	2.51	8.9	0.7	49.5
Tolicha Pk.#2	11/97	50	0.23	4.15	0.93	0.57	0.15	0.32	17.0	1.06	15.0	111	4.5	0.11	1.92	0.52	2.61	10	1.4	16.4
Tolicha Pk.#2D	11/97	51	0.26	4.10	0.89	0.49	0.14	0.32	16.9	0.91	14.5	109	4.5	0.12	1.82	0.51	2.67	10	2.1	15.1
U-20WW	11/97	62	0.23	2.42	0.25	49.9	3.43	0.42	5.9	0.50	4.9	24.3	3.2	0.14	0.08	1.15	2.15	41	1.7	8.9
U-20WWD	11/97	63	0.23	2.49	0.26	49.6	3.43	0.42	5.9	0.51	4.9	24.2	3.2	0.15	0.08	1.15	2.11	43	1.9	9.0
Uc-29a1	11/97	38	0.13	3.43	1.39	0.56	0.51	0.33	5.8	0.89	7.7	51.5	1.9	0.14	2.05	0.63	1.32	5.1	1.1	16.2

Uc-29a1D	11/97	40	0.12	3.40	1.35	0.54	0.49	0.33	5.9	0.92	7.5	52.4	1.9	0.13	2.02	0.62	1.34	4.3	1.0	15.6
Uc-29a2	11/97	41	0.12	0.62	0.23	4.90	0.26	0.13	2.0	0.49	5.6	28.6	2.1	0.06	0.12	0.56	0.54	3.3	0.8	9.0
Uc-29a2D	11/97	38	0.11	0.64	0.27	4.94	0.28	0.13	2.2	0.49	5.6	29.9	2.1	0.06	0.13	0.59	0.56	3.5	1.0	8.2
Coffer-WM#1	9/94	111	0.63	0.99	0.26	11.8	0.60	0.78	7.8	0.40	4.7	183	11.2	0.23	1.82	1.53	5.08	18	2.2	46.0
Coffer-WM#2	9/96	104	0.14	0.31	0.08	23.5	0.48	0.25	3.9	0.43	3.8	157	10.4	0.14	1.29	0.98	1.66	7.9	1.1	35.2
Coffer-WM#2D	9/96	104	0.14	0.19	0.12	22.4	0.48	0.25	2.4	0.43	3.8	159	10.6	0.15	1.13	1.02	1.78	7.4	0.9	36.3
Coffer-WM#3	2/97	107	0.34	0.20	0.23	15.6	0.41	0.14	2.8	0.22	4.6	137	10.5	0.10	1.08	0.88	1.12	11	2.1	35.8
Coffer-WM#3D	2/97	103	0.36	0.21	0.21	15.0	0.44	0.15	2.8	0.24	4.6	133	10.6	0.10	1.03	0.83	1.10	12	1.6	41.1
Coffer-WM#4	11/97	120	0.30	1.11	0.13	8.25	0.41	0.89	8.4	0.53	4.8	181	11.9	0.23	1.61	1.48	5.86	15	2.5	10.0
WW-8	11/97	27	0.10	1.03	0.18	1.17	0.30	0.17	1.8	0.70	6.7	4.9	0.9	0.05	0.09	0.21	0.36	15	1.3	12.9
WW-8D	11/97	29	0.12	1.03	0.17	1.11	0.28	0.18	1.7	0.74	6.7	4.9	0.9	0.05	0.09	0.21	0.35	14	1.0	11.7
Springs																				
1Baileys	11/97	235	0.61	2.09	0.41	0.17	0.16	1.76	18.4	0.29	28.8	117	12.2	0.4	18.3	2.74	9.84	75	2.8	84.6
1BaileysD	11/97	235	0.60	2.11	0.42	0.16	0.15	1.79	18.5	0.32	28.6	117	12.3	0.39	18.7	2.77	9.83	75	2.2	82.2
2Baileys#1	9/96	234	0.46	2.12	0.48	1.36	0.12	1.74	19.7	0.84	28.6	121	12.2	0.41	16.7	2.82	8.53	68	2.4	132
2Baileys#2	4/97	239	1.78	2.07	0.39	0.06	0.08	1.79	19.8	0.47	29.3	119	12.2	0.43	17.6	2.88	9.21	76	2.1	159
3Baileys#1	9/96	220	0.30	2.06	0.34	1.18	0.12	1.66	18.8	0.74	26.3	110	11.6	0.39	16.0	2.84	8.34	69	1.9	140
3Baileys#2	4/97	240	1.56	2.04	0.39	0.07	0.08	1.82	20.1	0.42	28.8	115	12.2	0.40	16.4	2.92	8.63	75	1.4	156
Fran's#1	11/97	129	0.26	1.72	0.30	0.05	0.03	1.17	23.1	0.60	11.6	21.0	9.2	0.50	0.48	1.63	5.55	83	2.3	48.1
Fran's#2	8/96	126	0.61	2.28	0.58	0.34	0.07	1.28	26.6	1.02	13.1	22.5	10.0	0.56	0.49	1.57	5.54	100	1.7	51.9
Fran's#3	2/97	128	0.45	1.81	0.36	0.29	0.25	1.14	23.7	0.53	12.2	21.1	9.2	0.53	0.53	1.60	5.97	83	2.4	18.5
Goss	11/97	143	0.70	2.38	1.33	0.05	0.13	1.07	7.5	0.50	12.4	85.3	8.1	0.19	4.97	0.52	9.28	37	2.7	19.4
GossD	11/97	146	0.59	2.26	1.32	0.05	0.16	1.06	7.2	0.54	12.5	87.5	8.2	0.19	4.89	0.52	9.53	35	2.9	21.6
Mullen	11/97	147	0.40	2.26	1.18	0.37	0.07	0.98	7.2	0.49	13.1	82.2	8.1	0.18	4.38	0.53	9.49	33	2.6	25.1
Tippipah	12/94	12	4.10	1.40	0.09	1.39	0.09	0.04	2.0	0.56	7.1	6.2	0.7	1.94	0.34	0.01	0.52	76	0.6	42.9

less variance is explained for each subsequent component.

Principal components analysis of the simulated and measured trace element data were performed using S-PLUS [22] and MINITAB [23], respectively.

2.3. Monte Carlo data simulation and single detection limit censoring

Three parent data matrices ($\mathbf{X1}_{51,19}$, $\mathbf{X2}_{21,14}$, and $\mathbf{X3}_{24,15}$) were used in this study. Identical data simulation methodology was applied to each of the three matrices. The data matrix, **X1**, consisted of 19 trace element measurements made on a total of 51 samples (see Table 1). Some of the sites, listed in Table 1, were sampled multiple times either on the same day (duplicates) or over a period of time (repeat samples). Duplicate samples are identified with a "D" following the sampling location name and repeat samples are sequentially numbered. Details of these data, including sampling locations, are reported elsewhere [6]. Although the concentration of a total of 58 trace elements was measured in each groundwater sample, only 19 were included in this study. The 19 trace elements were selected based on 50% agreement between duplicate samples. The 50% criteria were used to eliminate the elements with a greater uncertainty associated with their measurement. In most cases, the concentrations of the trace elements measured in the field blanks were less than 1% of the measured concentrations in the field samples.

Using the Kolmogorov–Smirnov Normality Test, the concentrations of the 19 trace elements in the groundwater samples were found to be neither multivariate normal nor multivariate lognormal. The p values for 15 of the variables and 8 of the log transformed variables were less than 0.01. Only three variables and six log-transformed variables showed p values that exceeded 0.15. Instead, subpopulations within the data were observed [6]. The data distribution within each of these subpopulations can then be modeled by a multivariate normal distribution. Data for this study were therefore generated using this mixture multivariate normal model.

Six clusters, within this data set, were identified using Hierarchical cluster analysis [24]. The program S-PLUS was then used for Monte Carlo simulation of the data contained in each cluster (i) as follows: First, Choleski Decomposition of the covariance matrix ($\mathbf{\Sigma}_i$) was performed to find $\boldsymbol{a}_i$ such that $\mathbf{\Sigma}_i = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_i$. An independent normal data matrix, $\mathbf{Z}_i$, was then generated. Mean centered data with a covariance structure similar to that of the original data were then calculated as $\mathbf{Z}_i\boldsymbol{a}_i^{\mathrm{T}}$. The mean was adjusted to that of the original data by adding the row vector of column means (mean of each variable in the original data contained in the cluster) to each row of the matrix $\mathbf{Z}_i\boldsymbol{a}_i^{\mathrm{T}}$. Finally, the data matrices for each of the six clusters were combined to produce the full data set.

Detection limits (DL) were then applied so that a given percentage (5%, 10%, 15%, 20%, 30%, 40%, 50%, 75%) of values within each column were below this value. Values <DL were substituted with 0, DL, and DL/2. PCA was then applied to the unsubstituted and each of the three substituted data matrices This process was repeated 100 times, generating 100 different independent normal data matrices, $\mathbf{Z}_i$, for each of the clusters. The differences between the PC1 and PC2 scores and loadings for the substituted data matrices and the unsubstituted data matrix were calculated and averaged.

The complete process for **X1** is described in the algorithm below. The other two data matrices (**X2** and **X3**) were treated the same as **X1** with the exception that four clusters were observed for **X2** ($n_1 = 3$, $n_2 = 3$, $n_3 = 4$, and $n_4 = 11$) and three clusters were observed for **X3** ($n_1 = 11$, $n_2 = 4$, $n_3 = 9$).

1. Input six data matrices ($\boldsymbol{X}_i$) containing n_i cases (samples) and $p = 19$ columns (variables), where $n_1 = 6$, $n_2 = 11$, $n_3 = 5$, $n_4 = 14$, $n_5 = 5$ (giving a total of 51 samples).
2. Input a 51×19 matrix of cluster means, which combines six data matrices $\mathbf{M}_i$ ($i = 1, 2, \ldots, 6$). Each row of the matrix $\mathbf{M}_i$ ($n_i \times 19$) is represented by the vector of column means for the corresponding cluster. For instance, rows 1 to n_1 consist of the column means for cluster 1, rows $(n_1 + 1)$ to $(n_1 + n_2)$ consist of the column means for cluster 2, etc.
3. For each cluster i, $i = 1, 2, \ldots, 6$:
   - 3.1. Read the data matrix $\mathbf{X}_i$;
   - 3.2. Compute $\mathbf{S}_i$ = covariance of matrix $\mathbf{X}_i$;
   - 3.3. Compute the Choleski decomposition of $\mathbf{S}_i$: $\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i = \mathbf{S}_i$;

3.4. Generate the data matrix $\mathbf{Z}_i^T$ $(19 \times n_i)$ from independent multivariate normal (MVN) distribution;
3.5. Generate MVN data for cluster i: $\mathbf{Y}_i^T$ $(19 \times n_i) = \mathbf{a}_i^T \mathbf{Z}_i + \mathbf{M}_i^T$;
3.6. Compute $\mathbf{Y}_i$ = transpose $(\mathbf{Y}_i^T)$ to obtain n_i observations from cluster i.

4. Combine the above six data matrices $\mathbf{Y}_i$ into one 51×19 data matrix $\mathbf{X}_{sim}$.
5. Compute (for $q = 5, 10, 20, 30, 40, 50, 75$) the 1×19 quantile vector $\mathbf{Q}$ such that q% of data in column j (of $\mathbf{X}_{sim}$) is less than or equal to the j-th element of $\mathbf{Q}$. Let the detection limit for the variable, j, be $DL_j = Q_j$.
6. If, for any sample ($k = 1, 2, \ldots, 51$) within $\mathbf{X}_{sim}$, the value in the column $j < \mathbf{Q}_j$, then replace the value by one of the three substitution values: 0, DL, DL/2. In the S-Plus program, the resulting matrices are called xsim0, xsimdl, xsimdlby2.
7. Perform PCA (using the correlation matrix) on the unsubstituted data matrix ($\mathbf{X}_{sim}$) and the three substituted matrices (xsim0, xsimdl, xsimdlby2).
8. Compute the various PC-objects (PC-loadings and PC-scores for the first two PCs).
9. Compute the distances between the PCA objects from xsim0, xsimdl, xsimdlby2 and those obtained from PCA of $\mathbf{X}_{sim}$. Also, compute the percentages of variance explained by the first two principal components.
10. Repeat (steps 3–9) 100 times and average the PCA results of step 9 over the 100 iterations.

2.4. Simulation of multiple detection limits

The difficulty with dealing with values below the detection limit is further complicated when multiple detection limits are reported for a given variable (i.e., trace element). Multiple detection limits were therefore applied to **X1**, **X2**, and **X3**. Data simulation was performed as described for the single detection limit case except that Step 5 is changed to the following:

(5) Compute (for $q = 10, 20, 30, 40, 50, 75$) the 1×19 quantile vector $\mathbf{Q}$ such that q% of data in column j (of $\mathbf{X}_{sim}$) is less than or equal to the j-th element of $\mathbf{Q}$. Let the maximum detection limit for the variable j be maximum $DL_j = \mathbf{Q}_j$ and generate the multiple detection limits from the uniform distribution on the interval $(\mathbf{Q}_{j,5}, \mathbf{Q}_{j,q})$.

2.5. Simulation for variable selection

Selection of variables used in the data set included in this study was based on 50% difference between duplicate samples. Out of the 55 trace elements measured for each of the 51 samples, only 19 were retained for use in the statistical evaluations [6]. Although 50% difference between duplicate samples appears quite significant, in some cases, the only difference is between 1.0 and 1.5 ppt. This difference then becomes insignificant when compared to the 100-ppt levels that may exist in samples collected from other locations. These criteria for variable selection may therefore result in the elimination of important variables or also in retaining variables that provide no information toward discrimination between waters from different locations. Rather than selecting variables based solely on agreement between duplicates, variance between samples from different locations must also be considered. This section describes the experiment used to select variables based on the percentage of variance between duplicate samples relative to the variance observed for samples over all sites.

(1) Replace all duplicate sample values by normally distributed values so that they are within ± 20% of the original sample. This gives the 'true' data matrix $\mathbf{X}_{simtrue}$. Perform PCA on $\mathbf{X}_{simtrue}$ and compute the various PC-objects (PC-loadings, PC-scores, and percentages of variance explained for the first two principal components).

(2) Randomly select 10 variables and replace the duplicate measurements for these 10 variables with original values $+ q \times$ normally distributed error with mean 0 and sd $= q \times$ sd of the variable across all the sites ($q = 0.1, 0.2, 0.3, 0.4, 0.5, 0.75$). This gives the perturbed data matrix $\mathbf{X}_{simdup}$. Perform PCA on $\mathbf{X}_{simdup}$, and compute the various PC-objects (PC-loadings, PC-scores, and percentages variance explained).

(3) Compute distance between the PCA objects obtained from the two data matrices $\mathbf{X}_{simtrue}$ and $\mathbf{X}_{simdup}$. The PCA objects chosen in this experiment are: (1) difference of PC1 and PC2 scores between the

only duplicate measurements, (2) the same differences for all of the cases, (3) the PC loadings, and (4) the percentages variance explained.

3. Results and discussion

PCA is a useful tool when searching for similarities between samples (represented by PC scores) and also between variables (represented by PC loadings). The impact of the substitution methods on both the PC scores and loadings was therefore evaluated. Because of the large number of graphs prepared for the simulations of the three data matrices (**X1**, **X2**, and **X3**), and because similar results were observed, only the graphs generated for **X1** are presented here.

3.1. Single detection limit censoring

A comparison between the scores resulting from PCA on the substituted and unsubstituted data matrices generated from **X1** is shown in Fig. 1. The difference between the PC1 scores resulting from the unsubstituted data matrix (full matrix) and those of the substituted data matrices is plotted on the y-axis and the percentage of values below the detection limit (nondetects) is plotted on the x-axis (Fig. 1a). Fig. 1a shows that substitution with either 0 (SM_0) or DL/2 (SM_DL/2) is superior to substitution with DL (SM_DL). Fig. 1a also shows a rise in slope after the percentage of nondetects is increased approximately above 30%. Similar results are observed for the PC2 scores (Fig. 1b); substitution with 0 and DL/2 are superior to substitution with DL. An identical pattern is observed for the **X1** PC loadings (Fig. 2a,b). The variance explained for the first two principal components are shown in Fig. 3. The variance explained for the data matrix substituted with DL/2 is quite similar to that of the full matrix when the percentage of nondetects is less than 30% which is not the case for those substituted with DL or 0.

The results for the data matrices **X2** and **X3** were quite similar to those for **X1**. The distances from the unsubstituted matrix for both the PC scores and loadings were generally lower when substituted with DL/2. However, the distances were slightly lower for **X2** when there were less than 10% nondetects present

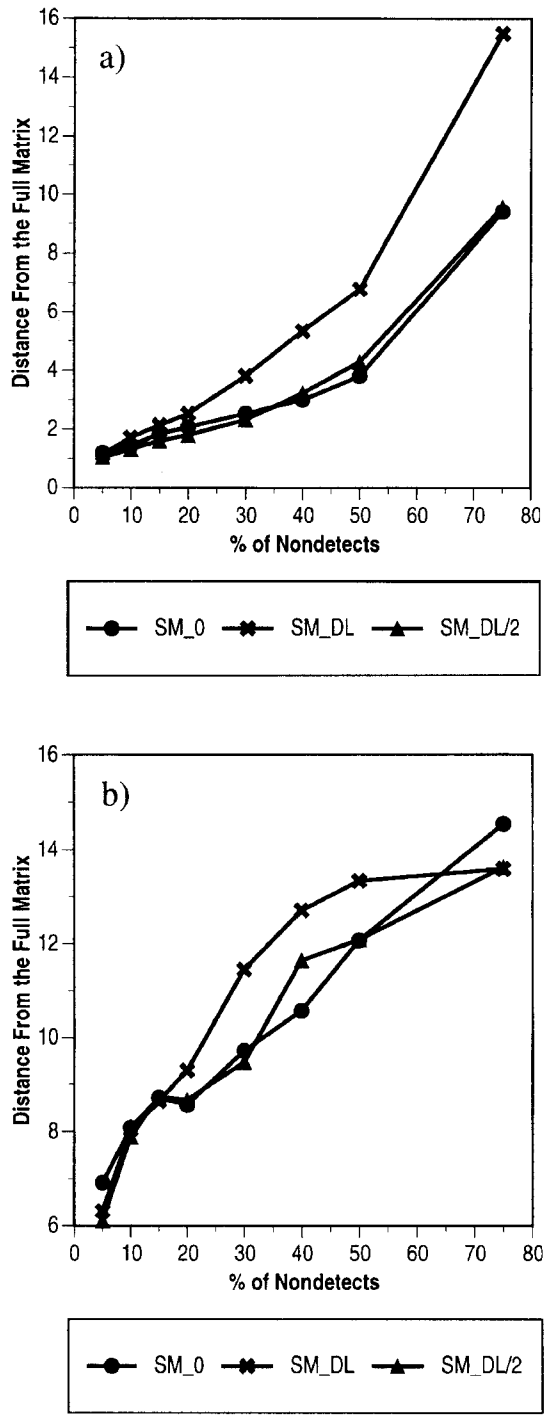

Fig. 1. Comparison of principal component scores for the three substitution methods using a single detection limit. (a) PC1 and (b) PC2.

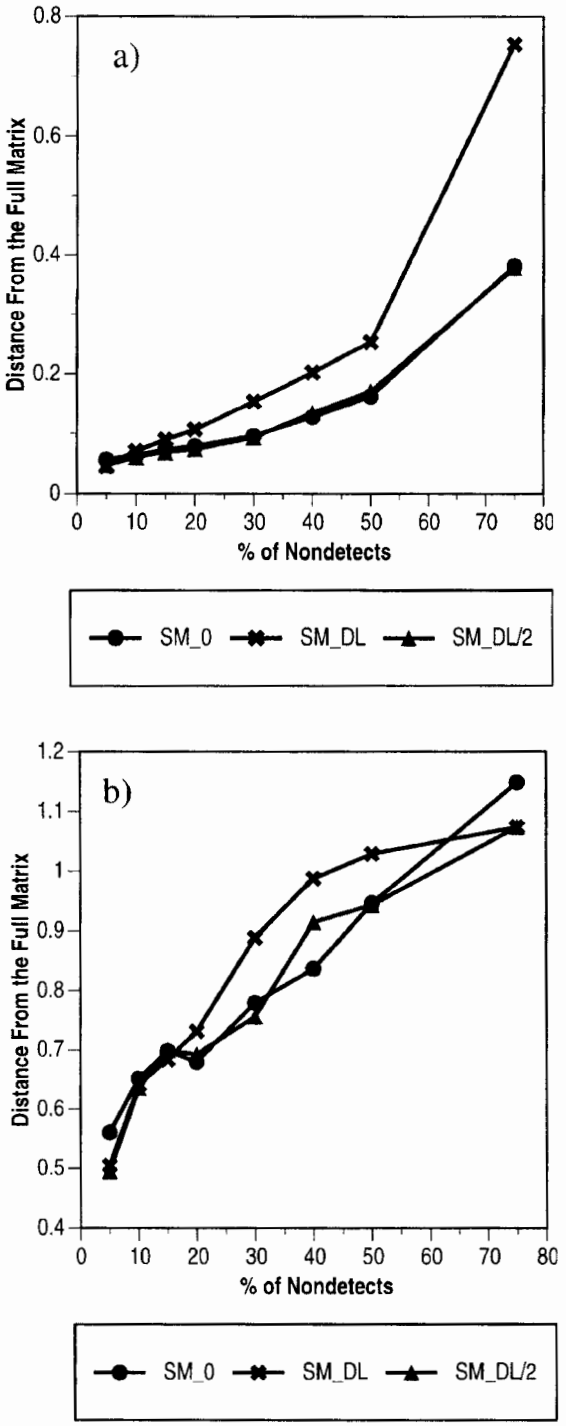

Fig. 2. Comparison of principal component loadings for the three substitution methods using a single detection limit. (a) PC1 and (b) PC2.

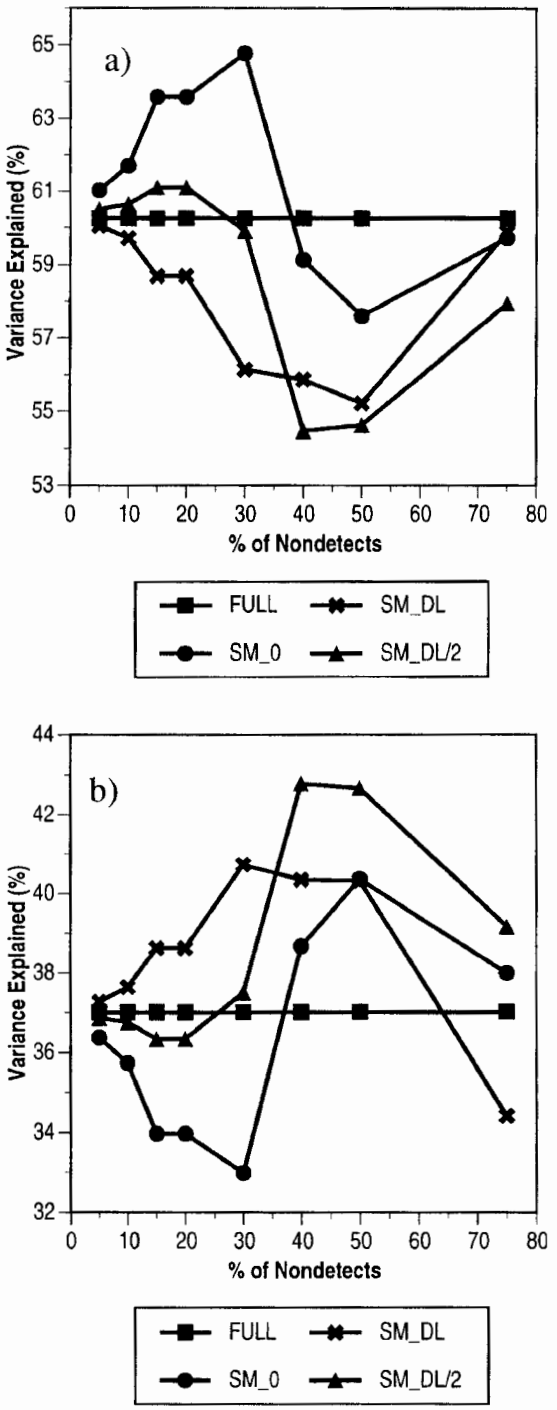

Fig. 3. Comparison of percentage variance explained for the three substitution methods using a single detection limit. (a) PC1 and (b) PC2.

and the DL substitution method was used. The impact of substituting the nondetects with DL/2 increased when the percentage of nondetects exceeded approximately 25%, in most cases. DL/2 performed the best in terms of percentage variance explained for **X2**. For **X3**, the DL substitution method appeared to be better than substitution with DL/2 when comparing percent variance explained by PC1, and the two substitution methods, DL and DL/2, were very similar for PC2. Substitutions with DL and DL/2 were superior to substitution with 0.

3.2. Simulation of multiple detection limits

The difference in PC scores between the substituted and unsubstituted (full matrix) data matrices, for **X1**, are shown in Fig. 4a,b. The x-axis for this multiple detection limit study is the maximum percentage of nondetects and corresponds to the maximum detection limit. The results are similar for substitution with 0 and DL/2 (PC1), and DL and DL/2 (PC2). Poor performance of all substitution methods was observed when the percentage of nondetects exceeded approximately 35%. The results for the **X1** PC1 loadings (Fig. 5a) were quite similar for all three substitution methods. The change in slope, in this case, occurred after the percentage of nondetects increased beyond 20% (Fig. 5a). For PC2, the two substitution methods, DL and DL/2, were superior to substitution with 0 (Fig. 5b). This was also observed with the percentage of variance explained (Fig. 6a,b). For both PC1 and PC2, the performance of all substitution methods deteriorated rapidly, when compared to the full data matrix, after the percentage of nondetects exceeded 30% (Fig. 6a,b). Comparing scores and loading for PC1 of **X2**, substitution by DL gave better results than substitution by DL/2 in the range of 10–18% nondetects, whereas substitution by DL/2 was better in the range of 18–30%. The picture changed considerably when comparing the explained percentage variance and also the scores and loadings for PC2; substitution by DL/2 was clearly the best method.

The use of DL/2 was further examined using a data set containing multiple measurements of the trace element chemistry of groundwater collected from several springs in Death Valley, CA. These samples were collected in November 1993, March 1994, and

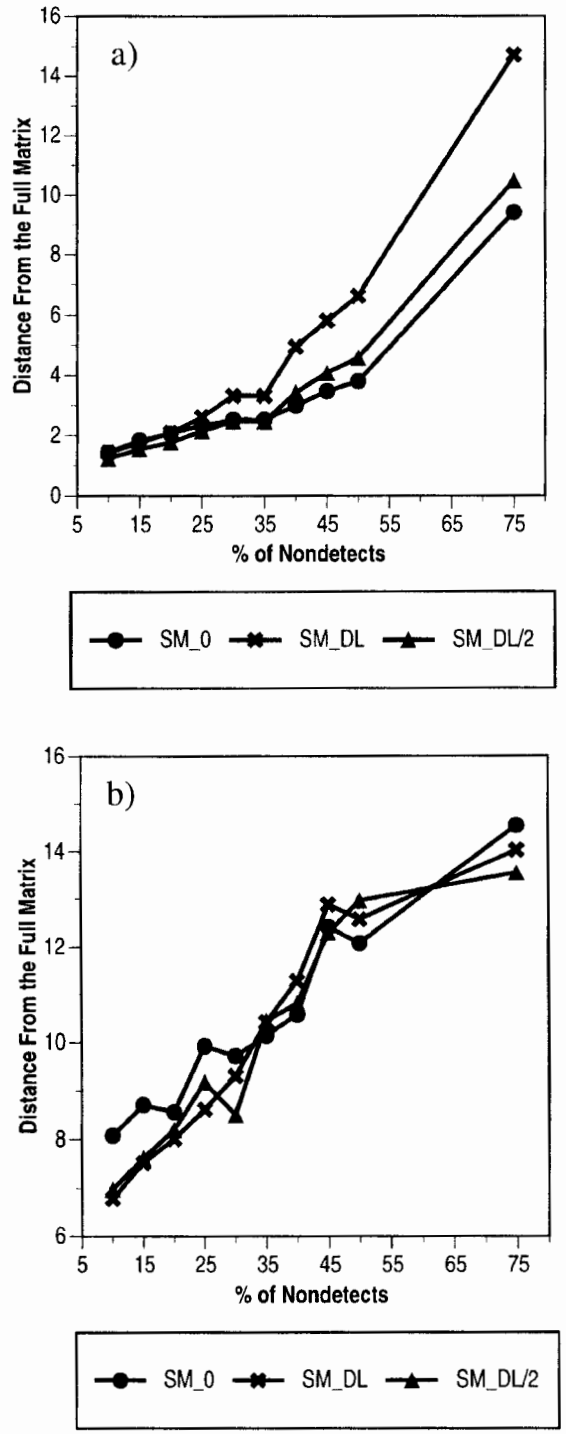

Fig. 4. Comparison of principal component scores for the three substitution methods using multiple detection limits. (a) PC1 and (b) PC2.

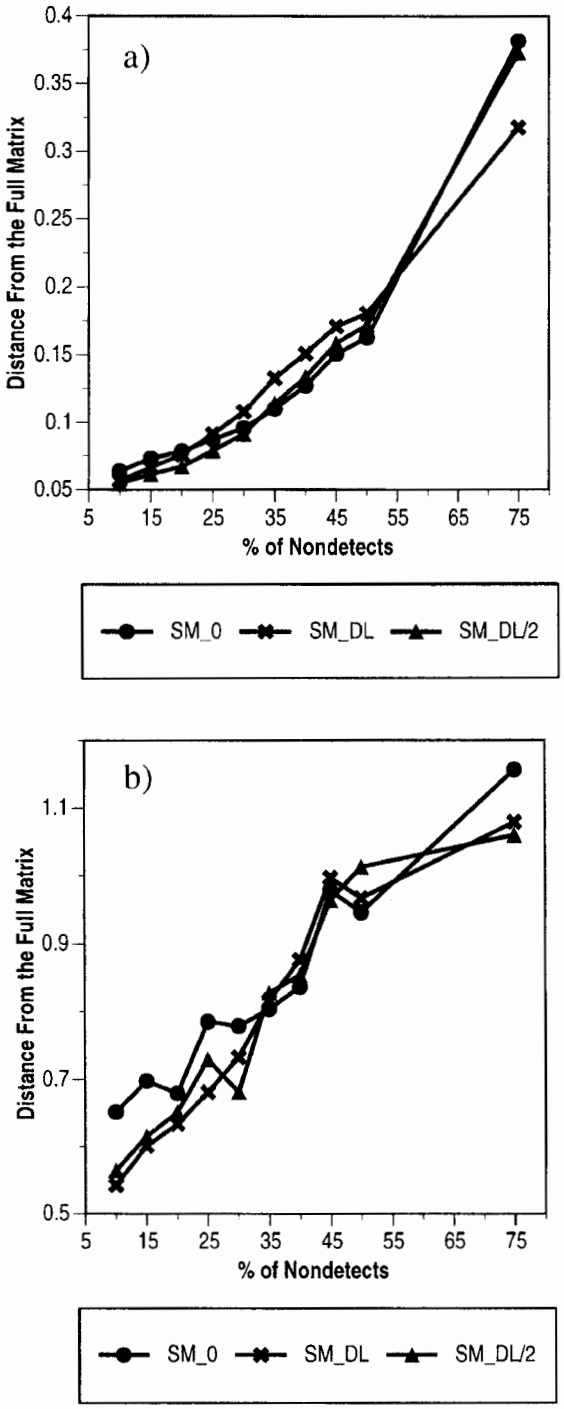

Fig. 5. Comparison of principal component loadings for the three substitution methods using multiple detection limits. (a) PC1 and (b) PC2.

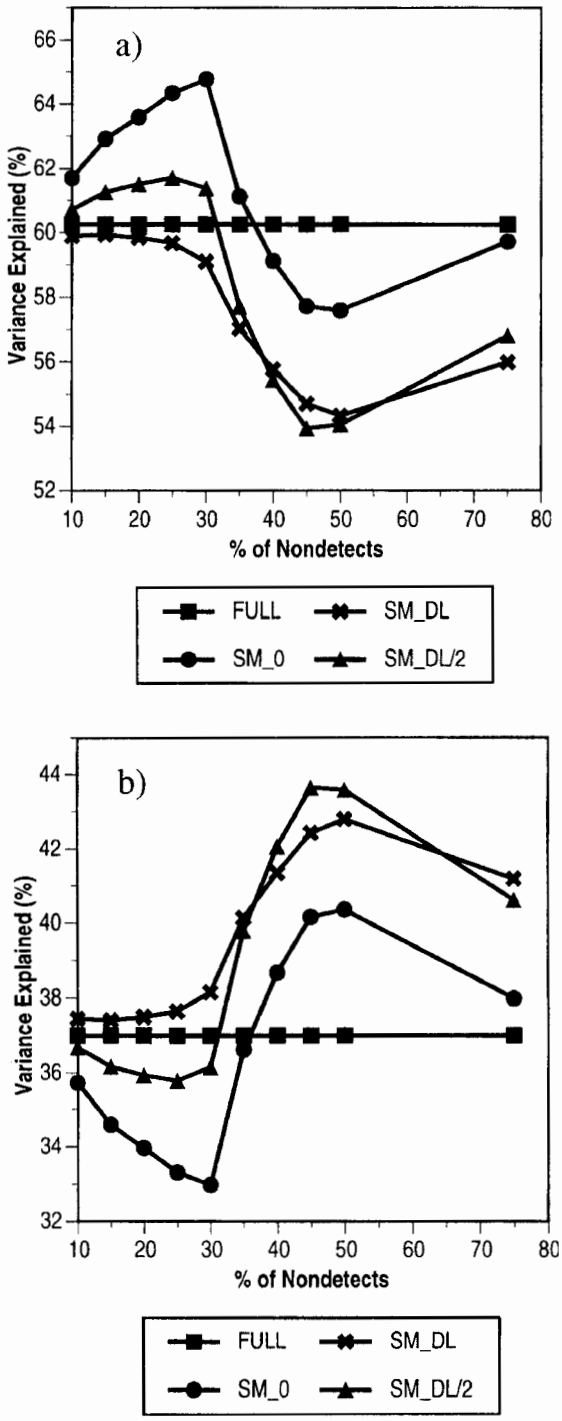

Fig. 6. Comparison of percentage variance explained for the three substitution methods using multiple detection limits. (a) PC1 and (b) PC2.

November 1994. In order to make an accurate comparison between waters from different locations, it must be first shown that waters from a single location can be measured reproducibly. Approximately 20% of these data were listed as '<DL' and were substituted with DL/2. A higher detection limit was associated with the third sample set. A plot of the scores for the first two principal components is shown in Fig. 7. This figure clearly illustrates the excellent reproducibility of the multiple measurements for each site. With the exception of Tex/TrvA and Scot/Surp, distinct clusters were formed that contain each of the multiple measurements and corresponding average for a given spring. The waters from Texas and TrvA cannot be distinguished from each other using these PCA results because the chemical composition of these springs is essentially identical. The same can be said for Scotty's and Surprise Springs. [1]. These results show that substitution of '<DL' values with DL/2 gives highly reproducible PCA results while maintaining the ability to distinguish waters from different sources even in the presence of multiple detection limits.

3.3. Simulation for variable selection

In general, the performance of the substitution methods deteriorated when the percentage of '<DL' values exceeded 30%. Depending on the trace element data set that is evaluated, the percentage of values below the detection limit may exceed this amount. For these cases, it may be preferred to use the uncensored data when available. Use of uncensored data can often be useful for evaluation of the trace element chemistry when the waters from different locations vary substantially. For instance, the concentration of certain trace elements in the groundwater from one region may be quite low (i.e., <DL) but may be substantially higher in another region (≫DL). These trace elements would therefore be quite useful in distinguishing groundwaters from these different regions. On the other hand, the concentrations of these trace elements may be quite low in the entire study area and including them in the evaluations may just add substantial noise to the analysis. Determination of criteria for the selection of

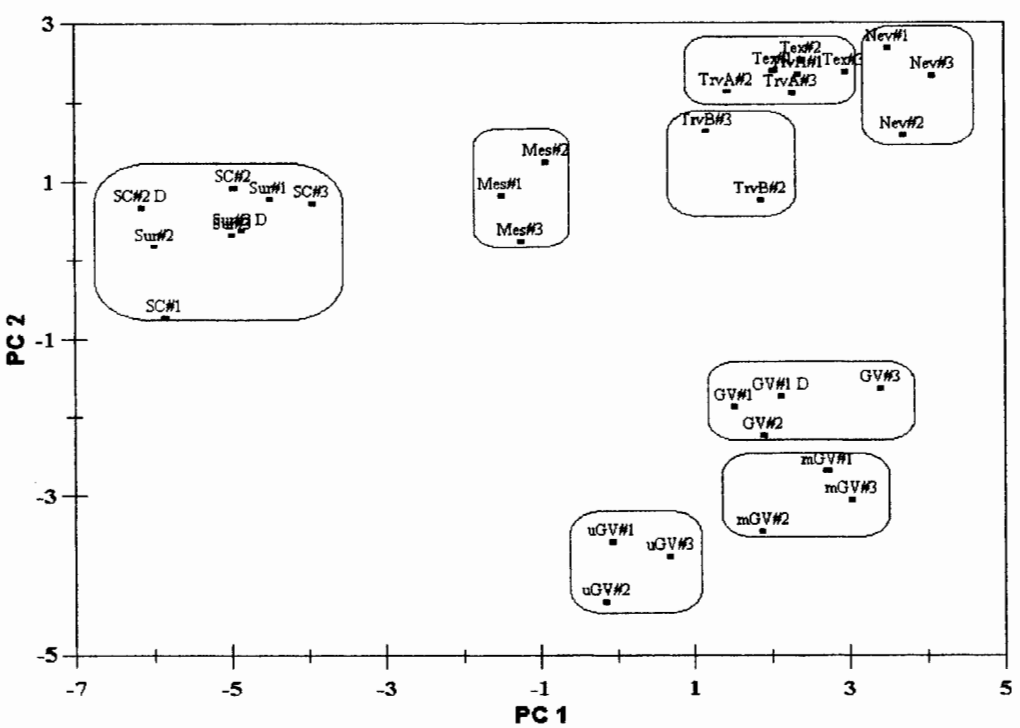

Fig. 7. PCA score plot for Death Valley trace element data (values<DL were substituted with DL/2).

variables based on variance between duplicates relative to that observed between all samples (percentage variability between duplicates) is therefore necessary to the success of our investigation. To this end, data sets containing duplicate samples were simulated based on 20% acceptability criteria between duplicate measurements (true data sets). A specified amount of variance, relative to the variance over all samples, was then added to 10 randomly selected variables within each of the duplicate samples. The results from PCA on the **X1** data set are shown in Fig. 8 (scores), Fig. 9 (loadings), and Fig. 10 (percentage variance explained). The distance between the scores for the true data set and the data set with additional variability added to the duplicate samples are shown in Fig. 8a for all samples and Fig. 8b when only the duplicate samples are compared. This distance increased linearly when comparing the PC scores for all samples (Fig. 8a) and also the PC loadings (Fig. 9). When considering the scores of the duplicate samples only (Fig. 8b), the distances increased exponentially once the variability between duplicates, relative to the total variability, exceeded 20%. The variance explained decreased exponentially, with no clear cutoff value, when the percent variability between duplicates was increased (Fig. 10).

For data sets **X2** and **X3**, the results were slightly different. The impact of the variability between duplicates increased when the differences between duplicates, relative to the total variability, were greater than 30%. For data set **X2**, the differences in PC scores, when all samples are compared, and the PC loadings increased rapidly for both PC1 and PC2 when the percent relative variability exceeded 30%. Similar results were observed for data set **X3**, with the exception that the differences in PC1 scores for the duplicates only, and the loadings showed a linear trend. A somewhat linear increase in distances was observed for both **X2** and **X3**, when the PC scores between the duplicate samples were compared. The variance explained decreased exponentially again; no apparent cutoff percentages were observed.

A data set containing trace element concentrations in groundwater samples collected from a group of wells in Nye County, NV was then used to test the effectiveness of selection of variables based on relative variability between duplicate samples. The con-

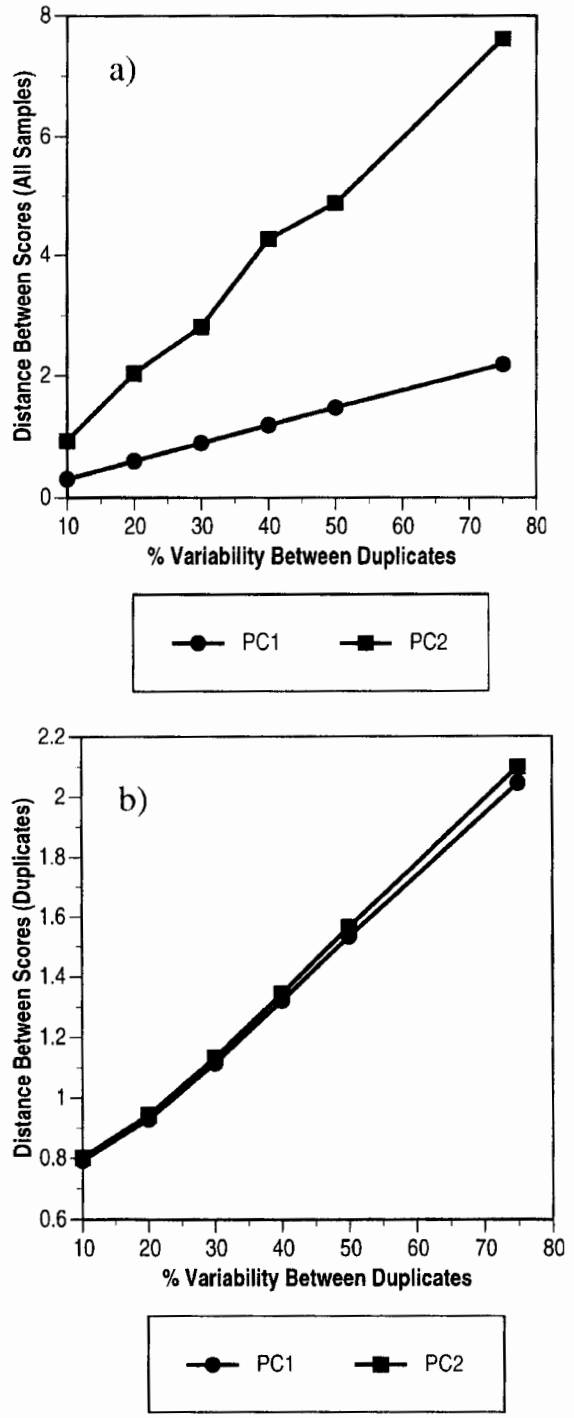

Fig. 8. Comparison of PC scores for data sets with added variability between duplicate samples. (a) Distance between scores for all samples. (b) Distance between scores for duplicate samples only.

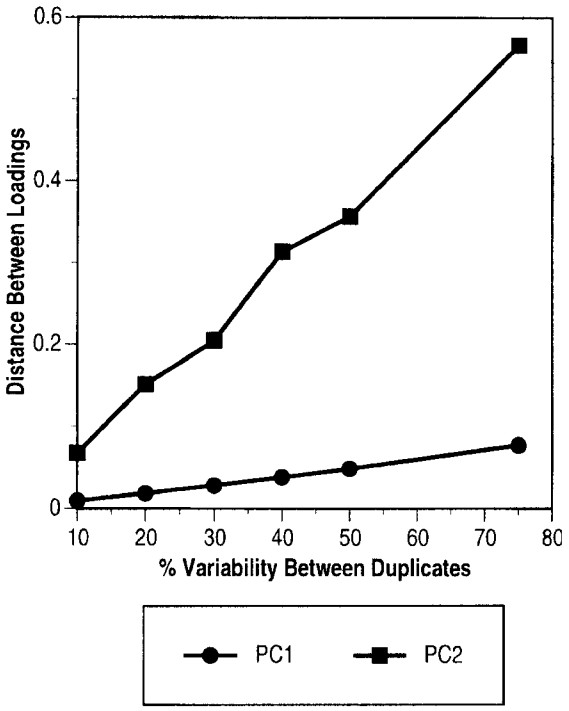

Fig. 9. Comparison of PC loadings for data sets with added variability between duplicate samples.

centrations of a total of 58 trace elements were measured for samples collected from six wells: NC-EWDP-1S (1S), NC-EWDP-3S (3S), NC-EWDP-9S (9S), NC-EWDP-1D (1D), Bond Gold Well #13 (BGW) and SD6ST1. Multiple zones within three of the wells (1S, 3S, and 9S) were sampled on two separate occasions (May and November 1999). May samples are identified as "#1" and November samples as "#2". A total of six samples (3SZn3, 9SZn1, 9SZn4, 1D, SD6ST, and BGW) were collected in duplicate during the May sampling. PCA was applied to the full data set and a total of 56% of the variance was explained in the first two principal components. The scores for the first two components are plotted in Fig. 11. Because of the similarities that exist in trace element chemistry of these groundwaters and the relatively high variability in some of the measurements, clear clustering of the different wells is not observed (Fig. 11). This is primarily seen in the overlap in clustering of the 9S, 1S, and SD6ST wells. PCA was then applied to the data set after eliminating all variables with a variance between duplicates that

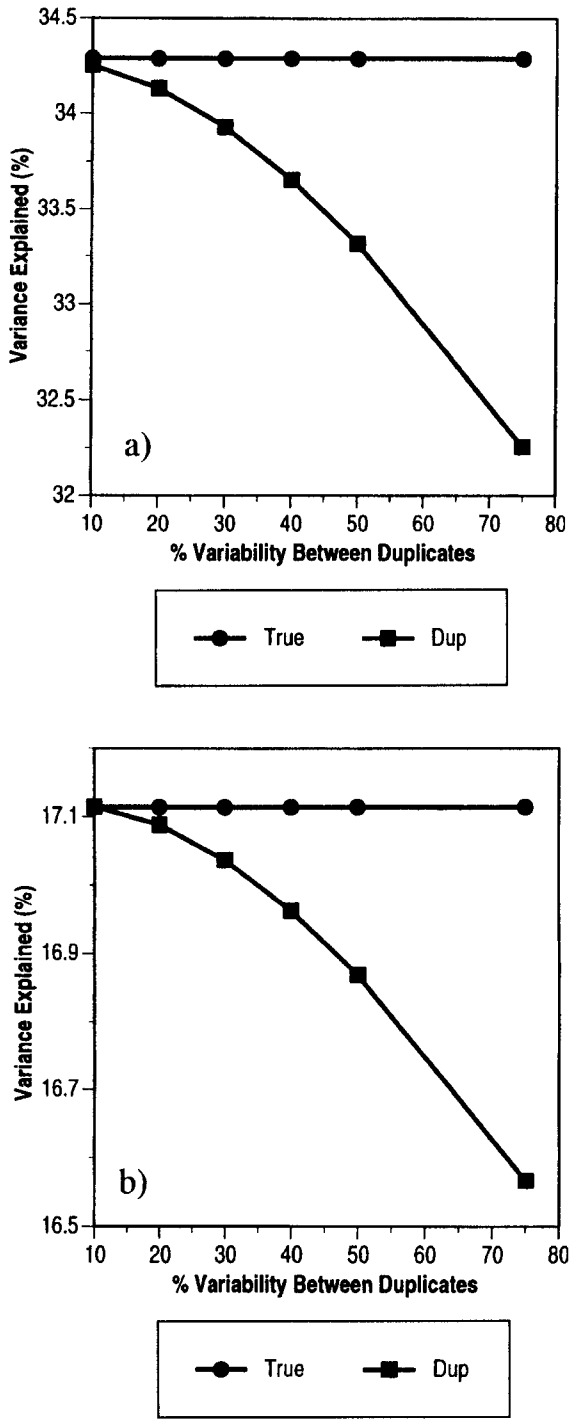

Fig. 10. Comparison of percentage variance explained for data sets with added variability between duplicate samples. (a) PC1 and (b) PC2.

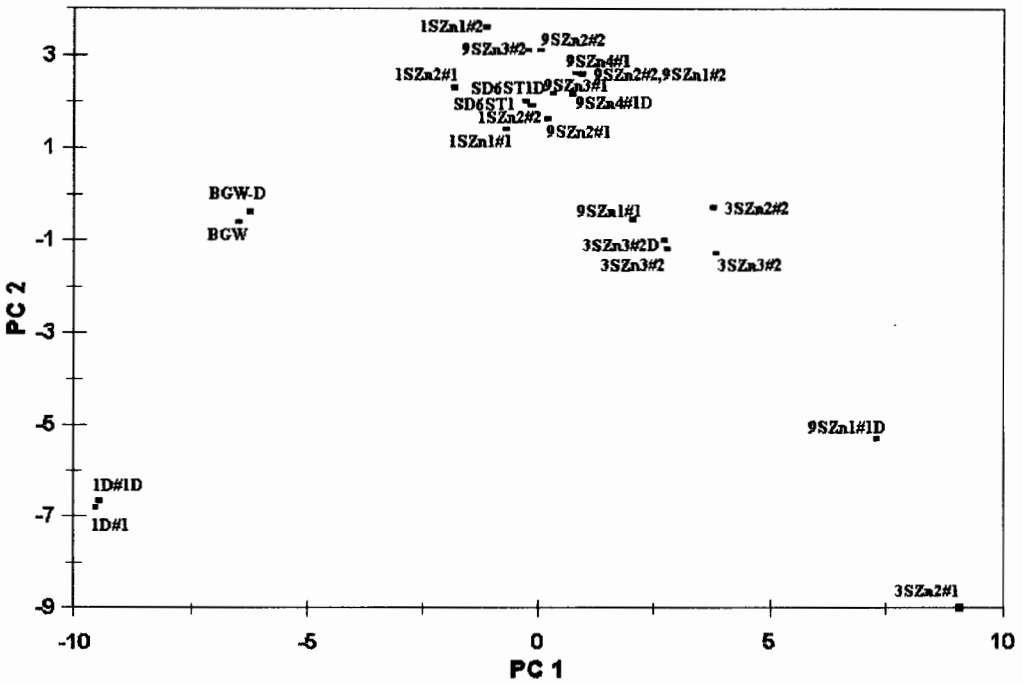

Fig. 11. Score plot for principal components analysis of the concentration of 58 trace elements in groundwater samples collected from Nye County, NV.

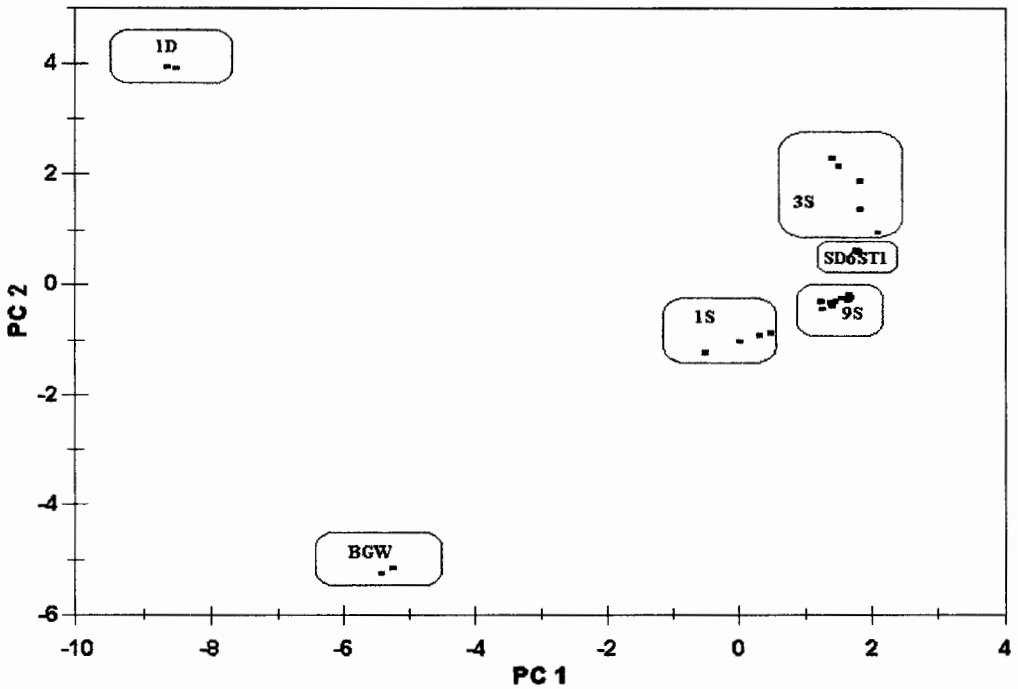

Fig. 12. Score plot for principal components analysis of the concentration of 19 trace elements (selected based on "30% relative variance between duplicates" criteria) in groundwater samples collected from Nye County, NV.

exceeded 30% of the variance observed between all 25 samples. A total of 19 trace elements were retained. For this data set, a total of 78% of the variance was explained in the first two components. The PC scores for the first two components are plotted in Fig. 12. It is shown in Fig. 12 that the selected variables can be used to clearly distinguish between all of the wells. All samples collected from the same well clustered quite tightly together with no overlap of clustering observed between the different wells.

4. Conclusions

New data simulation experiments were established to determine the best constant to use for substitution of '<DL' values within a given data set. Overall, for the data sets included in this study, the simulation experiments with single and multiple detection limits showed that substitution of '<DL' values with DL/2 was superior to substitution with 0 or DL. The results also suggested that all of the simulation methods behaved poorly when '<DL' observations exceed 30%. The effectiveness of the substitution of censored data with DL/2 (when less than 30% of the values are <DL) was demonstrated using a data set collected from Death Valley, CA. Approximately 20% of the data was censored and multiple detection limits existed between the different sampling dates. Tight clustering was observed between the PC scores (PC1 and PC2) of three measurements made on groundwater samples collected from the same springs or springs with very similar trace element chemistry.

A method for selection of variables was then developed for handling data sets with greater than 30% '<DL' values and also when uncensored data is available. Simulation experiments with duplicate measurements also showed that PCA results deteriorate quickly when the variation between duplicates exceeds 30% of the total variance over all measurements. This selection scheme was further evaluated using a data set from Nye County, NV. Elimination of variables with greater than 30% variance between duplicates when compared to the variance observed for all samples resulted in the ability to clearly distinguish each well from all others. The trace elements that are the best descriptors for this location were therefore identified and geochemical processes responsible for this discrimination can be examined.

As to be expected, the results may vary between different data sets. The methodology developed in this paper, however, can be used to determine the selection of substitution methods and variables for each particular data set.

Acknowledgements

This work was funded by the US Department of Energy (Grant #DE-FG-01-96EW56093). The authors are grateful to A. Smiecinski, C. Guo, J. Daniels, K. Lindley, T. Jankovic, V. Hodge, X. Zhou, and D. Baepler for their contributions.

References

[1] D.K. Kreamer, V.F. Hodge, I. Rabinowitz, K.H. Johannesson, K.J. Stetzenbach, Trace element geochemistry in waters from selected springs in Death Valley National Park, California, Ground Water 34 (1996) 95–103.

[2] M.J. Duane, G. Pigozzi, C. Harris, Geochemistry of some deep gold mine waters from the western portion of the Witwatersrand Basin, South Africa, J. Afr. Earth Sci. 24 (1997) 105–123.

[3] J.-L. Join, J. Coudray, K. Longworth, Using Principal Components Analysis and Na/Cl ratios to trace groundwater circulation in a volcanic island: the example of Reunion, J. Hydrol. 190 (1997) 1–18.

[4] K.J. Stetzenbach, I.M. Farnham, V.F. Hodge, K.H. Johannesson, Using multivariate statistical analysis of groundwater major cation and trace element concentrations to evaluate groundwater flow in a regional aquifer, Hydrol. Process. 13 (1999) 2655–2673.

[5] H. Suk, K.-K. Lee, Characterization of a ground water hydrochemical system through multivariate analysis: clustering into ground water zones, Ground Water 37 (1999) 358–366.

[6] I.M. Farnham, K.J. Stetzenbach, A.K. Singh, K.H. Johannesson, Deciphering groundwater flow systems in Oasis Valley, Nevada using trace element chemistry, multivariate statistics, and GIS, Math. Geol. 32 (2000) 943–968.

[7] Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, US Environmental Protection Agency Technical Report, EPA530-SW-89-026, Washington, DC, 1989.

[8] C.F. Chung, Estimation of covariance matrix from geochemical data with observations below detection limits, Math. Geol. 25 (1993) 851–865.

[9] D.R. Helsel, Less than obvious; Statistical treatment of data below the detection limit, Environ. Sci. Technol. 24 (1990) 1766–1773.

[10] J.U. Clarke, Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations, Environ. Sci. Technol. 32 (1998) 177–183.
[11] R. Aruga, Treatment of responses below the detection limit: some current techniques compared by factor analysis on environmental data, Anal. Chim. Acta 354 (1997) 255–262.
[12] S.W. Hinton, ΔLog-normal statistical methodology performance, Environ. Sci. Technol. 27 (1993) 2247–2249.
[13] L.H. Keith, Throwaway data, Environ. Sci. Technol. 28 (1994) 389A–390A.
[14] P.S. Porter, R.C. Ward, H.F. Bell, The detection limit, water quality monitoring data are plagued with levels of chemicals that are too low to be measured precisely, Environ. Sci. Technol. 22 (1988) 856–861.
[15] D.A. Chambless, S.S. Dubose, E.L. Sensintaffar, Detection limit concepts: foundations, myths, and utilization, Health Phys. 63 (1992) 338–340.
[16] R.J. Gilliom, R.M. Hirsch, E.J. Gilroy, Effect of censoring trace-level water-quality data on trend-detection capability, Environ. Sci. Technol. 18 (1984) 530–535.
[17] ASTM D 4210-89 Standard practice for intralaboratory quality control procedures and a discussion on reporting low-level data.
[18] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[19] K.J. Stetzenbach, M. Amano, D.K. Kreamer, V.F. Hodge, Testing the limits of ICP-MS determination of trace elements in groundwater at the parts-per-trillion level, Ground Water 32 (1994) 976–985.
[20] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.
[21] B.S. Everitt, G. Dunn, Applied Multivariate Data Analysis, Oxford Univ. Press, New York, 1992.
[22] S-PLUS, MathSoft, Seattle, WA.
[23] MINITAB, Release 13.1, MINITAB.
[24] STATISTICA '99, Release 5.5, StatSoft.