Combining multivariate statistical analysis with geographic information systems mapping: a tool for delineating groundwater contamination

Silas E. Mathes · Todd C. Rasmussen

Abstract Multivariate Statistical Analysis (MSA) has successfully been coupled with geographic information system (GIS) mapping tools to delineate zones of aquifer contamination potential. While delineating contaminants is key to site remediation, it is often compromised by a poor understanding of hydrogeologic conditions, and by uncertainties in contaminant observations. MSA provides improved estimates of contamination potential by augmenting observed contaminant concentrations with auxiliary information from other water-quality parameters. GIS is useful for organizing and managing water-quality information by visually communicating large amounts of information. The proposed method first establishes appropriate areal extents, GIS coverages, and scales for displaying groundwater contamination concentrations of tritium and the volatile organic contaminants trichloroethylene (TCE) and tetrachloroethylene (PCE) at the Savannah River Site, South Carolina, USA. Principal components analysis is used to group variables that are most indicative of contamination potential. Tritium contamination potential is best represented as the combination of tritium with the cations Al, Mg, Na and total dissolved solids, while PCE contamination potential is predicted using PCE and Cl. Maps of contamination potentials for 1993–1995 geochemical data compare favorably with measured contaminant concentrations during 1999. Cluster analysis of water-quality data groups geochemical and contaminant concentrations into zones of homogeneous behavior.

Received: 23 September 2005 / Accepted: 28 February 2006
Published online: 3 June 2006

S. E. Mathes · T. C. Rasmussen (✉)
Warnell School of Forestry and Natural Resources,
The University of Georgia
Athens, GA 30602-2152, USA
e-mail: trasmuss@uga.edu
Tel.: +1-706-542-4300
Fax: +1-706-542-8356

Present address:
S. E. Mathes
Tennessee Department of Environment and Conservation,
401 Church Street, Nashville, TN 37243-0435, USA

Résumé L'Analyse Statistique Multivariée (ASM) a été couplée avec succès avec les outils cartographiques des Sytèmes d'Information Géographique (SIG), afin de délimiter les zones à potentiel de contamination des aquifères. Alors que la connaissance de l'extension des contaminants est primordiale pour la réhabilitation des sites, elle est fréquemment compromise par une compréhension limitée du contexte hydrogéologique, et par les incertitudes sur l'observation des contaminants. L'ASM fournit une meilleure estimation du potentiel de contamination, en complétant les concentrations observées en contaminants par des informations annexes issues d'autres paramètres qualitatifs. Les SIG, par leurs capacités à conjuguer visuellement de grandes quantités d'information, sont utiles pour organiser et gérer les informations sur la qualité de l'eau. La première étape de la méthode proposée consiste à définir les limites spatiales appropriées, les couvertures SIG, ainsi que les échelles de visualisation des concentrations en contaminants des eaux souterraines, soit le tritium et les composés organiques volatiles trichloroéthylène et tetrachloroéthylène, sur le site de la rivière Savannah (Californie du Sud, USA). L'analyse en composantes principales est utilisée afin de regrouper les variables caractérisant de manière optimale le potentiel de contamination. Le potentiel de contamination en tritium est représenté le plus significativement par la combinaison du tritium, des cations Al, Mg et Na, et des solides dissous totaux (TDS). Le potentiel de contamination en tétrachloroéthylène est quant à lui estimé en utilisant le tetrachloroéthylène et Cl. Les cartes des potentiels de contamination établies sur les données géochimiques de la période 1993–1995 renvoient une image comparable aux concentrations en contaminants effectivement mesurées en 1999. Une analyse de groupement des données sur la qualité de l'eau rassemble les concentrations géochimiques et en contaminants dans des zones aux comportements homogènes.

Resumen El Análisis Estadístico Multivariado (AEM) se ha acoplado con éxito con herramientas cartográficas del Sistema de Información Geográfico (SIG), para delinear zonas de potencial contaminación de acuíferos. En tanto que la delimitación de contaminantes es importante para recuperación del sitio, ella está a menudo amenazada por una comprensión pobre de las condiciones hidrogeológicas, y por las incertidumbres en las observaciones del

 DOI 10.1007/s10040-006-0041-4

contaminante. El AEM proporciona estimaciones mejoradas de contaminación potencial al aumentar las concentraciones observadas del contaminante, con la información auxiliar de otros parámetros de calidad de agua. El SIG es útil para organizar y manejar la información de calidad de agua, a través de la comunicación visual de cantidades grandes de información. El método propuesto establece primero magnitudes de área apropiadas, el área cubierta por el SIG, y las escalas para mostrar las concentraciones de contaminación del agua subterránea, a partir de tritio y de contaminantes orgánicos volátiles como tricloroetileno (TCE) y tetracloroetileno (PCE) en el sitio del Río al Savannah, Carolina del Sur, EE.UU. El análisis de componentes principales se usa para agrupar variables que son más indicativas de contaminación potencial. El potencial de contaminación Tritio, se representa mejor como la combinación de tritio con los cationes Al, Mg, Na y los sólidos disueltos totales, mientras la contaminación potencial de PCE se predice usando PCE y Cl. Los mapas de potenciales de contaminación para los datos geoquímicos de 1993–1995 se comparan favorablemente con las concentraciones medidas del contaminante durante 1999. El análisis de racimo hecho en datos de calidad de agua, agrupa concentraciones geoquímicas y concentraciones del contaminante en las zonas de comportamiento homogéneo.

Keywords Contamination · Water quality · Statistical analysis · Geographic information systems · Savannah River Site

Introduction

Mapping of groundwater contamination is often complicated by infrequent and uneven distribution of monitoring locations, analytical errors in sample analyses, and large spatial variation in observed contaminants over short distances due to complex hydrogeologic conditions. While numerical simulation modeling is commonly used to delineate groundwater contamination plumes, this approach may be limited by insufficient knowledge of local hydrostratigraphic conditions. Also, managing and mapping extensive water-quality datasets can be difficult due to the multiple locations, times, and analytes that may be present.

An alternative to numerical simulation modeling uses statistical analysis of groundwater quality data to infer zones of potential contamination. Principal components analysis (PCA) is a multivariate statistical procedure designed to classify variables based on their correlations with each other. The goal of PCA, and other factor analysis procedures, is to consolidate a large number of observed variables into a smaller number of factors that can be more readily interpreted.

In the case of groundwater, concentrations of different constituents may be correlated based on underlying physical and chemical processes such as dissociation, ionic substitution or carbonate equilibrium reactions. Principal components analysis (PCA) helps to classify correlated variables into groups more easily interpreted as these underlying processes. The number of factors for a particular dataset is based on the amount of non-random variation that explains the underlying processes. The more factors extracted, the greater is the cumulative amount of variation in the original data.

Principal components analysis (PCA) has previously been used to generate accurate maps of monitoring wells grouped by their water-quality characteristics (Suk and Lee 1999; Ceron et al. 2000; Güler et al. 2002). Suk and Lee (1999) used multivariate analysis and geographic information systems (GIS) to correlate contaminant data with groundwater quality parameters for the purpose of identifying contaminated aquifer zones. They used this method to reduce several measured aquifer water-quality variables into a smaller series of underlying factors.

Cluster analysis is another multivariate statistical data reduction technique that can be used to group monitoring wells by aquifer water-quality behavior (Suk and Lee 1999). This method links variables hierarchically in the configuration of a tree with different branches. Branches that have linkages closer to each other indicate a stronger relationship among variables or clusters of variables. Suk and Lee ran factor scores generated by the PCA through a cluster analysis to group monitoring wells based on underlying water–rock interactions and recharge characteristics. These grouped wells were then mapped as aquifer zones using GIS software; zones identified by the researchers compared favorably with zones delineated with traditional hydrogeologic techniques.

Suk and Lee's (1999) multivariate analysis of geochemical data operated on the concept that each aquifer zone has its own unique groundwater quality signature, based upon the chemical makeup of the sediments that comprise it (Fetter 1994; Kehew 2001). Groups of water in aquifer zones delineated in this manner are known as hydrochemical facies (Fetter 1994). Groundwater dissolves minerals and other geochemical constituents from the geologic media that it inhabits. The dissolved mineral and chemical composition is unique to the water in each aquifer, forming a groundwater quality signature that can serve to identify the parent aquifer.

In northwestern Spain, Vidal et al. (2000) performed a principal components analysis (PCA) to reduce 14 water-quality variables to two factors correlated with saline and organometallic contamination. Vidal et al. plotted the two sets of factor scores from the PCA against each other, graphically labeling each observation according to spatial location (either a well or spring sampling point). Sampling points fell into different clusters on the graph, illustrating those that shared common groundwater quality signatures. The location of the sampling points on the graph ranked their respective aquifers according to vulnerability to saline and/or organometallic contamination.

Abu-Jaber et al. (1997) used a similar multivariate statistical exploration of geochemical data to identify predominant chemical interactions in known aquifer zones and to determine zone sensitivity to pollution from

 DOI 10.1007/s10040-006-0041-4

domestic sewer leakage. Meng and Maynard (2001) processed geochemical data using cluster and factor analysis; these groundwater classifications were then used as a basis for developing a conceptual geochemical model of their study area. Ochsenkühn et al. (1997) performed a cluster analysis on groundwater geochemical data to identify major trend axes representing dominant groundwater flow pathways. Other studies have used similar principles to correlate groundwater pesticide contamination with different crop rotations and for the inference of groundwater flow direction (Grande et al. 1996; Zanini et al. 2000).

Güler et al. (2002) compared a wide range of graphical and multivariate statistical techniques for classifying water chemistry data. Their analysis focused on data collected in the south Lahontan hydrologic regime within the basin and range geologic province of southern California, USA. Eleven water-quality variables were used to develop a robust classification scheme for partitioning water chemistry samples into homogeneous groups. They note that a combination of graphical and statistical methods provided more consistent and objective classification.

The goal of this study is to demonstrate the methodology for generating GIS maps of groundwater contamination using multivariate statistical analysis (MSA) of water-quality data. While GIS is routinely used for displaying map data, the use of statistical indicators of contaminant distributions is not. GIS is an important tool that is used to organize and manage large amounts of information for use in decision support systems. The methodology presented here is intended to more effectively utilize the emerging use of GIS for water-quality data analysis and interpretation.

Rather than mapping the observed contaminant distribution, a map of contamination potential is created using auxiliary water-quality data. Correlations between water-quality analytes are evaluated using principal components analysis (PCA) to generate water-quality factors. The water-quality factor containing each contaminant is then mapped using GIS to display the likely locations of groundwater contamination.

Adding auxiliary water-quality data provides additional information about local groundwater conditions, and reduces the reliance on individual contaminant observations. By incorporating the correlation between the contaminant and local water-quality variation, improved maps of likely contamination can be obtained. The goal is not so much to find relationships between water-quality variables and the contaminant, but rather to most effectively predict contaminated zones using all the observed data. In these situations, the combined variables yield a factor that more closely matches the likely contamination at the site. This approach is limited to those areas where auxiliary water-quality data are available, and does not reduce or eliminate the need for water-quality monitoring. Instead, it makes better use of existing water-quality information. The resulting aquifer water-quality maps provide insight into both the extent and history of groundwater contamination problems.

Water-quality data from the Savannah River Site (SRS) are used to demonstrate the methodology. Savannah River Site is a US Department of Energy facility located near Aiken, South Carolina, USA, near the Georgia–South Carolina border. Areas of groundwater contamination are present at multiple locations on the site due to the release of industrial and radioactive contaminants as by-products of nuclear weapons materials production from the 1950s until the 1990s. Most of the groundwater contamination is generally limited to the vicinity of its point of release.

Because groundwater near SRS is the major source of water for human consumption (Arnett et al. 1995), understanding the location of contaminated groundwater at SRS is imperative to maintaining both public safety and mitigating risk perception. To this end, SRS investigators have installed thousands of monitoring wells and maintain a quarterly (every 3 months) sampling program for the detection of contaminants (Arnett et al. 1995; Bollinger 1999). Despite extensive data collection efforts, monitoring data have been difficult to process for integration into historical or current maps of the contamination. Further, the underlying aquifer systems of the Southeastern Coastal Plain are complex; numerical groundwater flow-modeling efforts traditionally used at SRS have been limited to relatively small portions of the site where detailed hydrostratigraphic characterizations are available.

Methods

Study site

The Savannah River Site is a 790-km^2 area operated by the US Department of Energy. Savannah River Site is located on the Atlantic Coastal Plain in southwestern South Carolina along the Georgia–South Carolina border (Fig. 1). During the Cold War (1950s–1980s), SRS was used to manufacture nuclear materials including tritium and plutonium for the nation's defense. Manufacturing, waste disposal, and reactor facilities are scattered across SRS, but cover only a small proportion of the entire SRS area. This study focuses on two areas, the administration and manufacturing area (A/M Area) and the general separations area (GSA). Three representative contaminants were studied in these areas: tetrachloroethylene (PCE) and trichloroethylene (TCE), both found in groundwater beneath the A/M Area, and tritiated water (tritium), found within the GSA. Historically, most radioactive and industrial wastes were generated and stored here, and the greatest groundwater contamination is also found in these areas.

Volatile organic compounds (VOCs)—commonly used solvents for cleaning in the metals fabrication process—are found in groundwater beneath the A/M Area (Bollinger 1999). Two VOCs found at particularly high concentrations, PCE and TCE, are cleaning solvents used in nuclear-fuel manufacturing and metals-machining pro-

 DOI 10.1007/s10040-006-0041-4

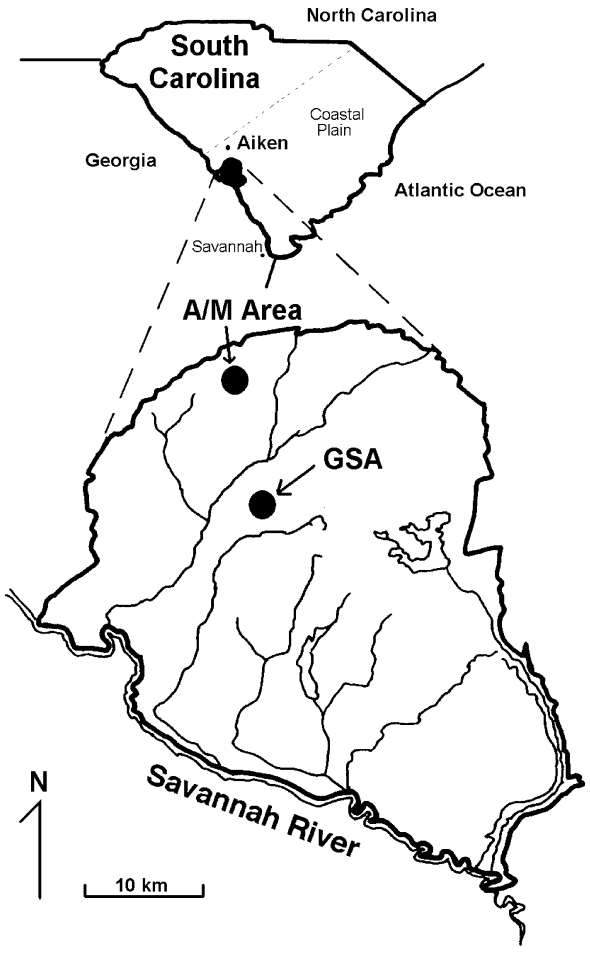

Fig. 1 Savannah River Site location map showing hydrographic features and the two main operations areas, the administration and maintenance Area (A/M Area) and the general separations area (GSA)

cesses (Arnett et al. 1995; Bollinger 1999). Both solvents, as well as other volatile organic compounds and heavy metals were disposed in shallow, unlined seepage basins from 1952 until the mid-1980s.

The general separations area (GSA), located in the central portion of SRS, is comprised of waste disposal sites including high-level radioactive waste tank farms and mixed-waste burial grounds, and facilities for the separation of specific radionuclides from targets produced at the reactor areas. Groundwater beneath the GSA is contaminated with a wide variety of chemicals and radionuclides, most notably tritium.

The complex behavior of nonaqueous phase liquids (including TCE and PCE) in groundwater can be simulated if the hydrostratigraphy is homogeneous and well understood (Kehew 2001). Unfortunately, the complex layering of aquifers and confining units present at SRS hampers the accurate calculation of aquifer hydraulic properties, causing large uncertainties in groundwater flow and contaminant transport model predictions (Harris et al. 1997; Miller et al. 2000; Kehew 2001).

The complexity of SRS hydrogeology is due to its location on the Upper Atlantic Coastal Plain. The SRS sediments form a complex, stacked lithology of unconsolidated sands, clayey sands, sandy clays, and calcareous muds deposited by periodic oceanic transmigration and by river and stream channel migration (Aadland et al. 1995). Adding to the complexity are ancient buried stream channels that conduct flow along preferential pathways.

The several thousand groundwater monitoring wells that were installed over the last 30 years (Arnett et al. 1995, Bollinger 1999) were screened at varying depths and are thought to correspond with the locations of specific aquifers and/or aquitards. Historically, groups of monitoring wells were placed and constructed as part of small-scale projects at SRS. Thus, concentration data from monitoring wells at SRS are clustered primarily around known contaminated areas, leaving gaps where water-quality data are unavailable.

Wells at SRS are generally sampled every 6 months for a wide range of analytes. To manage these water-quality data, SRS developed a geochemical information management system (GIMS). GIMS is an Oracle database maintained by a private contractor, Exploration Resources (Athens, Georgia, USA). The database is secure and can only be accessed by authorized personnel at the Savannah River Site.

Data selection

Although water-quality data are available from the earliest period of SRS operations, many analytes were not measured until the mid-1980s, resulting in smaller datasets for many groundwater constituents. In addition, wells were sampled at different time intervals. Wells located in areas of particularly high concern or at the location of groundwater remediation projects were often sampled once or more during a 3-month period (Q1 indicates data collected between January and March, Q2 indicates data collected between April and June, etc.). Monitoring wells in lower priority locations often were only sampled yearly. Some wells are installed when new remediation or characterization efforts begin while others are abandoned as projects on site reach completion.

A single parameter not measured over a quarter excluded a well from subsequent statistical analysis. This is especially evident for the entire A/M Area where many wells were not re-measured for tritium after the first quarter of 1993; the A/M Area was, for practical purposes, excluded from PCA analysis of subsequent quarters. Unfortunately, such changes in the number of observations make inter-quarter comparisons unclear because they introduce variation that is difficult to separate from the natural fluctuations of analyte concentrations. Substituting the mean for a missing observation is one strategy for handling missing values, yet this was not done because it may result in an incorrect correlation structure.

Data analysis focused on quarterly data collected between 1993 and 1995. Over these quarters, wells were

 DOI 10.1007/s10040-006-0041-4

Table 1 Summary of groundwater analyte concentrations for 3,914 observations from 1993 to 1995 in the administration and manufacturing area and general separations area, Savannah River Site, South Carolina, USA

Analyte	Units	Mean	Min.	Percentiles					Max.
				10th	25th	50th	75th	90th	
pH		4.6	3	4.4	4.9	5.5	6.5	8	12.8
TDS	mg/L	126.6	23	23	33	60	128	265	1,785
Al	mg/L	2.6	0.004	0.02	0.03	0.09	0.38	2.1	155
Ca	mg/L	13	0.01	0.62	1.32	3.83	13.6	34	482
Cl	mg/L	3.3	0.25	1.62	2.03	2.55	3.47	5.5	44.6
Fe	mg/L	0.5	0.004	0.005	0.01	0.04	0.17	0.7	48.1
K	mg/L	2.1	0.049	0.5	0.5	0.83	1.55	3.3	145
Mg	mg/L	1.4	0.002	0.26	0.4	0.66	1.22	2.7	40
Na	mg/L	14.4	0.495	1.82	2.57	4.66	12.8	34.2	360
Si	mg/L	13.3	0.152	6.06	7.37	9.38	13.6	26.2	158
SO_4	mg/L	5.6	0.094	1	1	1.67	5.47	10.6	440
PCE	µg/L	26.9	0.03	1	1	1	2.5	5	19,700
Tritium	pCi/mL	1,800	0.002	0.7	1.63	11.4	263	3,600	286,000

sampled frequently and groundwater was analyzed for the widest variety of constituents during any period of the SRS groundwater monitoring program. The program performed its most extensive site-wide analysis during the early 1990s. The peak of the groundwater monitoring occurred during the first quarter of 1993, and during this time over 1,000 wells were sampled for the analytes chosen to represent groundwater quality in this study.

The structure of the groundwater monitoring data led to the selection of a reduced number of analytes for statistical analysis. These variables were selected based on three criteria: (1) availability, (2) high spatial and temporal frequency of measurement, and (3) their likeliness to represent naturally occurring chemical conditions, i.e., providing a clear aquifer water-quality signature.

Thirteen water-quality analytes were selected for use in obtaining water-quality signatures. Variables meeting criteria include total dissolved solids (TDS), pH, aluminum (Al), calcium (Ca), chloride (Cl), iron (Fe), potassium (K), magnesium (Mg), sodium (Na), silica (Si), and sulfate (SO_4). Two additional variables, tritium and tetrachloroethylene (PCE), were selected because they represent radioactive and industrial groundwater contamination, respectively. Eleven of the 13 groundwater analytes are natural constituents of groundwater. Of these, six (Na, Ca, K, Mg, SO_4, and Cl) are major ions, while three are minor or trace constituents (Fe, Al, and Si), and the remaining two (TDS and pH) are common measurements.

Table 1 is a summary of groundwater analyte concentrations for all data collected within the A/M Area and GSA between 1993 and 1995. The table lists minimum and maximum concentrations, concentrations ranked by percentile, and mean concentrations for 3,914 observations. The table is useful because it indicates that there is substantial variation in the observed water-quality conditions at the site—at least four orders of magnitude for some variables. Principal components analysis (PCA) relies on this variation in water-quality parameters to assist in the identification of contaminant mapping. Note that the distribution is strongly skewed to the right, as indicated by the percentile and extreme observations, as well as by the ratio of the mean to the 50th percentile (i.e., median). This skew was removed by taking the logarithm of the water-quality variables, except for pH which is already defined as a logarithm.

The spatial extent of the data varied according to the wells sampled each quarter. For the first quarter of 1993, 744 wells were sampled for all 13 variables in this study. These wells were located in both the GSA (400 observations) and the A/M Area (275 observations). For subsequent quarters, tritium monitoring efforts focused heavily on the GSA and not on the A/M Area; approximately 90% of all observations that were selected for this study by matching analytes were located in the GSA vicinity during these quarters. As a result, analysis and mapping for the A/M Area relies heavily on data from the first quarter of 1993 and on the aggregate dataset.

Methods for mapping analyte concentrations were developed by utilizing the extensive data sets available from the SRS quarterly groundwater well-monitoring program. Scripts and shapefiles were developed within an ArcView-based GIS to produce maps depicting tritium and TCE contamination at SRS Areas during 1999. These maps suggest that most contamination at SRS is located near production facilities and that outlying areas have groundwater with little or no tritium and TCE.

These basemaps established the sizes, scales, and aerial extents that best communicate levels of contamination for different areas at SRS. These maps are intended for use as basemaps for the display of analyte concentrations from other time periods. These basemaps were created for later use in displaying the results of the aquifer water-quality analysis.

Principal components analysis (PCA)

Six quarters yielded sufficient monitoring data for stable PCA of the correlation matrix between observed water-quality variables. These quarters ranged from the beginning of 1993 to the beginning of 1995 and contained from 343 to 744 observations in total per analyte. The

 DOI 10.1007/s10040-006-0041-4

correlation matrix was obtained using logarithmic transformations of observed water-quality parameters (except for pH). Records with any missing values were excluded.

As a summary of time-averaged aquifer water-quality conditions, a PCA was also performed on an aggregate group of all monitoring well observations where the 13 analytes of interest were measured. There were 3,914 acceptable observations were identified for the aggregate PCA ranging from the fourth quarter of 1992 to the fourth quarter of 1999. All water-quality data were log-transformed (except for pH).

SPSS factor analysis module (SPSS Inc. 2005) was used with data for each quarter, specifying the principal components method with varimax rotation. Varimax rotation is a commonly employed method for assisting in factor identification, which was useful, in this application, to place the contaminants on a minimum number of factors. After several trial runs, it was found that extracting four factors during the analysis was sufficient to account for at least two thirds of the variation in six of the seven datasets. For all quarterly datasets, four factors were extracted and were used to generate four corresponding groups of factor scores by multiplying the original observations by the appropriate factor-score coefficients.

The factor scores were merged as four variables into the quarterly datasets for subsequent cluster analysis and for interpolation in ArcView. Factor loading tables were produced to show the strength of the relationship between each variable and factor. After examination of the PCA results for several quarters, it was found that high factor loadings repeatedly grouped the same variables together by component. However, due to rotation and differences in the amount of variation explained, the component position of these variable groups changed from quarter to quarter.

To characterize patterns among quarters more clearly, each numbered component (1–4) was assigned a letter (A–D) based on the three quarters with similar PCA results. For these three quarters, variables with absolute factor loadings greater than 0.4 for the first principal component were assigned to group A, group B for the loadings in the second component, C for the third, and D for the fourth. Using this nomenclature, subsets of analytes were identified that were related by similarly varying concentrations. Subsets containing the contaminant analytes (i.e., tritium and PCE) were reserved for subsequent mapping.

Cluster analysis

Cluster analysis was performed using the previously saved principal components for the purpose of grouping monitoring wells by geochemical zone (i.e., similar analyte behavior during a quarter). The hierarchical cluster analysis option in SPSS was selected to process observations from the four saved factor-score variables in each dataset. (Factor-score variables correspond to components identified by the PCA.) Factor scores were labeled using the well name field, and specified the 'Ward' method with squared Euclidean distance calculations for clustering.

In most cases, the Euclidean distance (defined as the square root of the sum of the squared differences) is used to determine the distance between observations, and then Ward's method is used to analyze the distances among linkages for the entire group of observations. Ward's method is a regression approach designed to minimize the sum of squared errors between any two clusters at each hierarchical level (Statsoft Inc. 2002).

The dendrogram generated from tree clustering provides a useful graphical tool for determining the number of clusters that adequately describe underlying processes that lead to the identification of homogeneous groups. Cluster membership can be saved for each observation and then mapped to show the spatial variation of these homogeneous groups. Results were saved as a cluster membership group number for each well. It was not possible to cluster the dataset with pooled observations from all quarters because it contained multiple observations for wells.

Cluster memberships for a range of classes (from three to ten) were determined. To identify the number of classes necessary to distinguish between aquifer zones, cluster tree diagrams, or dendrograms, were generated by importing each dataset into the SAS statistical package (SAS Institute Inc. 2005). Saving cluster membership for a range of classes provided for flexibility in later identifying the practical limits for resolving differences in aquifer water-quality behavior among wells on GIS maps.

GIS interpolation and mapping

Table 2 summarizes how GIS mapping was performed. Inverse distance weighting (IDW) is an algorithm for spatially interpolating, or estimating values between measurements. Inverse distance weighting (IDW) is implemented in ArcView 3.2 GIS software. Each value estimated in an IDW interpolation is a weighted average of the surrounding sample points. Weights are computed by taking the inverse of the distance from an observation's location to the location of the point being estimated (Burrough and McDonnell 1998). The inverse distance can be raised to a power (e.g., linear, squared, cubed) to model different geometries (e.g., line, area, volume) (Guan et al. 1999). In a comparison of several different deterministic interpolation procedures, Burrough and McDonnell (1998) found that using IDW with a squared distance term yielded results most consistent with original input data.

Inverse distance weighting (IDW) interpolation was performed using factor scores for components correlated with the contaminant variables, tritium and PCE. The radial interpolation method created continuous surfaces for 500 m around each well based on the factor score values at the 12 nearest wells. No interpolations were generated for portions of the site outside the 500-m radius of any well. Interpolated surfaces reflected the coverage of

 DOI 10.1007/s10040-006-0041-4

Table 2 Procedure for aquifer water-quality signature mapping

Step	Activity
1	SRS well monitoring program provides groundwater samples and subsequent laboratory analysis
2	Measurements are stored in the geochemical management system (GIMS) database
3	Use ArcView Interface to query GIMS, creating single analyte dBase files
4	Select analyte files (eleven) based on availability and likelihood to represent natural aquifer water quality. Select contaminants (two)
5	Write custom ArcView script to average measurements by quarter (3 months), generating new files for each of the 13 analytes
6	Import files into Microsoft Access tables to link measurements based on common well location and analyte
7	Import linked measurements into Microsoft Excel to sort by measurement and to remove univariate outliers
8	Separate data into quarters containing sufficient (n>100) observations for analysis
9	Import each file into SPSS to calculate Euclidean distance scores, and to remove multivariate outliers
10	Perform PCA with Varimax rotation, and save factor scores
11	Perform cluster analysis on the factor scores to identify wells with similar water-quality behavior patterns
12	Interpolate and map factor scores in ArcView for the two PCA groups containing contaminant variables
13	Overlay wells color-coded by cluster onto factor score contaminant maps

wells; areas with high contamination potential often end abruptly on these maps.

Interpolated factor scores were generated using raster-based grid layers shaded to nine levels. Shades of orange were used to represent PCE contamination and shades of red to represent tritium contamination. The lightest shades corresponded to the lowest contamination potential, the middle shade (fifth in sequence) corresponded to the median contamination potential, and the darkest shades indicated the highest level of contamination potential. Factor scores moved into the positive range from the median to the highest contamination potential shade. Positive factor scores indicated a positive correlation with their respective contaminant-bearing component.

Tritium grids were symbolized using a nine-class graduated color scale of light pink to dark red. Low factor scores in each grid indicated low potential for contamination and were colored pink, while the highest factor scores suggested high potential for contamination and were colored dark red. PCE was symbolized similarly, using shades of orange instead of red. Two maps for each contaminant were generated to overlay both cluster-analysis results and raw contaminant concentrations from 1999.

Well points were also color coordinated according to their cluster membership for the shape files discussed earlier. Based on results from cluster diagrams, five levels were selected to represent geochemical zones and labeled wells accordingly. This layer was placed on top of one set of the tritium and PCE potential maps for each quarter. A total of 1,999 contaminant concentrations were overplayed onto the other set of contamination potential maps for comparison. Additional details regarding methods used for GIS interpretation can be found in Mathes (2002).

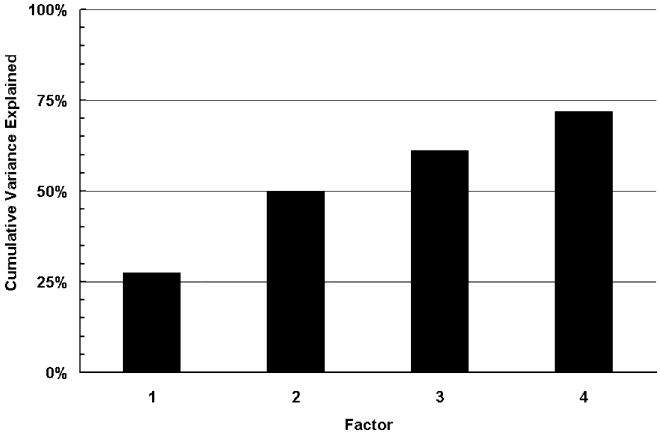

Fig. 2 Cumulative variance explained using principal components analysis, fourth quarter 1993

Results

Principal components analysis

Four factors were sufficient to explain at least two thirds of the variance for each quarterly correlation matrix. As an example, Fig. 2 presents the cumulative variation in observed water-quality data for the first four factors during the fourth quarter 1993. Figure 3 shows the PCA factor loadings for each of the analytes for the fourth quarter (Q4)1993: Al, Mg, Na, pH, TDS, and tritium were strongly correlated with the first factor; Ca, K, pH, Si, SO_4, and TDS were strongly correlated with the second factor; the third factor was strongly correlated with Al, Fe, and SO_4; and the fourth factor was best correlated with Cl and PCE. Note that the factor loading for pH is negative, indicating a positive loading with more alkaline conditions, and that SO_4 is found in both the second and third factors.

Table 3 summarizes the relative amount of variance explained by each factor for each quarter analyzed. Components in this table are labeled by letters that correspond to common variable groupings obtained by examining the factor loadings generated for each quarterly PCA and the aggregate dataset.

This pattern of variable grouping by component repeatedly occurred in PCAs for other quarters and for the aggregate dataset (i.e., all quarters shown in Table 3). Table 4 lists components with the absolute value of factor scores greater than 0.4 for all PCAs that were performed in this study. This clearly illustrates a strong quarterly pattern of repeating component-variable groups. The PCAs suggest that for the different hydrostratigraphic zones sampled by monitoring wells in this study, tritium concentrations behaved similarly to concentrations of the

 DOI 10.1007/s10040-006-0041-4

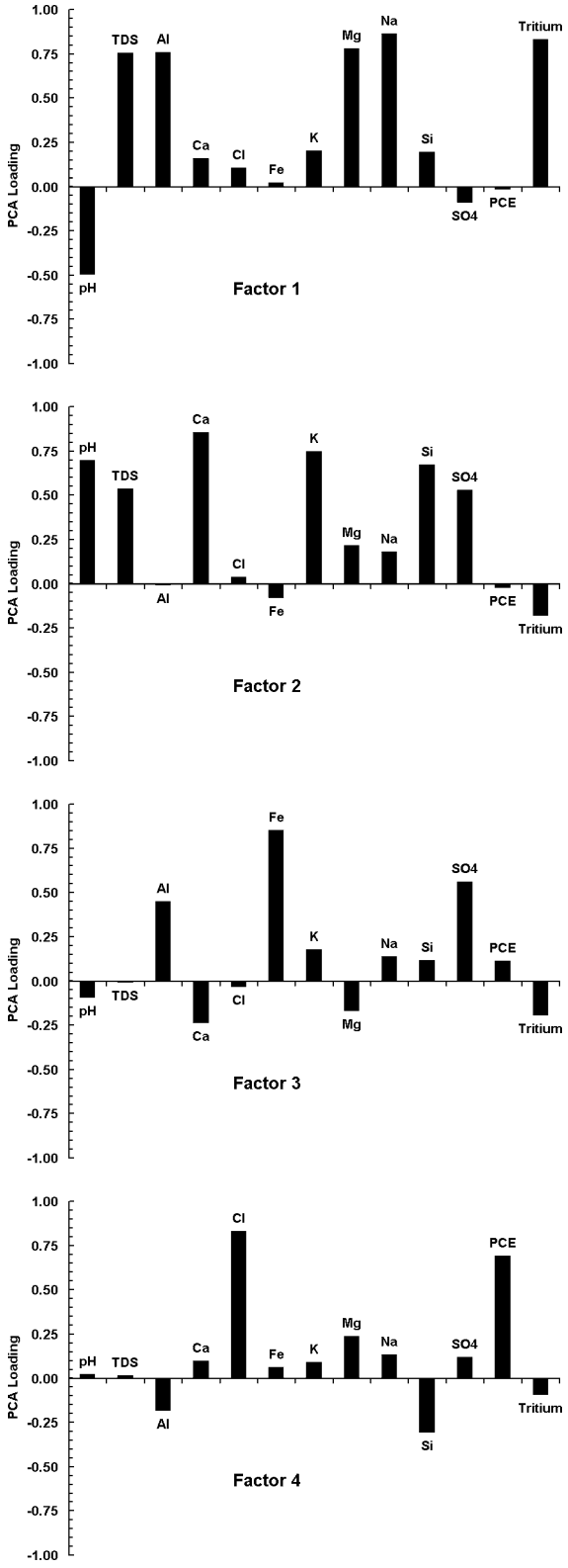

Fig. 3 Factor loadings calculated using principal components analysis, fourth quarter 1993

cations Al, Mg, and Na, along with TDS. PCE concentrations behaved similarly to those of the Cl anion.

Factor scores were calculated using the 13 original analytes observed at each well. For each PCA, this procedure yielded four sets of factor scores corresponding to the four components of the PCA. These scores were saved for subsequent cluster analysis. Factor scores for groups A and D, the two component-variable groups containing the contaminants tritium and PCE, were stored in a GIS for later interpolation and mapping.

Cluster analysis

Dendrograms (tree clusters) generated by preliminary cluster analysis implied that monitoring wells could be best separated into five homogeneous groups of wells. Cluster analysis divided wells into these different groups based on aquifer water-quality measurements. The Euclidean distance separating each of the five major clusters is greater than 100. In addition, an exploratory correlation analysis of SRS-identified aquifer zones with different levels (3–10) of cluster organization was also performed. The best correlation was observed between aquifer zones and a five-cluster system.

GIS maps

Groundwater concentration levels of the two factors associated with the most reliably measured analytes, tritium and TCE, were mapped for year 1999 sampling quarters and show the approximate areal extent of contamination at SRS. These maps are interpolations of factor scores, not the raw contaminant data. Thus, they incorporate both contaminant concentrations along with auxiliary information provided by the correlated water-quality variables. To roughly gauge the accuracy of contamination potentials generated from 1993–1995 data, unprocessed concentration data from all four quarters of 1999 were overplayed (shown in Figs. 4 and 5) where 1999 TCE and tritium concentration data (respectively) are symbolized using graduated color legends.

TCE and PCE were often discharged at the same locations and times at SRS, share many chemical characteristics, and behave similarly as dense nonaqueous phase liquids (DNAPLs) in groundwater. Despite a five- to six-year time lag, areas with high contamination potential for both tritium and PCE strongly coincided with later point observations of relatively high-tritium and TCE concentrations. Factor scores interpolated for components uncorrelated with either of the contaminants did not correspond to 1999 locations of elevated tritium and TCE concentrations.

The structure and extents of potentially contaminated zones changed slightly from quarter to quarter. In addition to the natural variation of analyte concentrations, the number of wells monitored changed over time. Well monitoring activity was reduced subsequent to the first quarter of 1993; wells on the periphery of the two SRS areas were not monitored during every quarter examined, thus changing the extent of these interpolations. After the first quarter of 1993, fewer wells were monitored overall. While quarterly changes did not have a strong effect on the component structure of the PCAs in this study (Table 4), the reduction in data points is reflected in

 DOI 10.1007/s10040-006-0041-4

Table 3 Order of original principal components and their component group letter assignment for six quarters, and for all data aggregated

Component group	Analytes	All data	1993Q1	1993Q2	1993Q3	1993Q4[a]	1994Q1	1995Q1
A	Al, Mg, Na, Tritium, TDS, pH	1	2	1	1	1	1	1
B	Ca, K, pH, Si, SO_4, TDS	2	1	2	2	2	2	2
C	Al, Fe, SO_4	4	3	3	4	3	3	3
D	Cl, PCE	3	4	4	3	4	4	4

[a] Data for 1993Q4 are represented in Fig. 3

shifted extents and locations of potential contamination zones on the maps.

Groundwater quality zones at well locations were also mapped. These zones were derived from cluster analysis of the factor scores generated by each PCA. Earlier work has established that clustering factor scores can help delineate geochemical facies with unique aquifer water-quality signatures (Suk and Lee 1999).

The mapping of the cluster analysis results in five levels for both the GSA and A/M Areas. These different levels suggest that there are five distinct groundwater quality zones in both areas. In several portions of the GSA, groups of wells repeatedly shared a common cluster membership over subsequent quarters. The largest repeating clusters of wells were located around the burial grounds and the F-area and H-area seepage basins. It is important to note that the statistical software assigned cluster number labels to the cluster groups based on analysis for a single quarter. Cluster number labels and their colors on the maps do not necessarily match each other from one quarter to the next.

On these maps, clusters showed a distinct spatial pattern for the first quarter of 1993, shown in Figs. 6 and 7. GSA monitoring wells were grouped into four of the five cluster categories, called well groundwater quality zones. Wells of the category not represented at the GSA were prevalent at the A/M Area and one of the GSA category wells was not present at the A/M Area. These differences indicate that potentially diverse aquifer geochemistry separates the A/M Area from the GSA. In subsequent quarters, data were limited to the GSA. While all five cluster categories (well groundwater quality zones in Figs. 6 and 7) were represented at the GSA, the same three categories identified during the first quarter of 1993 were dominant. On a more local scale, neighboring wells screened at different depths within both areas often varied by cluster category.

In general, map interpolations suggest that portions of the A/M Area were most likely to receive PCE contamination, while portions of the GSA possessed the highest potential for tritium contamination. On a finer scale, interpolations for both PCE and tritium indicate a stronger likelihood of contamination near seepage basins and other waste disposal sites. Mathes (2002) provides additional GIS coverages for time periods and analytes not shown here.

Discussion

The results of the PCAs in this study suggest that a smaller, optimized number of groundwater quality variables may be measured to gain the same insight into aquifer water-quality behavior. Matching a reduced number of geochemical variables would unquestionably incorporate more wells and provide a larger, higher resolution spatial picture of aquifer signatures and contamination potential.

Despite quarterly spatial differences in well monitoring, the variable-component structure of PCAs for different quarters was remarkably similar. The comparable temporal behavior of variables validates the assumption

Table 4 Principal component membership by analyte and quarter and all data aggregated

Analyte	All data	1993Q1	1993Q2	1993Q3	1993Q4	1994Q1	1995Q1
No. of wells	3,914	744	362	368	343	383	347
Al	A, D	A, C	A	A, C	A, C	A, C	A
Ca	B	B	B	B	B	B	B
Cl	D	D	D	D	D	D	D
Fe	D	C	−C	C	C	C	C
K	B	B	B	B	B	B	B
Mg	A	A	A	A	A	A	A
Na	A	A	A	A	A	A	A
pH	B, −A[a]	B	B, −A[a]	B, −A[a]	B, −A[a]	B, −A[a]	B, −A[a]
Si	B	B	C	B, −D[a]	B	B	B
SO_4	D, C	B, C	B	B, C	C, D	C	C
PCE	D	D	D	D	D	D	D
TDS	A, B	B, A	A, B	A, B	A, B	A, B	A, B
Tritium	A	A	A	A	A	A	A

Negative signs (−) indicate a negative factor score for the corresponding component. All data category includes repeat measurements of the same wells for different quarters

[a] Indicates second factor with a loading greater than 0.4; the larger factor loading of each pair is in the first position

 DOI 10.1007/s10040-006-0041-4

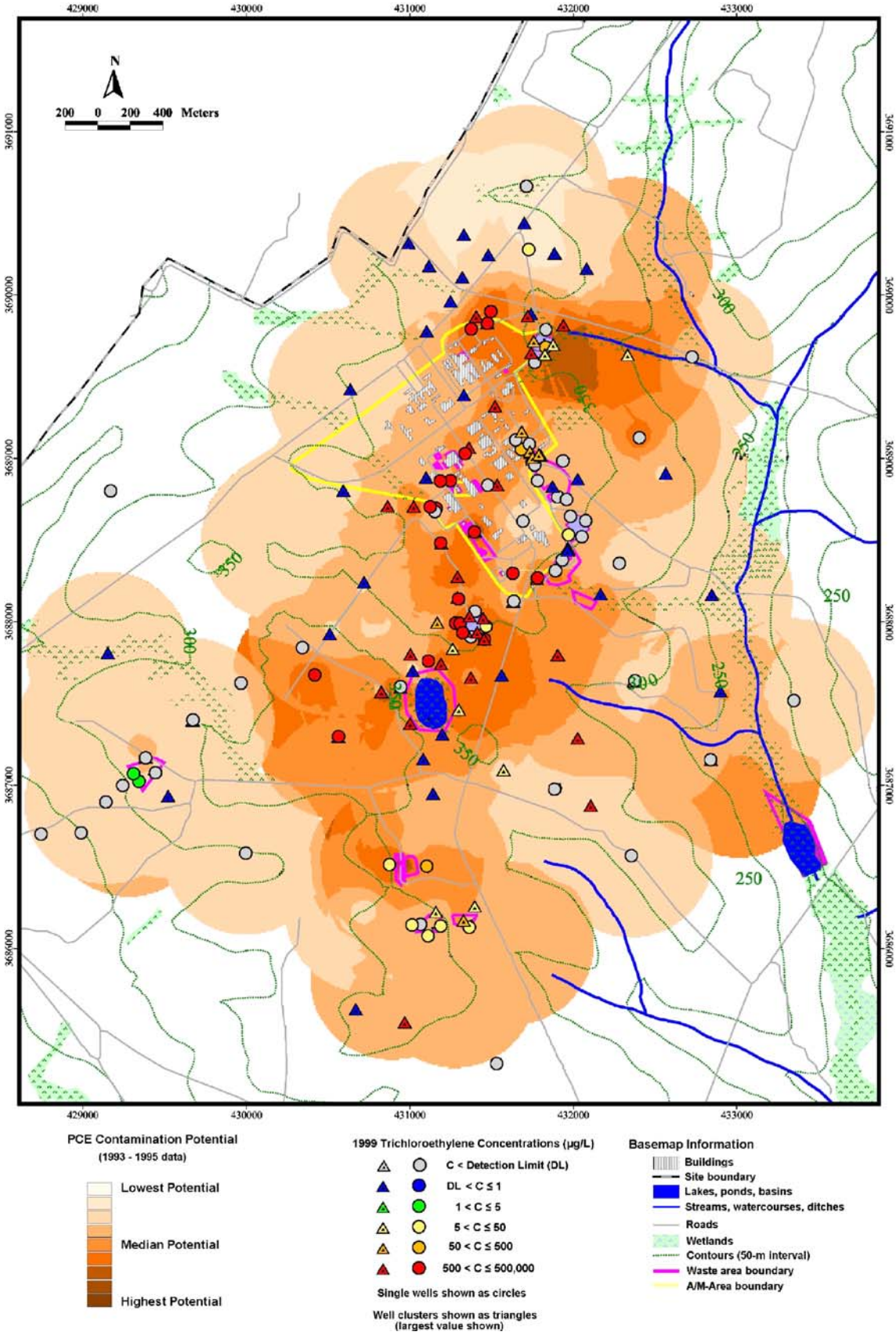

Fig. 4 A/M Area PCE contamination potential using inverse-distance squared weighting of PCA factor scores for 1993–1995 with 1999 TCE concentration overlay

 DOI 10.1007/s10040-006-0041-4

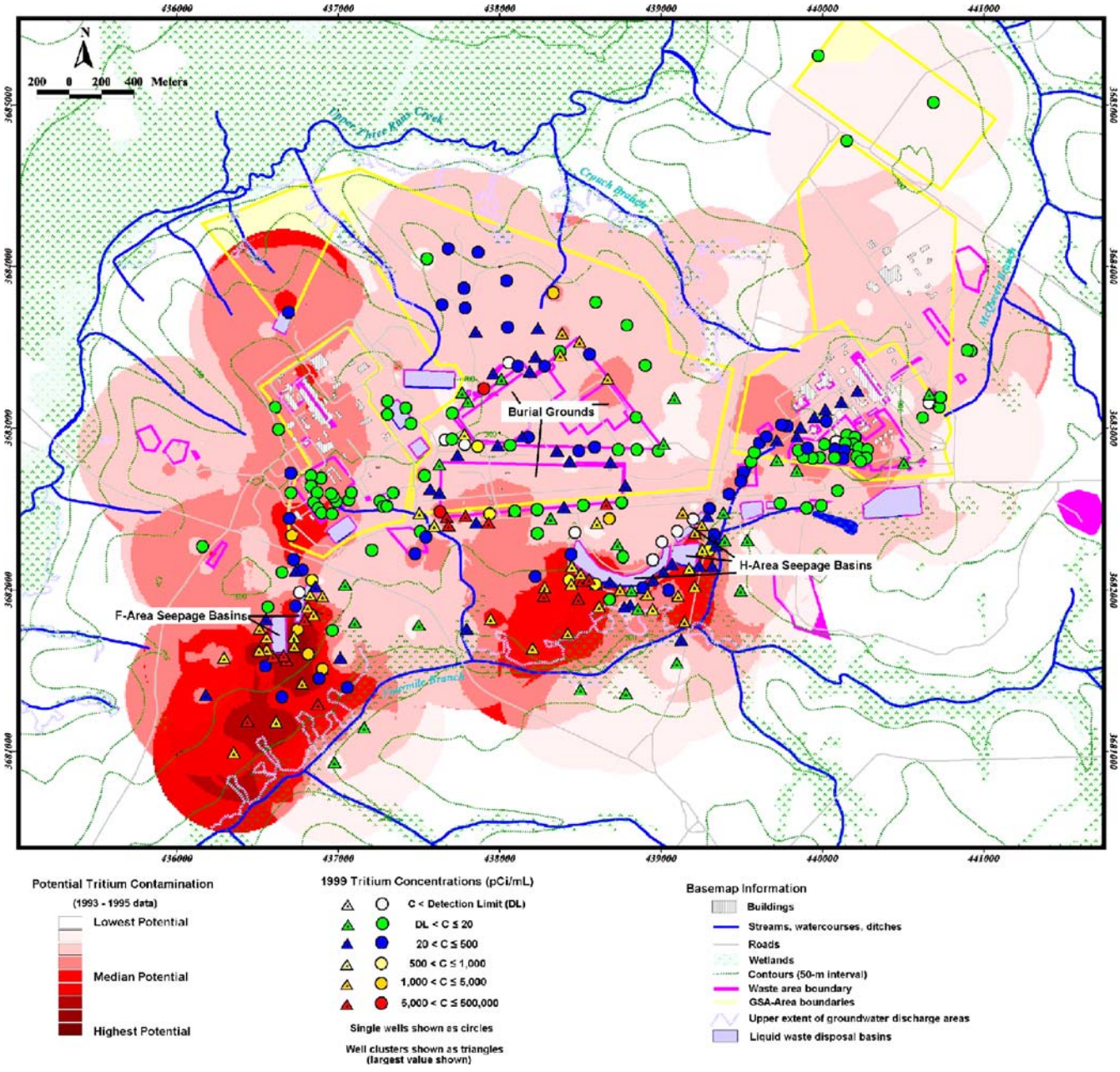

Fig. 5 GSA tritium contamination potential using inverse-distance squared weighting of PCA factor scores for 1993–1995 with 1999 tritium concentration overlay

that aquifer water-quality signatures were governed by underlying processes such as water–rock interactions and recharge–discharge relationships. The aquifer water-quality signature assumption is further validated because contamination concentrations measured in 1999 strongly corresponded to mapped locations of high contamination potential. These locations possessed positive factor scores for principal components highly correlated with contaminant variables.

Interpolation of factor scores incorporates underlying aquifer processes better than simple interpretation of raw contaminant concentrations. Using factor scores smoothes spikes in concentrations and identifies areas that may receive future contamination because of correlation with variables other than contaminants. The interpolations are meant to visualize the potential for contamination; interpolation to a larger distance was not performed because the coverage of monitoring wells became increasingly sparse with increased distance from SRS areas. As distance from large groups of monitoring wells increased, predicting the potential for contamination was less reliable because interpolations at these wells included fewer factor score observations.

Four water-quality factors were identified, including: (1) total dissolved solids plus aluminum, sodium, magnesium, and tritium; (2) total dissolved solids plus calcium, potassium, silica, sulfate, and alkaline conditions; (3)

 DOI 10.1007/s10040-006-0041-4

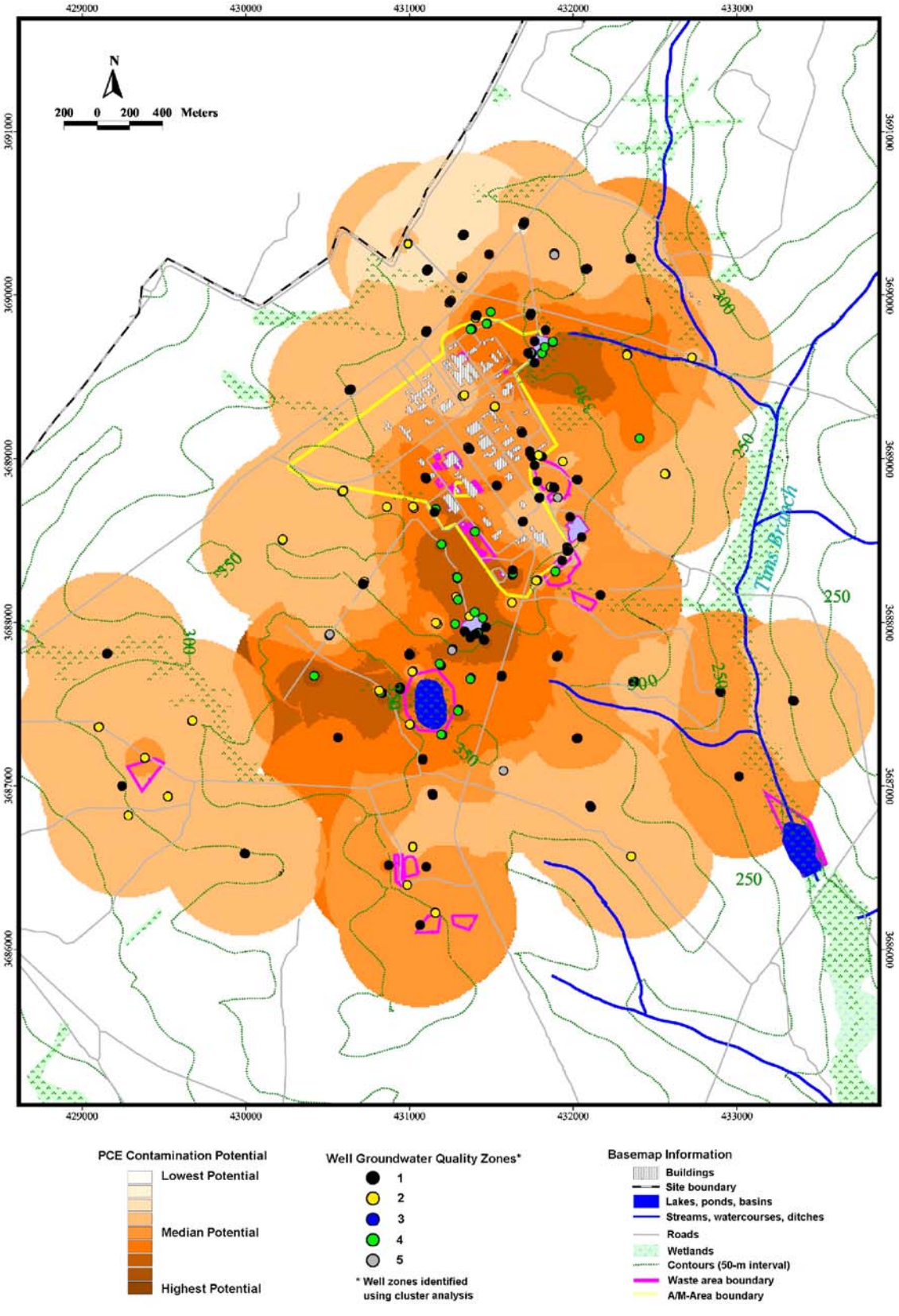

Fig. 6 A/M Area PCE contamination potential using inverse-distance squared weighting of PCA factor scores for first quarter 1993. Aquifer zone overlay based on cluster analysis also shown

 DOI 10.1007/s10040-006-0041-4

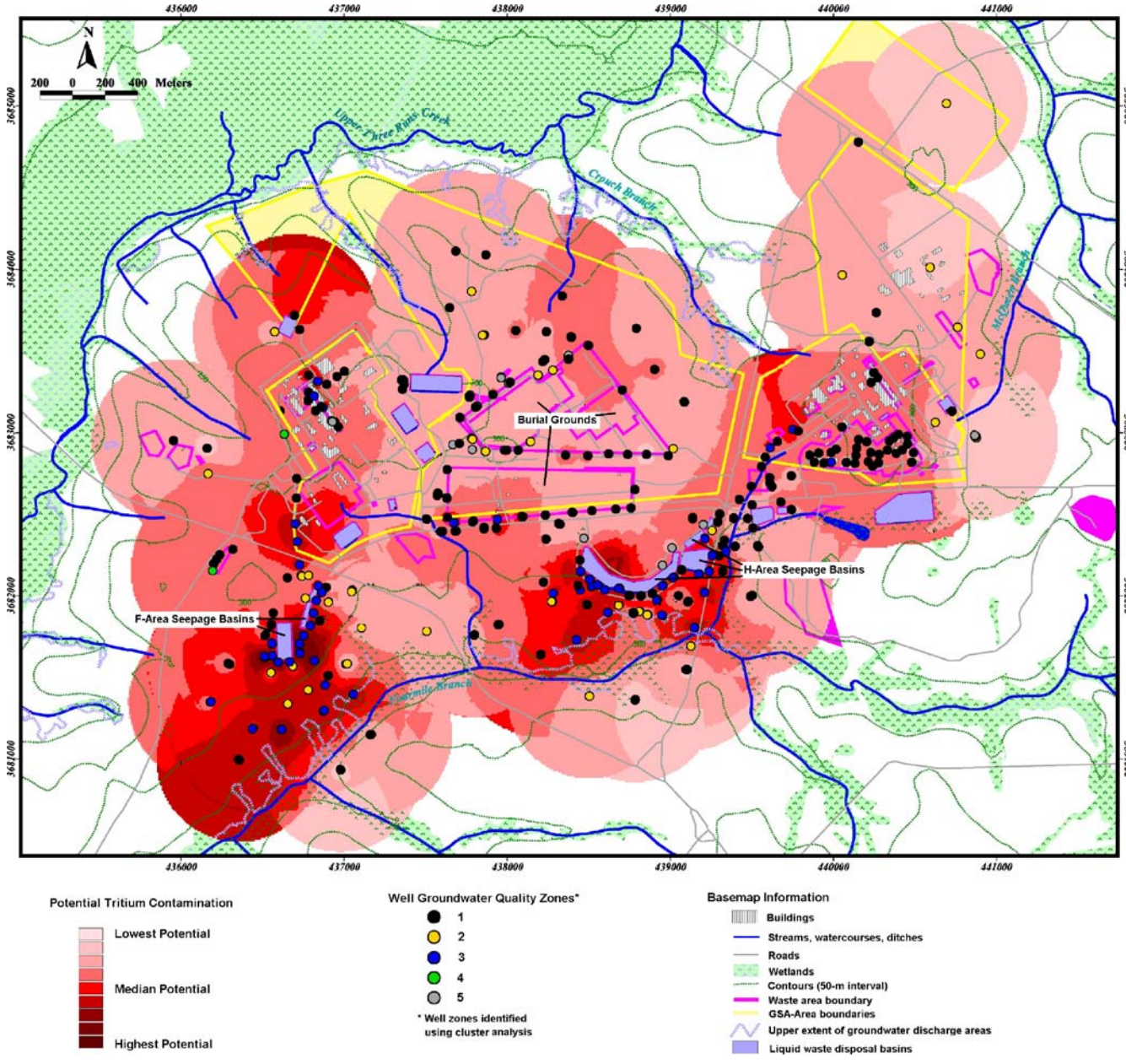

Fig. 7 GSA tritium contamination potential using inverse-distance squared weighting of PCA factor scores for first quarter 1993. Aquifer zone overlay based on cluster analysis also shown

aluminum, iron, and sulfate; and (4) chloride plus tetrachloroethylene. Total dissolved solids (TDS) is a measure of soluble ionic minerals or chemicals in water, and is elevated whenever the major cations or anions are present, which is the case for the first two factors.

For the SRS site, tritium appears to be related to the presence of sodium, magnesium, and aluminum. This could result from either co-release of these compounds, or their presence in the shallow aquifers near where tritium is released. The third factor contains minor cations plus sulfate, but not TDS, which indicates that these components do not dominate the whole water chemistry. The fourth factor is related to tetrachloroethylene and chloride. Again, this could be the result of degradation, co-release, or the type of groundwater present near the release point.

The identification of statistically distinct water-quality signatures assists in tracking the migration of contamination plumes, in that the occurrence of a group of correlated water-quality indicators may assist in delineating possible contaminant zones when the contaminant itself is absent due to field and laboratory errors. The use of augmented set of water-quality variables helps to improve the predictive accuracy of the delineated plume.

GIS mapping provides a graphical representation of the observed contamination utilizing all water-quality data. The spatial variation in the factor loading supplements the target contaminant with auxiliary information from correlated variables. While the presentation of both factor information with the original water-quality data is duplicative, it can be of assistance in providing an improved

 DOI 10.1007/s10040-006-0041-4

prediction of contamination extent in those areas where uncertain contaminant data are present.

Overall, the water-quality analysis methods described in this research show promise for supplementing previous studies of groundwater flow and contaminant transport at SRS. Mapping contamination potentials using GIMS, the SRS data management system, proved inexpensive and relatively efficient. Widely available statistical and GIS software packages were used to complete this analysis. The final data preparation and analysis steps took less than a week after refinement of the methodology. While mapping the results was time consuming and tedious, scripts were generated to automate interpolations and other repetitive tasks. Although the methods presented here compared favorably with subsequent data, a full validation may require focused analysis on a selected portion of SRS where fine-scaled hydrostratigraphic information is available.

Further efforts are needed to overcome questions related to the physical basis of the water-quality relationships presented here. One recommendation would be to compare the results reported here with geochemical model predictions. Another recommendation is to perform a test application of this methodology to a portion of the GSA where wells have been intensively monitored and the hydrostratigraphy has been well characterized. Known aquifer processes may be matched with aquifer water-quality signatures and these relationships could then be extrapolated to other areas of the site. Discriminant analysis is another possible tool that might be employed to identify distinct water-quality zones.

An in-depth comparison of hydrogeology with the aquifer water-quality zones delineated by cluster analysis is necessary to understand the aquifer processes represented by the zones. Seasonal differences in aquifer water quality may explain why some wells changed their cluster zone association over different quarters. These wells may draw water from shallow aquifers affected by local recharge and discharge. Many wells belonged to the same cluster zone for all quarters analyzed; these wells may draw water from deeper aquifers with more stable geochemistry. Nested monitoring wells screened at different depths fell into separate clusters, further indicating the sensitivity of the statistical techniques to vertical differences in groundwater geochemistry.

Conclusion

The purpose of this study has been to demonstrate the use of geographic information systems (GIS) for mapping groundwater contamination zones using statistical estimates of water-quality signatures. The maps presented here depict aquifer water-quality signatures as contamination potential for both the general separations area (GSA) and the administration and manufacturing area (A/M Area) at the Savannah River Site (SRS). The procedures outlined here may assist hydrogeologic characterization, contaminant plume delineation, and future well construction and monitoring decisions.

Principal components analysis (PCA) was used to identify four water-quality groups: (1) tritium, Al, Mg, Na, and total dissolved solids (TDS); (2) Ca, K, Si, sulfate, alkaline conditions, and TDS; (3) Al, Fe and SO_4; and iv) PCE and Cl.

Once the water-quality signatures were estimated, cluster analysis of the water-quality data was used to group geochemical and contaminant concentrations into zones of homogeneous behavior at monitoring wells. Maps of tritium and TCE concentrations for 1999 show the extent of contamination at SRS. The observed areal extent of groundwater contamination is limited to the immediate vicinity of disposal areas, and does not affect large areas of the Savannah River Site.

This approach has several key advantages. First, the methods rely on data that have already been collected; no additional parameters need to be measured to gain insight into historical and present groundwater conditions at SRS. Second, this project utilizes observed water-quality data, avoiding the uncertainty involved with contaminant transport modeling in the complex southeastern coastal plain environment.

The resulting GIS maps provide great utility in site-wide contamination management. Having the ability to quickly visualize contamination zones should assist in identifying the scope and nature of the problem, as well as the allocation of resources for site remediation. GIS maps are routinely used by management, and integrated water quality in this format will likely improve the overall effectiveness of environmental restoration efforts.

Acknowledgments This research was funded by a grant from the US Department of Energy, Westinghouse Savannah River Corporation through the Education, Research, and Development Association of Georgia Universities. Technical support was provided by John Reed and Jim Bollinger of the Savannah River Site. Editorial assistance provided by Sue Duncan substantially improved the quality of this manuscript.

References

Aadland RK, Gelici JA, Thayer PA (1995) Hydrogeologic framework of west-central South Carolina. Report 5. South Carolina Department of Natural Resources, Water Resources Division, Columbia, SC

Abu-Jaber NS, El Aloosy AS, Jawad Ali A (1997) Determination of aquifer susceptibility to pollution using statistical analysis. Environ Geol 31(1–2):92–106

Arnett MW, Karapatakis LK, Mamatey AR (1995) Savannah River Site environmental report for 1995. Westinghouse Savannah River Corporation, Savannah River Site, Aiken, SC

Bollinger J (1999) ArcView geographic information systems interface to the geochemical information management system. Westinghouse Savannah River Corporation, Savannah River Technology Center, Aiken, SC

Burrough PA, McDonnell RA (1998) Principles of geographical information systems for land resources assessment. Oxford University Press, New York

Ceron JC, Jimenez-Espinosa R, Pulido-Bosch A (2000) Numerical analysis of hydrogeochemical data: a case study. Appl Geochem 15:1053–1067

Fetter CW (1994) Applied hydrogeology, 3rd edn. Prentice Hall, New York

 DOI 10.1007/s10040-006-0041-4

Grande JA, Gonzalez A, Beltran R, Sanchez-Rodas D (1996) Application of factor analysis to the study of contamination in the aquifer system of Ayamonte-Huelva (Spain). Ground Water 34(1):155–163

Guan W, Chamberlain RH, Sabol BM, Doeringand PH (1999) Mapping submerged aquatic vegetation in the Caloosahatchee Estuary: Evaluation of different interpolation methods. Mar Geod 22:69–91

Güler C, Thyne GD, McCray JE, Turner AK (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol J 11:607–608

Harris MK, Flach GP, Thayer PA (1997) Groundwater flow and tritium migration in coastal plain sediments, Savannah River Site, South Carolina. WSRC-MS-97-0075, Westinghouse Savannah River Corporation, Aiken, SC

Kehew AE (2001) Applied chemical hydrogeology. Prentice Hall, Upper Saddle River, NJ

Mathes SE (2002) Geographic information systems (GIS) mapping of groundwater contamination at the Savannah River Site (SRS). MSc Thesis, The University of Georgia, Athens, GA

Meng SX, Maynard JB (2001) Use of statistical analysis to formulate conceptual models of geochemical behavior: water chemical data from the Botucatu aquifer in Sao Paulo state, Brazil. J Hydrol 250:78–97

Miller RB, Castle JW, Temples TJ (2000) Deterministic and stochastic modeling of aquifer stratigraphy, South Carolina. Ground Water 38(2):284–295

Ochsenkühn KM, Kontoyannakos J, Ochsenkühn-Petroulu M (1997) A new approach to a hydrochemical study of groundwater flow. J Hydrol 194:64–75

SAS Institute Inc. (2005) Website for SAS Institute Inc., http://www.sas.com. Accessed December 2005

SPSS Inc. (2005) Website for SPSS Inc., http://www.spss.com. Accessed December 2005

Statsoft Inc. (2002) Electronic statistics textbook, principal components and factor analysis, http://www.statsoftinc.com/textbook/stfacan.html. Accessed August 2005

Suk H, Lee KK (1999) Characterization of a ground water hydrochemical system through multivariate analysis: clustering into ground water zones. Ground Water 37(3):358–366

Vidal M, Melgar J, Lopez A, Santoalla MC (2000) Spatial and temporal hydrochemical changes in groundwater under the contaminating effects of fertilizer and wastewater. J Environ Manage 60:215–225

Zanini L, Novakowski KS, Lapcevic P, Bickerton GS, Voralek J, Talbot C (2000) Ground water flow in a fractured carbonate aquifer inferred from combined hydrogeological and geochemical methods. Ground Water 38(3):350–360

 DOI 10.1007/s10040-006-0041-4