Geotech Geol Eng (2007) 25:79–102
DOI 10.1007/s10706-006-0008-1

ORIGINAL PAPER

# Hierarchical probabilistic regionalization of volcanism for Sengan region, Japan

**Pinnaduwa H. S. W. Kulatilake · Jinyong Park · Pirahas Balasingam · Sean A. Mckenna**

Received: 28 April 2005 / Accepted: 18 April 2006 / Published online: 5 October 2006


**Abstract** A 1 km square regular grid system created on the Universal Transverse Mercator zone 54 projected coordinate system is used to work with volcanism related data for Sengan region. The following geologic variables were determined as the most important for identifying volcanism: geothermal gradient, groundwater temperature, heat discharge, groundwater pH value, presence of volcanic rocks and presence of hydrothermal alteration. Data available for each of these important geologic variables were used to perform directional variogram modeling and kriging to estimate geologic variable vectors at each of the 23949 centers of the chosen 1 km cell grid system. Cluster analysis was performed on the 23949 complete variable vectors to classify each center of 1 km cell into one of five different statistically homogeneous groups with respect to potential volcanism spanning from lowest possible volcanism to highest possible volcanism with increasing group number. A discriminant analysis incorporating Bayes' theorem was performed to construct maps showing the probability of group membership for each of the volcanism groups. The said maps showed good comparisons with the recorded locations of volcanism within the Sengan region. No volcanic data were found to exist in the group 1 region. The high probability areas within group 1 have the chance of being the no volcanism region. Entropy of classification is calculated to assess the uncertainty of the allocation process of each 1 km cell center location based on the calculated probabilities. The recorded volcanism data are also plotted on the entropy map to examine the uncertainty level of the estimations at the locations where volcanism exists. The volcanic data cell locations that are in the high volcanism regions (groups 4 and 5) showed relatively low mapping estimation uncertainty. On the other hand, the volcanic data cell locations that are in the low volcanism region (group 2) showed relatively high mapping estimation uncertainty. The volcanic data cell locations that are in the medium volcanism region (group 3) showed relatively moderate mapping estimation uncertainty. Areas of high uncertainty provide locations where additional site characterization resources can be spent most effectively. The new data collected can be added to the existing database to perform

P. H. S. W. Kulatilake (✉)
Geological Engineering Program, Department of Materials Science & Engineering, University of Arizona, Tucson, AZ 85721, USA
e-mail: kulatila@u.arizona.edu

J. Park · P. Balasingam
Department of Mining & Geological Engineering, University of Arizona, Tucson, AZ 85721, USA

S. A. Mckenna
Geohydrology Department, Sandia National Laboratories, 5800 MS 0735, Albuquerque, NM 87185, USA

future regionalized mapping and reduce the uncertainty level of the existing estimations.

**Keywords** Cluster analysis · Discriminant analysis · Entropy · Japan · Kriging · Probability · Regionalized mapping · Sengan · Variogram modeling · Volcanism

## Introduction

Nuclear Waste Management Organization of Japan (NUMO) is responsible for developing approaches to screen and locate a long-term site for a high-level nuclear waste repository in Japan. Any site chosen must meet the requirement that it is a location free from potential disruption from volcanic and fault activities. The ultimate goal is to screen the entire country of Japan to identify those areas that should be excluded from consideration for hosting a repository. In this paper, multivariate statistical techniques and geo-statistical interpolation techniques are applied on geologic variable data that are linked to volcanism to perform hierarchical probabilistic regionalized mapping of volcanism for Sengan region, Japan. The Sengan region was chosen as a test case because of the availability of geologic data and the multiple volcanic centers.

Many problems in geo-engineering and earth sciences involve an attempt to discretize the physical space into regions that are relatively homogeneous in a statistical sense with regard to some set of variables measured within them. *Regionalized classification* is a technique that provides a quantitative means of transferring a multivariate classification of a set of observations onto the physical, geographic space from which the observations were taken (Bohling et al. 1990; Harff and Davis 1990). Figure 1 illustrates the basic idea of regionalized classification. Two variables, A and B, are measured at a number of stations distributed throughout a study area (Fig. 1a). These observations can be plotted in variable space and classified into statistically homogeneous sets by multivariate statistical techniques such as cluster analysis (Anderberg 1973; Anderson 1984; Everitt 1993; Davis 2002). This process, referred to as *typification*, might be used to identify, say, three groups of interest (Fig. 1b). The groups identified contain observations that are simultaneously as similar as possible to other observations in the same group and as distinct as possible from observations in other groups.

After the groups are defined, each observation can be assigned a probability of membership in each group, which is essentially a transformation of the distance from each observation to a given group mean, or centroid, in variable space. This can be accomplished by using discriminant analysis (McLachlan 1992; Davis 2002) along with Bayes' theorem in probability. Each observation is then assigned to the group that produced the highest probability. These probabilities are then used to construct a probability of membership map for each group obtained at the typification step. These maps automatically delineate boundaries between different groups in the physical space. These spatial mapping steps are the essential features of *regionalization step* (Fig. 1c). The concepts of regionalized classification have been applied to the determination of the spatial distribution of formation thickness (Harff and Davis 1990), groundwater chemistry (Bohling et al. 1990), petroleum (Harff and Davis 1990), oil (Harff et al. 1989, 1990, 1993), gas (Harff et al. 1990, 1993), mineral resources (Harff et al. 1991), electrofacies properties (Moline and Bahr 1995) and grain size properties (Fernandez et al. 1997).

In Section 'Selected coordinate system and the grid system to show volcanic and geologic variable data for Sengan region' of the paper describes the

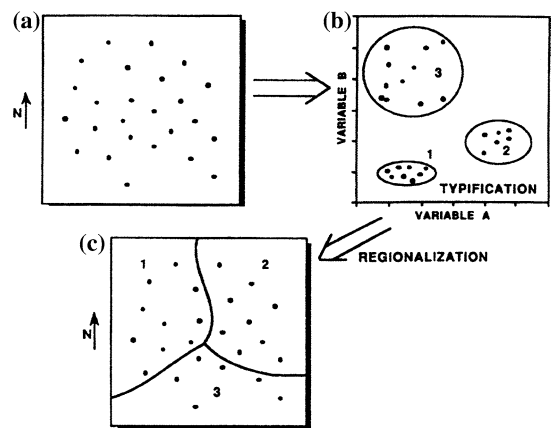


**Fig. 1** Schematic diagram of regionalized classification (from Bohling et al. 1990)

coordinate system and the regular grid system selected to work with the volcanism data and the geologic variable data (linked to volcanism) for Sengan region. The regionalized classification procedure used for hierarchical regionalization of volcanism in Sengan region, Japan consists of three steps: (1) *variable selection*, (2) *typification* and (3) *regionalization*. The *variable selection step* deals with identifying the most important geologic variables from all the available geologic variables that relate to volcanic risk, This step is covered in Section 'Available data for volcanism and geologic variables strongly linked to volcanism in Sengan region' of the paper.

Data on the most important geologic variables for Sengan region, Japan were obtained from NUMO. As is expected in a situation where the different data sets have been collected by different organizations for different purposes and at different times, the available data for geologic variables are not necessarily sampled at the same locations. In other words, for a given location, the geologic variable vectors were incomplete. However, complete variable vectors are required to perform multivariate statistical analyses. The following three options were considered to overcome the problem of incomplete vectors: (1) to perform multivariate statistical analyses on a minimum number of complete variable vectors coming from the sampling stations; (2) to use variogram modeling and kriging (Matheron 1971; Journel and Huijbregts 1978; Isaaks and Srivastava 1989; Deutsch and Journel 1998) to interpolate all variables to all sampling stations and then to use complete variable vectors to perform multivariate statistical analyses; (3) to use variogram modeling and kriging to interpolate all variables to a set of selected locations (points in a regular grid system) and then to use these complete variable vectors to perform multivariate statistical analyses. Because of the irregularity of the locations and the large differences of the available numbers of the data for different geologic variables, the third option was used to construct complete variable vectors. Section 'Geostatistical analysis' reports the variogram modeling and kriging performed for the selected important geologic variables to construct the complete variable vectors for Sengan region.

The constructed complete geologic variable vector data are classified into different groups of volcanism in Section 'Multivariate classification (typification) of volcanism for Sengan region' through cluster analysis in the *typification step*. The procedures stated before for the *regionalization step* are performed in Section 'Regionalized mapping of volcanism for Sengan region' to produce probability of membership maps of the identified volcanism groups for Sengan region. Entropy of classification is suggested to assess the uncertainty of the allocation process of the selected grid point locations based on the calculated posterior probabilities. The spatial distribution of calculated entropy in the Sengan region is shown in Section 'Regionalized mapping of volcanism for Sengan region.' Maps obtained in Section 'Regionalized mapping of volcanism for Sengan region' are compared with locations of recorded volcanism in Sengan region to evaluate the reasonableness of the predictions and to determine the locations where future data collection is needed to improve reliability of the predictions.

## Selected coordinate system and the grid system to show volcanic and geologic variable data for Sengan region

The Universal Transverse Mercator (UTM) projection coordinate system (PCS) available in ArcGIS 8.x software package is used in the paper to work with volcanic and geologic variable data available for Sengan region. The mid point of the Sengan region has a longitude close to 141 degrees and latitude close to 40 degrees. The Sengan region is within UTM zone 54 (138–144 degrees longitude). UTM zone 54 has the following properties: Ellipsoid: GRD 80; Central meridian: 141.00000; Reference longitude; 0.00000; Scale factor: 0.99960; False easting: 500000.00000; False northing: 0.00000.

As the study area, longitudes between 139.667 and 142.0844 degrees and latitudes between 39.333 and 40.667 degrees have been used with Greenwich as the prime longitude to cover Sengan region. The aforementioned region covers 209 km in the E–W direction and 149 km in N–S direction. Table 1 shows the starting and ending

**Table 1** X and Y Coordinates of the study area according to the UTM zone 54 PCS

| | Starting (km) | Ending (km) | Difference (km) |
|---|---|---|---|
| X-coordinate | 385.086 | 594.086 | 209 |
| Y-coordinate | 4,354.616 | 4,503.616 | 149 |

X and Y coordinates of the study area according to the UTM zone 54 PCS. This study area was divided into 1 km square cells as shown in Fig. 2. The total number of 1 km cells in the study area turned out to be 31141. Out of that total number, 23949 one km cells (76.9 percent) occupied the land portion. The rest of the 1 km cells are located in the oceanic portion. All of the geologic data provided by NUMO use a geographic coordinate system (GCS) to specify the locations. Data of all the geologic variables considered in this study have been converted from a GCS to UTM zone 54 PCS with the datum GRD 80.

## Available data for volcanism and geologic variables strongly linked to volcanism in Sengan region

### Recorded volcanism in Sengan region

The available data on observed volcanism can be separated into three groups as follows: 'edifice by vent'=80 points; 'edifice by topography'=5 points; 'volcano center by topography'=30 points. Figure 3 shows the locations for the aforesaid observed volcanism in the study area. It is noted that these observations of volcanism are not used in the classification and mapping process, but are only used afterwards as a check on the results.

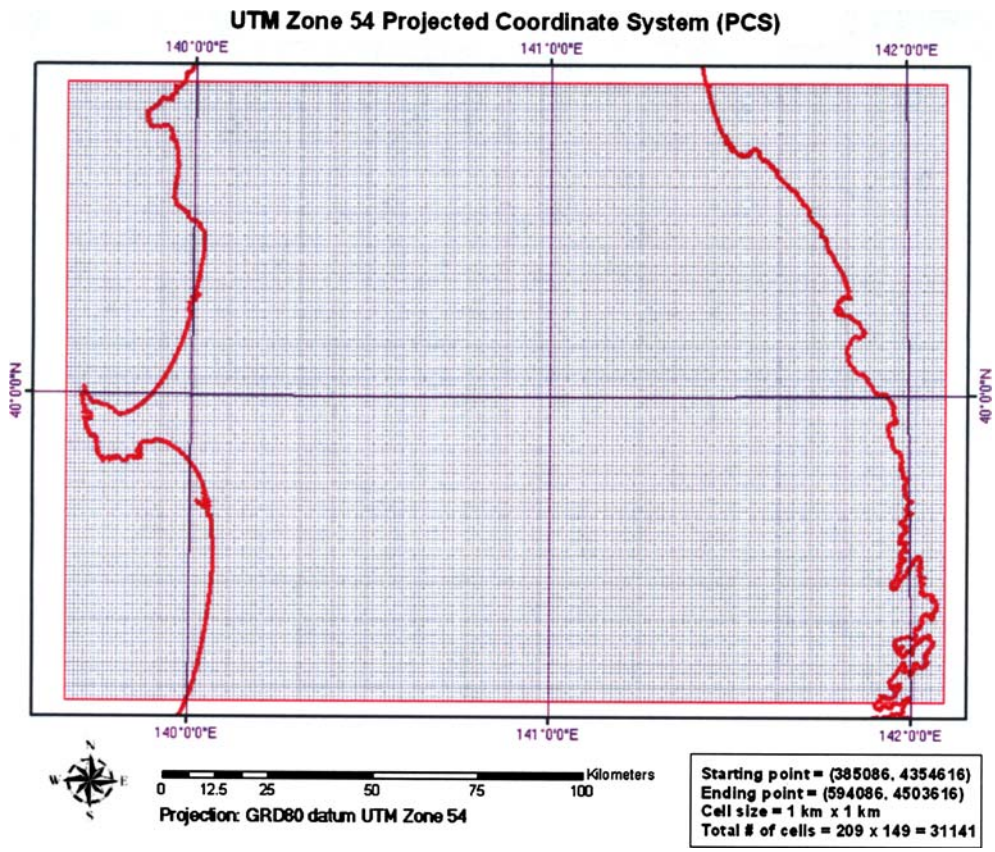


**Fig. 2** One km square grid system on UTM zone 54 PCS for the Sengan region

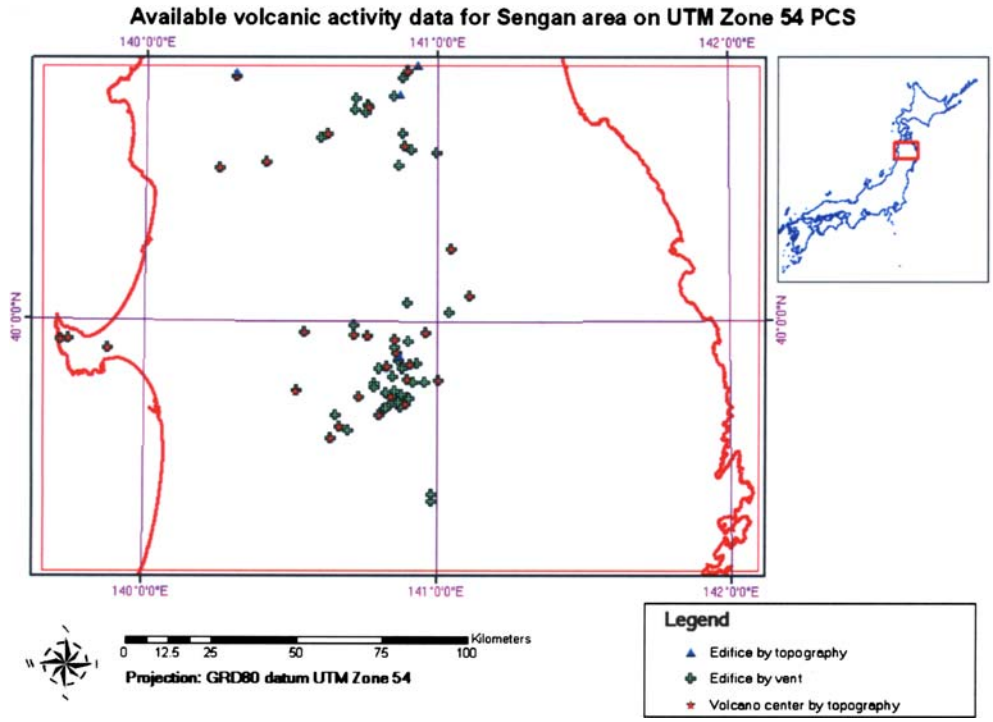


**Fig. 3** Available volcanic activity data for Sengan region

## Geologic variables most important to volcanism

A volcanism workshop held to discuss on conceptual models of volcanic activity (Arnold et al. 2003) identified the following geologic variables as the most important variables related to volcanism: geothermal gradient, geothermal heat flow, groundwater temperature, presence of quaternary volcanic rocks, presence of hydrothermal alteration and groundwater pH value. Seismicity (shallow), magnetic, teleseismic, radar interferometry, spectral satellite, gravity, horizontal shear strain, groundwater chemistry, elevation, slope magnitude, and slope orientation were identified as variables of secondary importance in evaluating volcanic risk at the regional scale.

The map obtained for the volcanism in Sengan region (Fig. 3) was visually compared to the map obtained for each of the aforesaid most important as well as secondary important geologic variables to determine whether the considered geologic variable is strongly correlated to volcanic activity. This comparison confirmed that the aforesaid most important geologic variables are strongly correlated and the rest of the geologic variables are either poorly or weekly correlated to volcanism. Therefore, the geostatistical and multivariate statistical analyses reported in Section 'Geostatistical analysis' were conducted only for the geologic variables that were labeled as most important to volcanism. The data available for these most important geologic variables are given below in rest of this Section.

## Groundwater temperature

Groundwater temperature data were obtained through the following sources: groundwater databases; hot spring databases; fumaroles and geothermal wells. In these databases, several temperature values were sometimes available for one horizontal location at different depths. The average of the available data with depth was used to represent the groundwater temperature at the considered horizontal location in such a situation. Figure 4 shows the constructed data base available for groundwater temperature on UTM zone

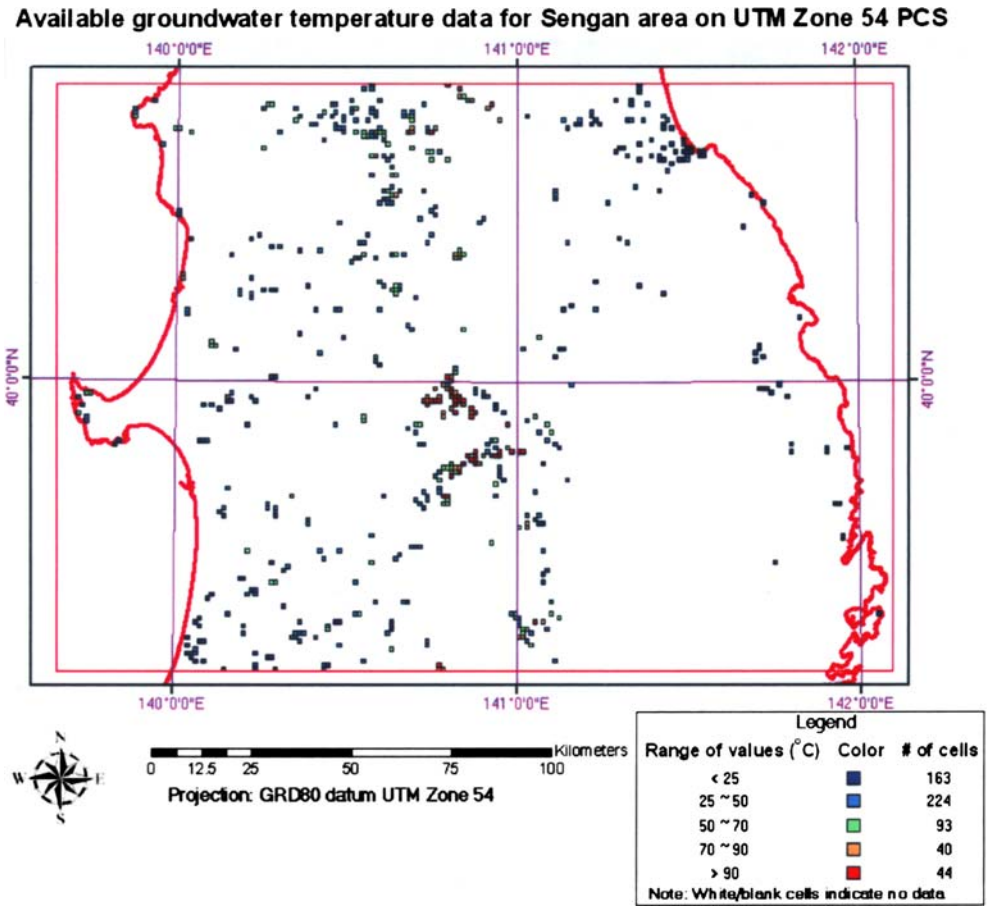


**Fig. 4** Available groundwater temperature data for Sengan region

54 PCS. A few 1 km square cells have more than one data point and many cells do not have a single data point. In Fig. 4, the groundwater temperature data are separated into five arbitrary groups according to the level of temperature. Note that the higher the groundwater temperature, the higher the chance of volcanism. Also note that there is a large area in the southeastern region of the study area with a few or no data.

## Groundwater pH value

Note that even though temperature values are available from the groundwater databases, hot spring databases, fumaroles and geothermal wells, pH values are only available from the groundwater databases. Figure 5 shows the available pH values for the study area on UTM zone 54 PCS. A few 1 km square cells have more than one data point and many cells do not have a single data point. In Fig. 5, pH data are separated into five arbitrary groups based on the level of pH value. Note that the lower the groundwater pH, the higher the chance of volcanism. Also note that there is a large area in the southeastern region of the study area with a few or no data.

## Geothermal gradient

Figure 6 shows the available geothermal gradient values for the study area on UTM zone 54 PCS in units of C°/km. A few 1 km square cells have more than one data point and many cells do not have a single data point. In Fig. 6, the geothermal gradient data are separated into five arbitrary groups as for the groundwater temperature. Note

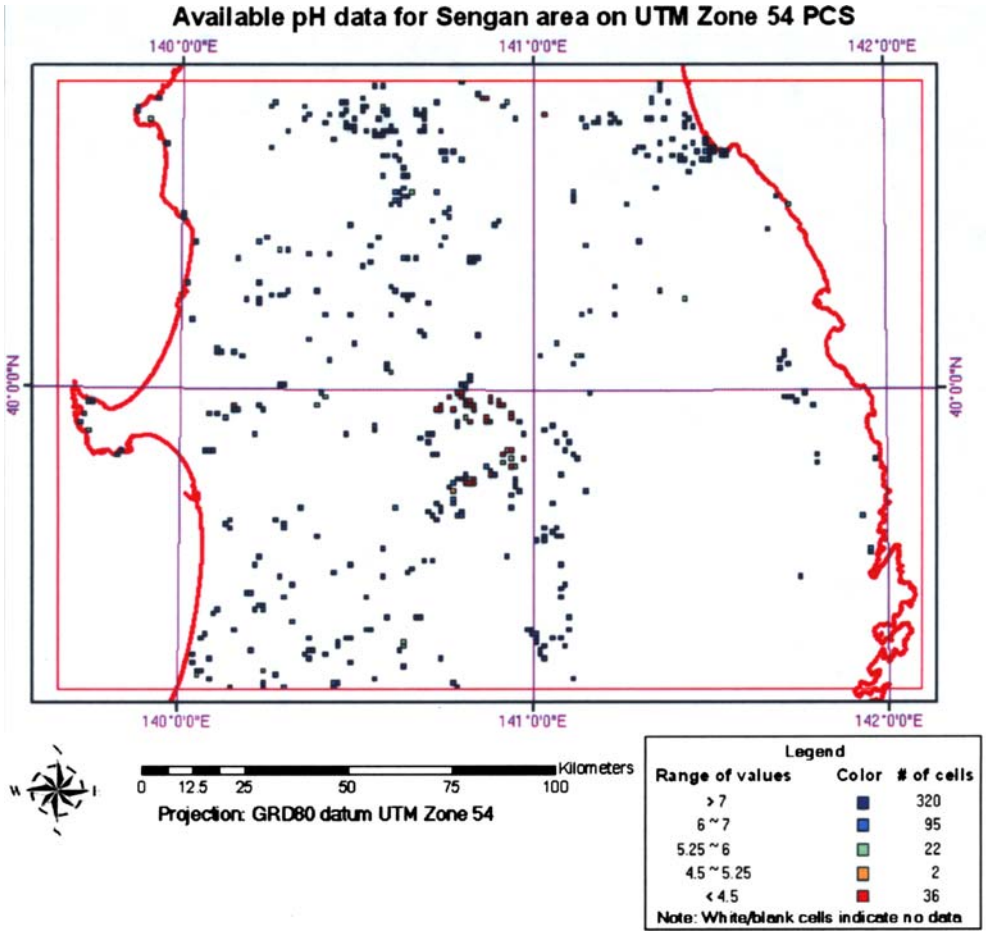


**Fig. 5** Available groundwater pH data for Sengan region

that the higher the geothermal gradient, the higher the chance of volcanism. Also note that there is a large area in the eastern region of the study area with no data.

Heat discharge

Heat discharge data available through hot springs on a raster image were converted to a 1 km square grid system using UTM zone 54 PCS. The obtained map is shown in Fig. 7. Numbers given in the map are the logarithms of the heat discharge in units of $\log_{10}(\mu \mathrm{Wm}^{-2})$. In Fig. 7, the heat discharge data are separated into five arbitrary groups as for the groundwater temperature. Note that the higher the heat discharge, the higher the chance of volcanism. Also note that there is a large area in the southeastern region of the study area with a few or no data.

Presence of quaternary volcanic rocks

Available images of the distribution of quaternary volcanic rocks were used in arriving at the polygonal coverage shown in Fig. 8 to represent the distribution of volcanic rocks in the study area.

Presence of hydrothermal alteration

Available images of the distribution of hydrothermal alteration were used in arriving at the polygonal coverage shown in Fig. 9 to represent the distribution of hydrothermal alteration for the study area.

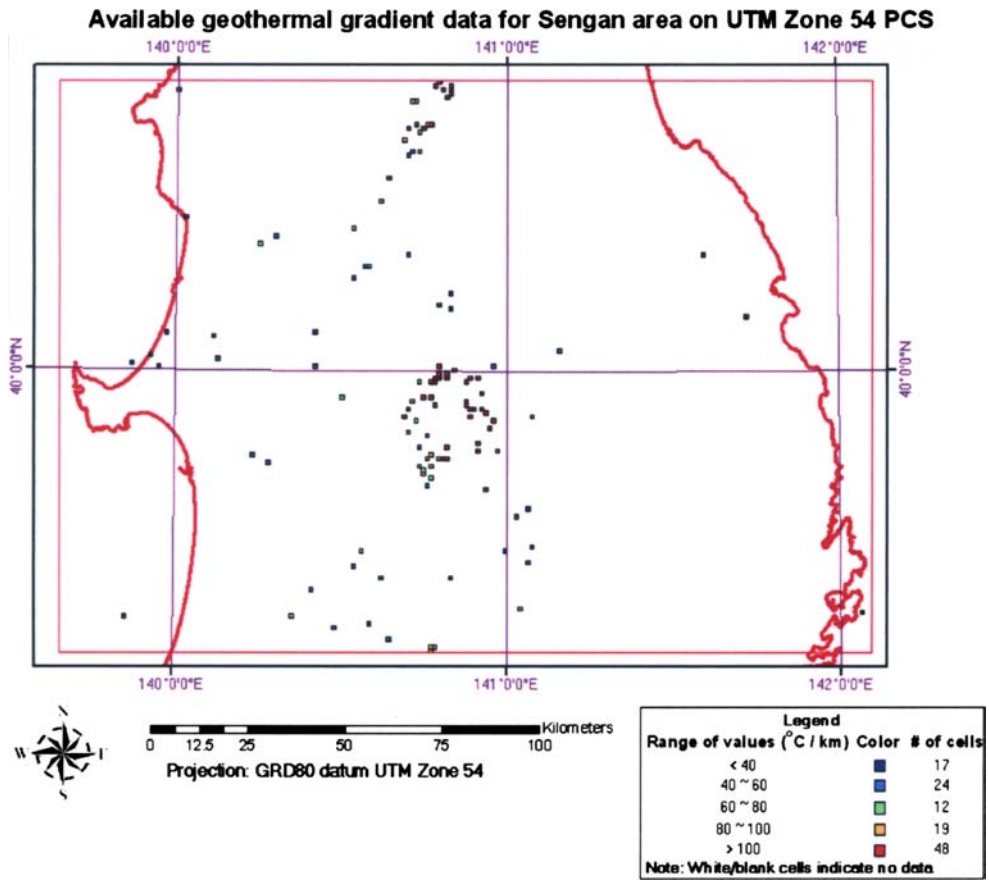


**Fig. 6** Available geothermal gradient data for Sengan region

## Geostatistical analysis

The previous section provided maps for available data on UTM zone 54 PCS for the following geologic variables: groundwater temperature, groundwater pH value, geothermal gradient, heat discharge, presence of volcanic rocks and presence of hydrothermal alteration. Each map was shown on a 1 km square grid system. Each 1 km cell can be considered as a sampling station for geologic variables. When all of the 6 geologic variables are taken together at any one location, the existing data vector is incomplete. However, complete geologic variable vectors are required to perform multivariate statistical analysis. Spatial variation of a geologic variable including the anisotropy for the study area on the two-dimensional horizontal plane can be studied by constructing the variogram surface and directional variograms. Variogram modeling then can be performed to capture the essential properties depicted by the directional variograms. The variogram model then can be used along with the kriging technique to estimate the geologic variable value at the center of each 1 km cell where no data are available for the considered geologic variable. In addition, kriging can be applied to refine the geologic variable value at a center of a cell where data are available for the considered geological variable. This procedure was applied to each of the first four of the six geologic variables to complete the geologic variable vector for the said four variables in the study area. Variogram modeling was performed using the computer program VARIOWIN version 2.2 (Pannatier 1996). The study area was extended by 50 km to both north and south to use available data beyond the boundary of the study area to better estimate the geologic variable values at the cells located at the north and south boundaries of the study region.

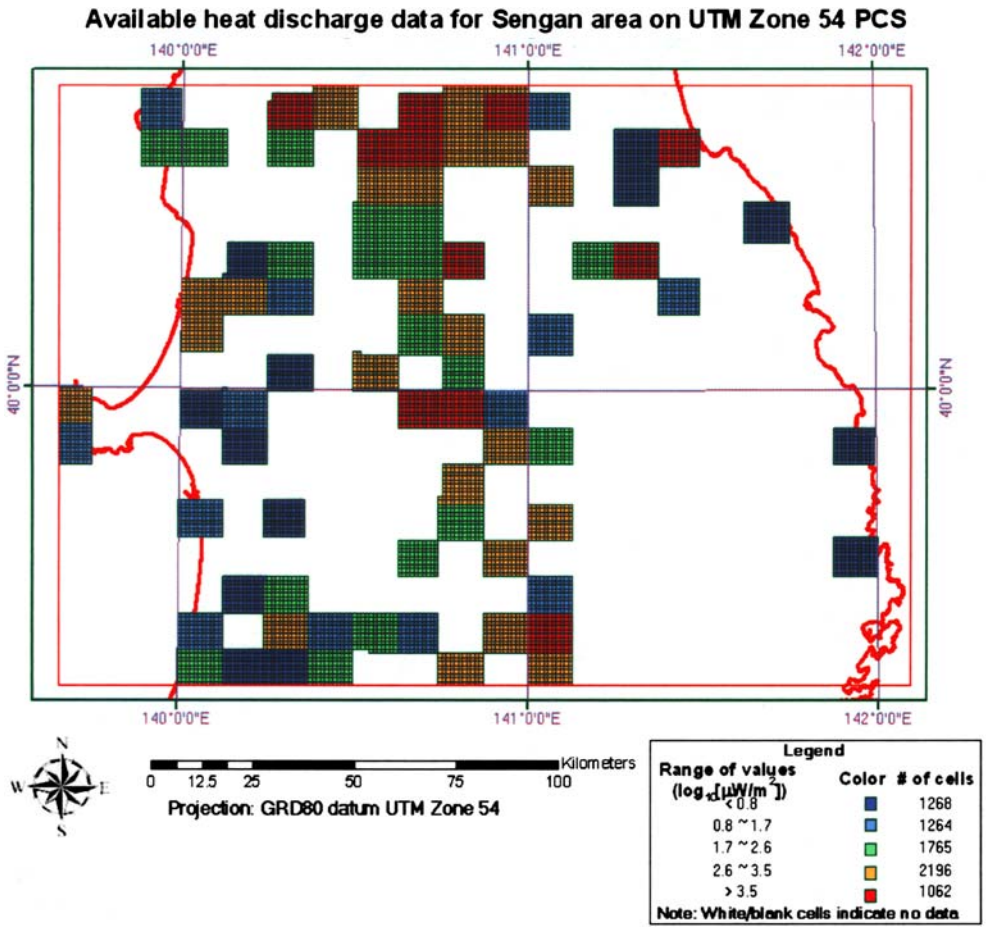


**Fig. 7** Available heat discharge data for Sengan region on a 1 km square grid using UTM zone 54 PCS

ArcGIS Geostatistical Analyst (Johnston et al. 2001) was used to perform kriging. For the last two variables, a more simple re-mapping analysis was done to estimate the geologic variable value for each 1 km cell in the study area. Through this way, geologic variable value estimation was completed for all 23949 one km cells in the study area for each of the six geologic variables. The results obtained for groundwater temperature and presence of volcanic rocks are given below to illustrate the two different procedures.

## Groundwater temperature

In the extended data base region, groundwater temperature data were available for 999 one km square cells. These data were used to obtain the variogram surface shown in Fig. 10. The variogram surface plot shows that the major axis direction for correlation distance is around N 5–15° W. Figure 11a shows the variogram (1 km lag spacing) calculated for groundwater temperature assuming isotropic spatial variation (omni directional) in the considered region. Directional variograms were calculated at every 15° counterclockwise starting at 0° (East) using a lag spacing of 1 km (Fig. 11b through m). The following directional parameters were used in calculating the directional variograms: angular tolerance = 22.5° and maximum bandwidth = 3× lag spacing (Pannatier 1996).

Directional variograms clearly show the presence of anisotropy. Exponential, spherical, Gaussian and power functions were considered in fitting the directional variograms. An exponential function turned out to be the best fit according to

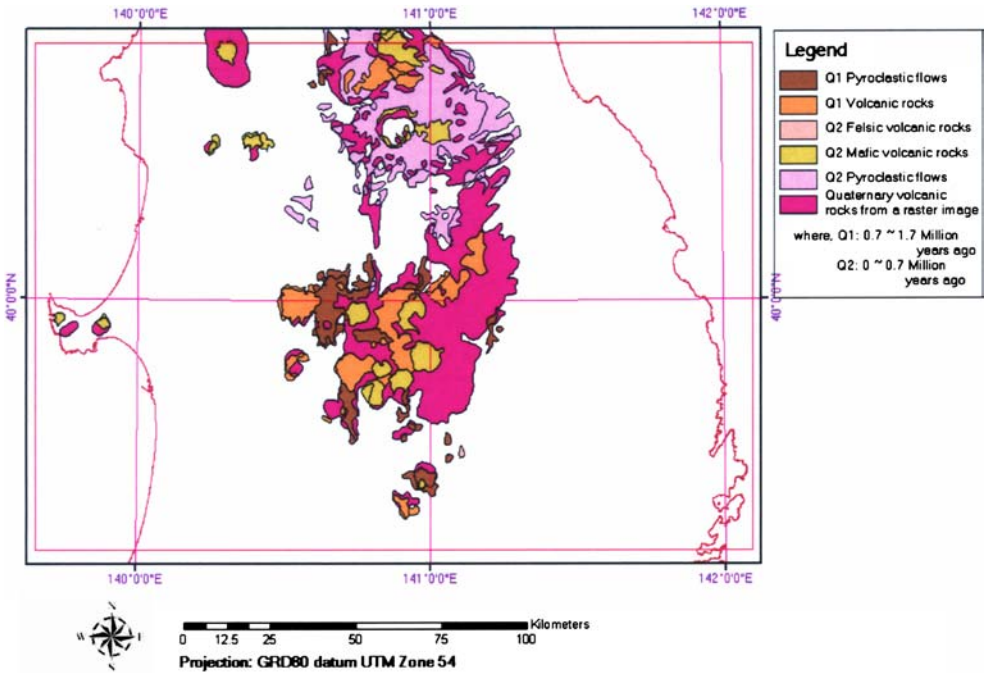


**Fig. 8** Polygonal coverage of available volcanic rock areas in Sengan region from two different sources

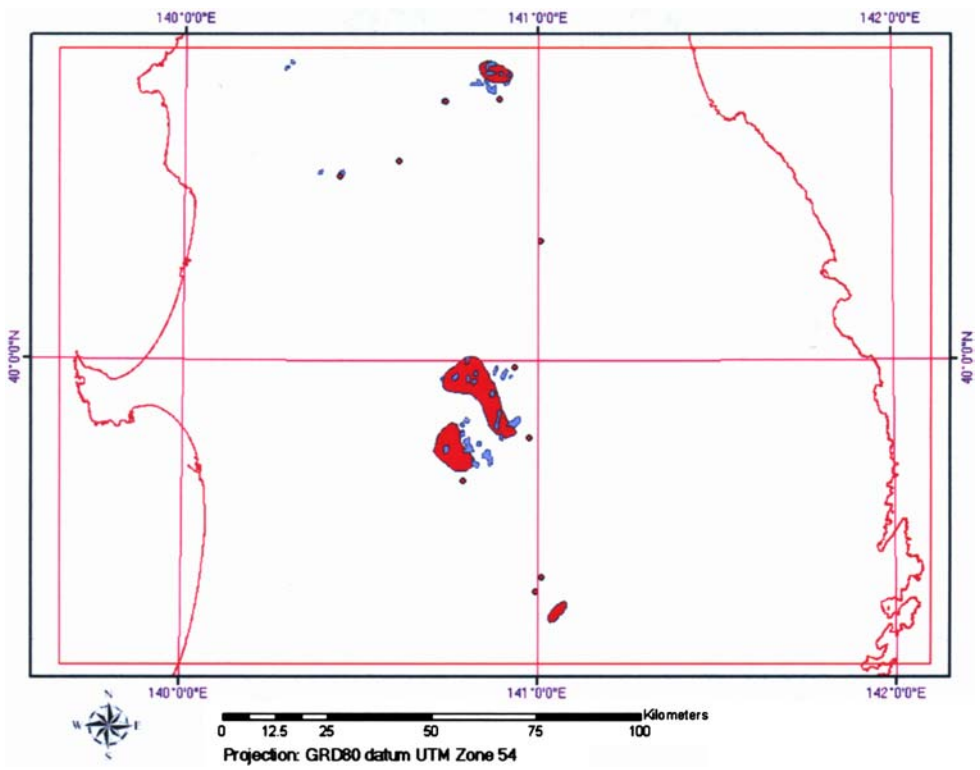


**Fig. 9** Polygonal coverage of available hydrothermal alteration areas in Sengan region from two different sources

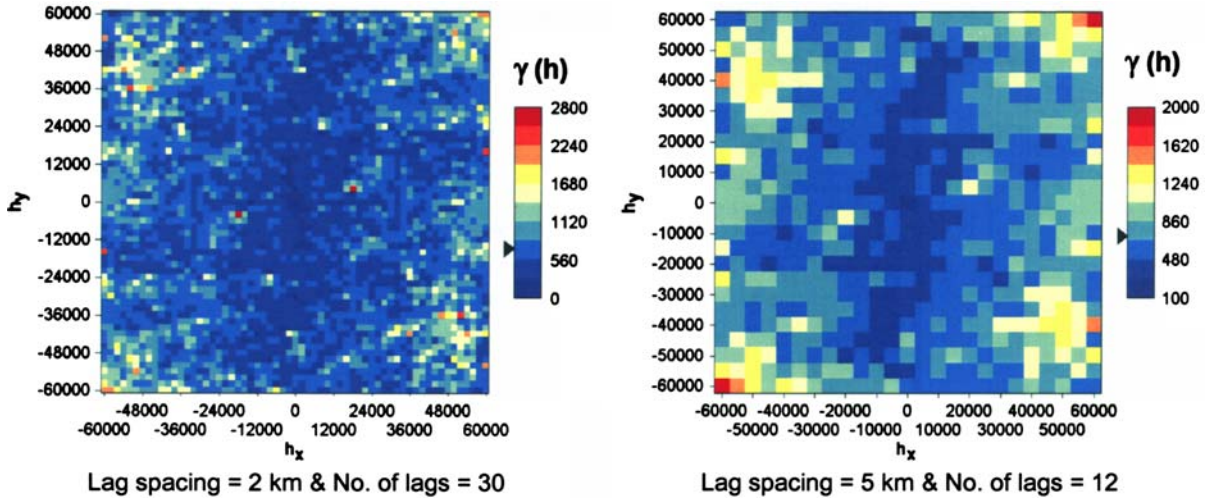


**Fig. 10** Variogram surface plots of groundwater temperature at different lag spacings and number of lags. Note: the black triangle shown on the color legend indicates the total variance value

the goodness of fit test results. The obtained variogram models for directional variograms are shown in Figs. 11b through m. Note that the nugget and partial sill values for the variograms were found to be 120 and 600, respectively. Correlation distances were found to vary with the direction.

The correlation distance obtained for every 15-degree is plotted on a polar coordinate system in Fig. 12. The least square elliptical fit obtained for the directional correlation distance is also shown in the same figure. Note that the fit is quite close to the calculated directional correlation distances. The figure also shows the correlation distance obtained through the assumption of isotropic spatial variation (through omni directional variogram). This figure clearly shows that anisotropy cannot be neglected in modeling the spatial variation of groundwater temperature. Major axis direction, and semi-major and semi-minor axis lengths obtained for the correlation distance model are also given in the figure.

The variogram model was used to perform kriging. Figure 13 shows the map obtained for the predicted values of groundwater temperature. The map obtained for the standard error (the square root of the kriging variance) of the prediction is shown in Fig. 14. Note that the standard error (uncertainty) increases in areas with a lack of data.

### Presence of volcanic rocks

To express the presence of volcanic rocks in the 1 km square grid system, the 1 km grid mesh was superimposed on the combined map obtained in Section 'Presence of quaternary volcanic rocks.' For each 1 km cell that was fully within the volcanic rock area, a value of one was assigned to the cell. For each 1 km cell that was fully out of the volcanic rock area, a value of 0 was assigned to the cell. If only a portion of 1 km cell fell within the volcanic rock area, the cell was assigned a value between 0 and 1 according to the proportional coverage of the cell area by the volcanic rock area. This final map obtained is shown in Figure 15.

## Multivariate classification (typification) of volcanism for Sengan region

### General features of cluster analysis (CA)

Cluster analysis is a technique designed to perform classification by assigning observations to groups or ''clusters'' so each group is more or less homogeneous and distinct from other groups. CA procedures can be separated into four general types (Sneath and Sokal 1973; Gordon 1999): (1) Partitioning methods, (2) Arbitrary origin meth-

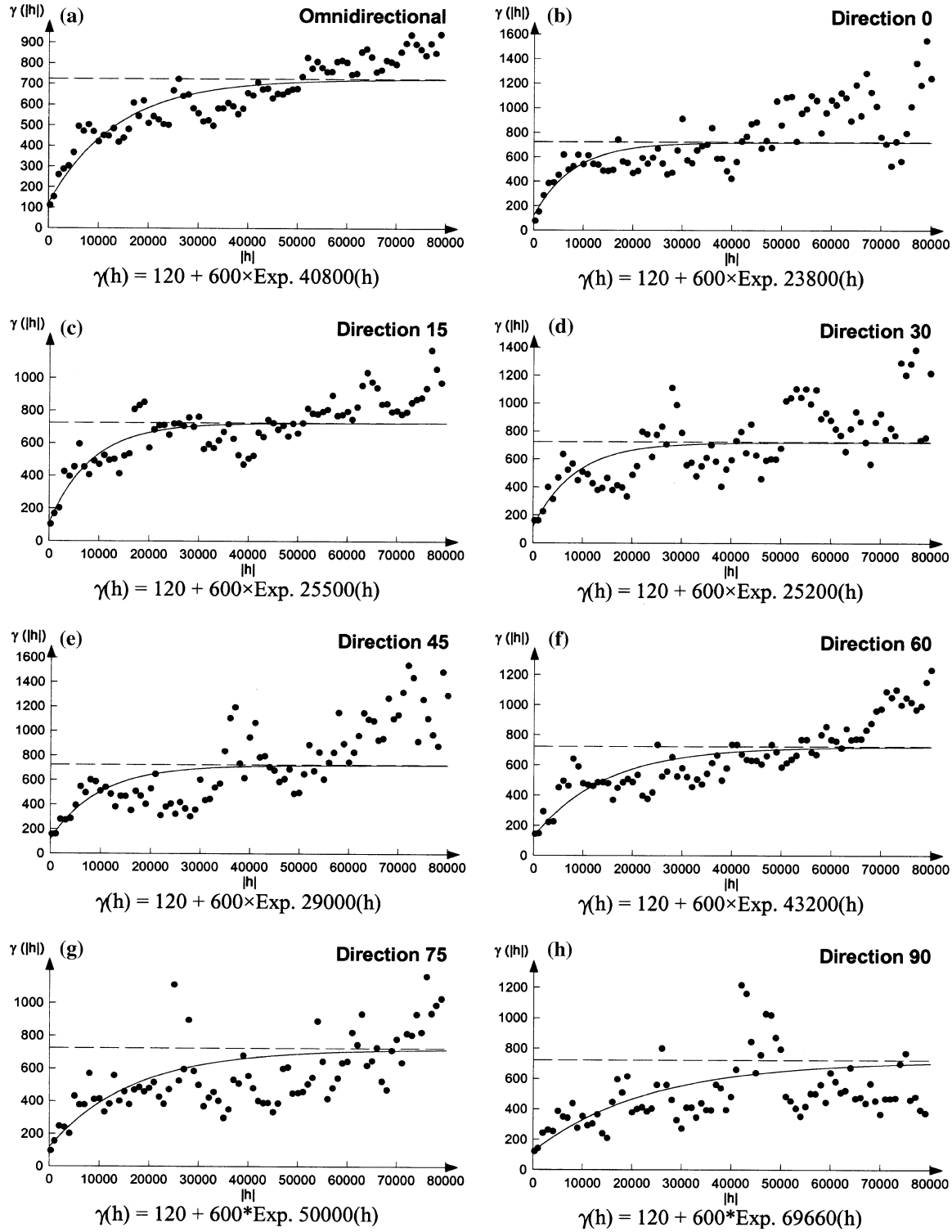


**Fig. 11** (**a**) Omni directional variogram and (**b–m**) directional variograms for groundwater temperature at 1 km lag spacing

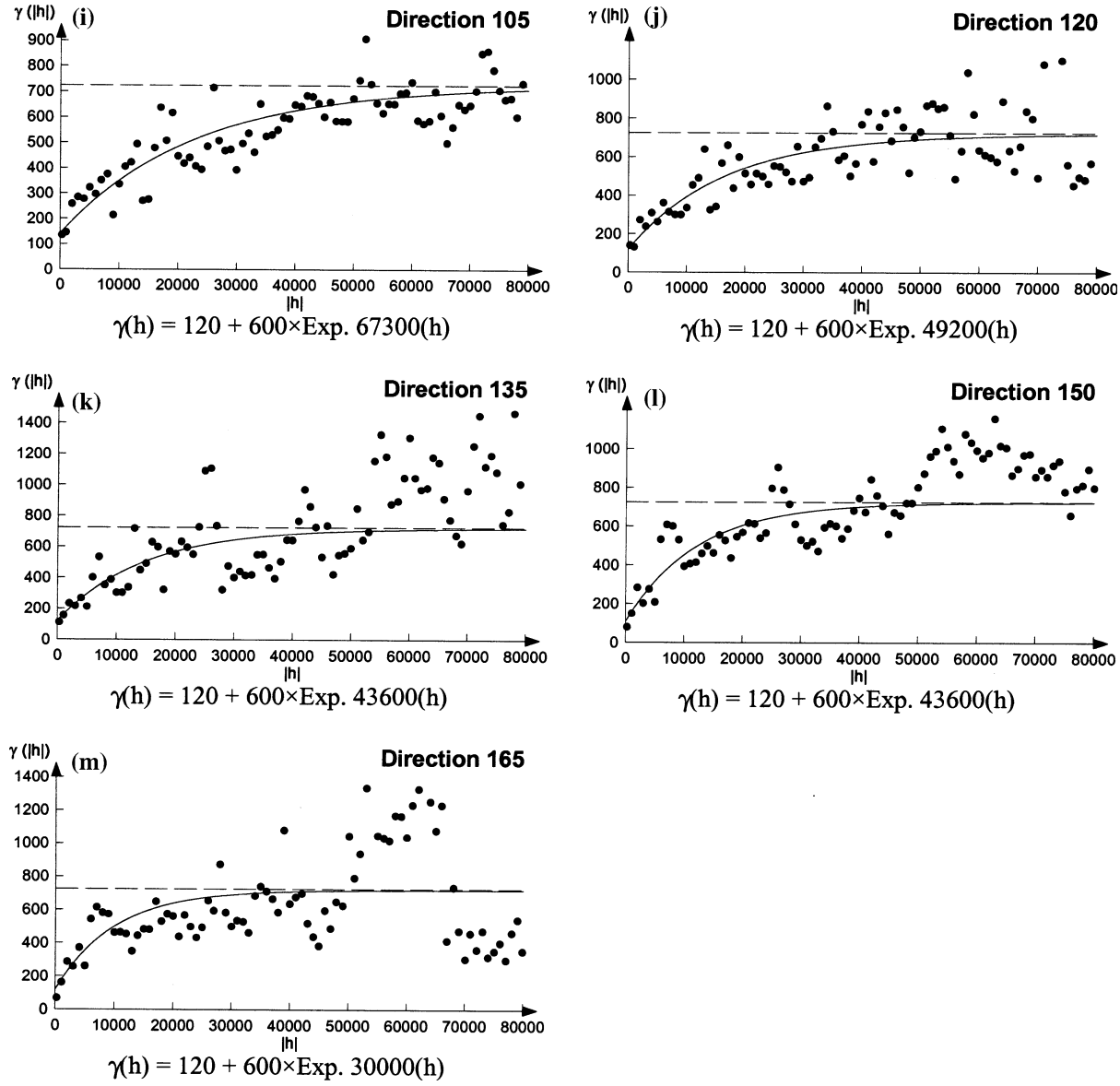


**Fig. 11** continued

ods, (3) Mutual similarity procedures and (4) Hierarchical clustering. The hierarchical clustering technique is the most popular technique in earth sciences. Therefore, some details of this technique are given below.

Consider $n$ objects having $m$ measurable characteristics. The observations will form an $n\times m$ data matrix, **X**. Some measure of resemblance or similarity is computed between every pair of objects; that is, between every pair of rows of the data matrix. A popular similarity measure between objects is a standardized $m$-space Euclidean distance, $d_{ij}$. This is computed by

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}{m}} \quad (1)$$

where, $X_{ik}$ denotes the $k$th variable measured on object $i$ and $X_{jk}$ is the $k$th variable measured on object $j$. In all, $m$ variables are measured on each object, and $d_{ij}$ is the distance between

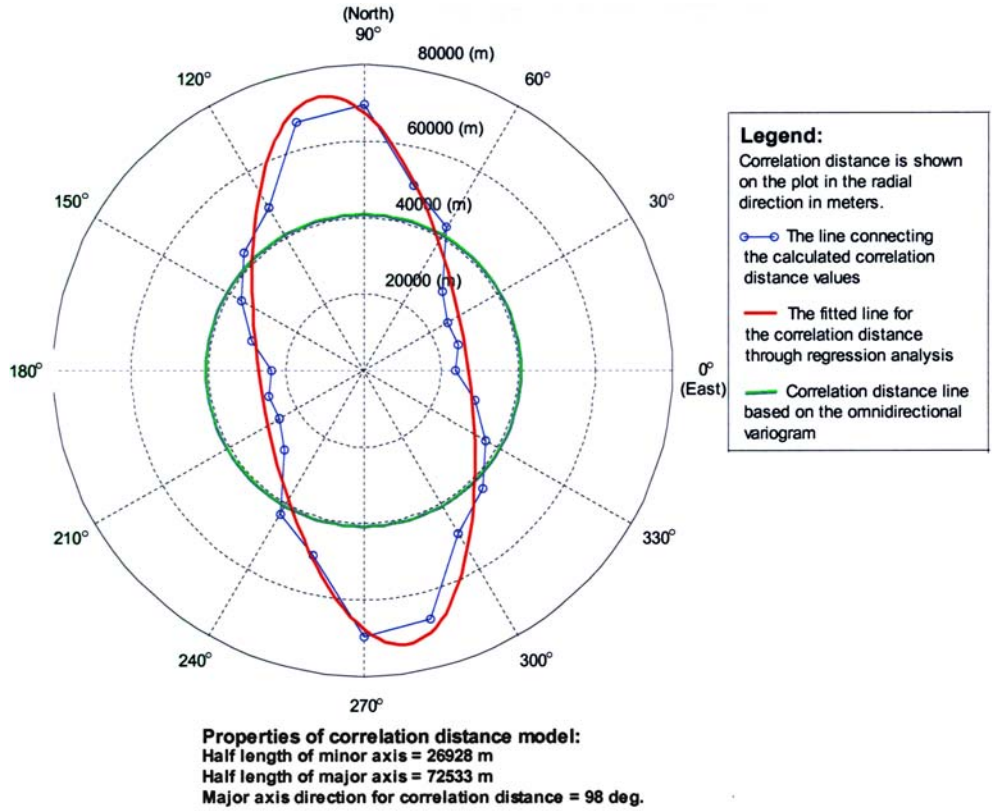


**Fig. 12** Variation of correlation distance with direction for groundwater temperature

object *i* and object *j*. A small distance indicates the two objects are similar, whereas a large distance indicates dissimilarity. Usually, to weigh each variable equally and to remove the effects of different units of measurement across the different variables, each element in the data matrix **X** is standardized by subtracting the column means and dividing by the column standard deviations prior to computing $d_{ij}$. In other words, each value in the data matrix is expressed as a deviation from the mean in terms of a rational number of standard deviations.

Computation of a similarity measurement using the Euclidean distance between all possible pairs of objects produces an $n \times n$ symmetrical matrix, **C**. Each coefficient $c_{ij}$ in the matrix indicates the resemblance between objects *i* and *j*. Next, the objects are arranged into a hierarchy so that objects with the highest mutual similarity are placed together to form groups or clusters. Then the groups having closest resemblance to other groups are connected together until all of the objects are placed into a classification scheme named as a dendrogram. Different procedures are available in the literature to form these groups or clusters (Sneath and Sokal 1973; Backer 1995; Gordon 1999).

Although several measures of similarity have been proposed in the literature, only two are widely used: the Euclidean distance and the correlation coefficient in performing cluster analysis. If the raw data are standardized as explained above prior to computing the similarity coefficient, the correlation coefficient and Euclidean distance can be directly transformed from one to another. Dendrograms constructed from the two measures generally are similar (Davis 2002). However, the Euclidean distance is not constrained within the range plus or minus one as is the correlation coefficient, so it may produce more effective dendrograms if a few of the objects are very dissimilar from the other as in the case of volcanism range.

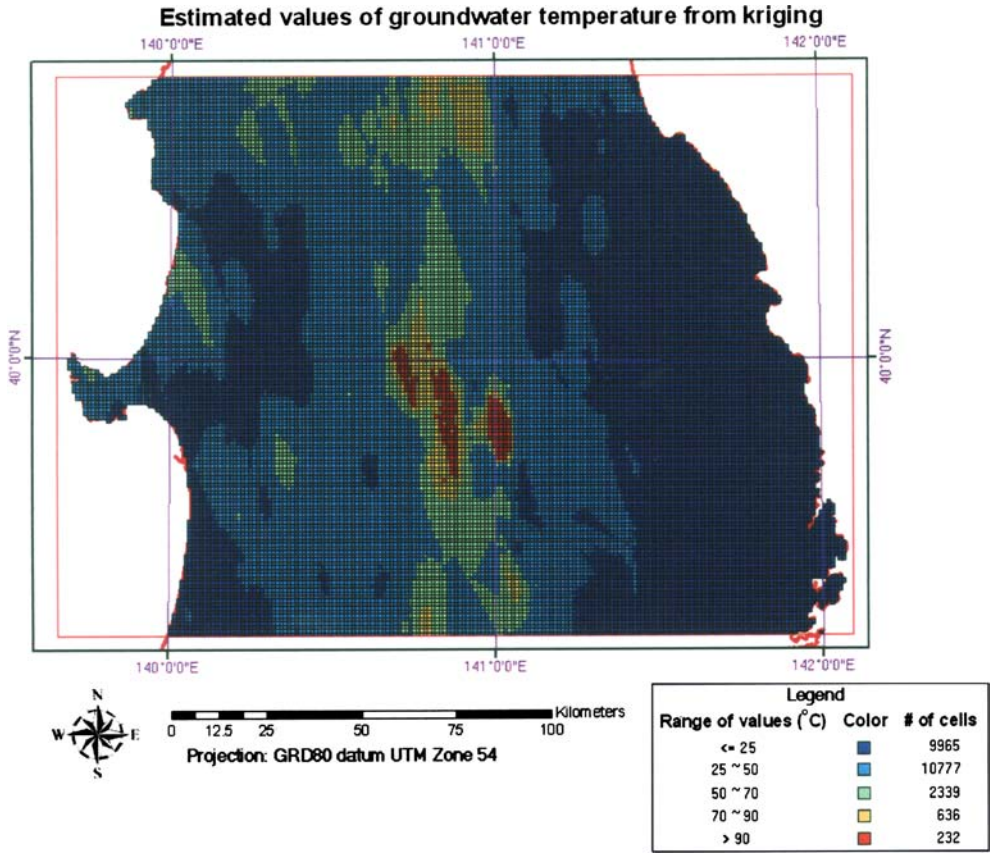


**Fig. 13** Estimated values of groundwater temperature from kriging for the study area

Multivariate classification of volcanism for the Sengan region through CA

The land area of Sengan is divided into 23949 one km square cells (see Section 'Selected coordinate system and the grid system to show volcanic and geologic variable data for Sengan region'). In Section 'Geostatistical analysis,' at the center of each cell, values were calculated for each of the following geologic variables: groundwater temperature, geothermal gradient, heat discharge, groundwater pH value, presence of volcanic rock and presence of hydrothermal alteration. These variables were identified as the variables most important to volcanism. This means that to perform cluster analysis a complete vector of values for six geologic variables at 23949 one km cell center locations (i.e. cases or objects) are available. The Tree clustering method available in STATISTICA software package (StatSoft 1997) cannot be used for more than 300 cases. However, the K-Mean clustering method available in SAS (2002) can be used to perform cluster analyses for 23949 cases. Therefore, first the K-Mean clustering method was performed to reduce 23949 cases to 300 groups. A complete variable mean vector was obtained for each group. Then these 300 groups were treated as cases and Tree clustering was performed using the STATISTICA software package. Figure 16 shows the dendrogram obtained for the final 50 groups. Figure 16 shows clearly that the number of groups selected can be lowered by increasing the value chosen for the Euclidean distance. As an example, it is possible to reduce 23949 cases to five groups by selecting a value of about 2.65 for the Euclidean distance. If needed, the number of groups can be reduced further to three by selecting a value slightly lower than 3.0 for the Euclidean distance (Fig. 16). This

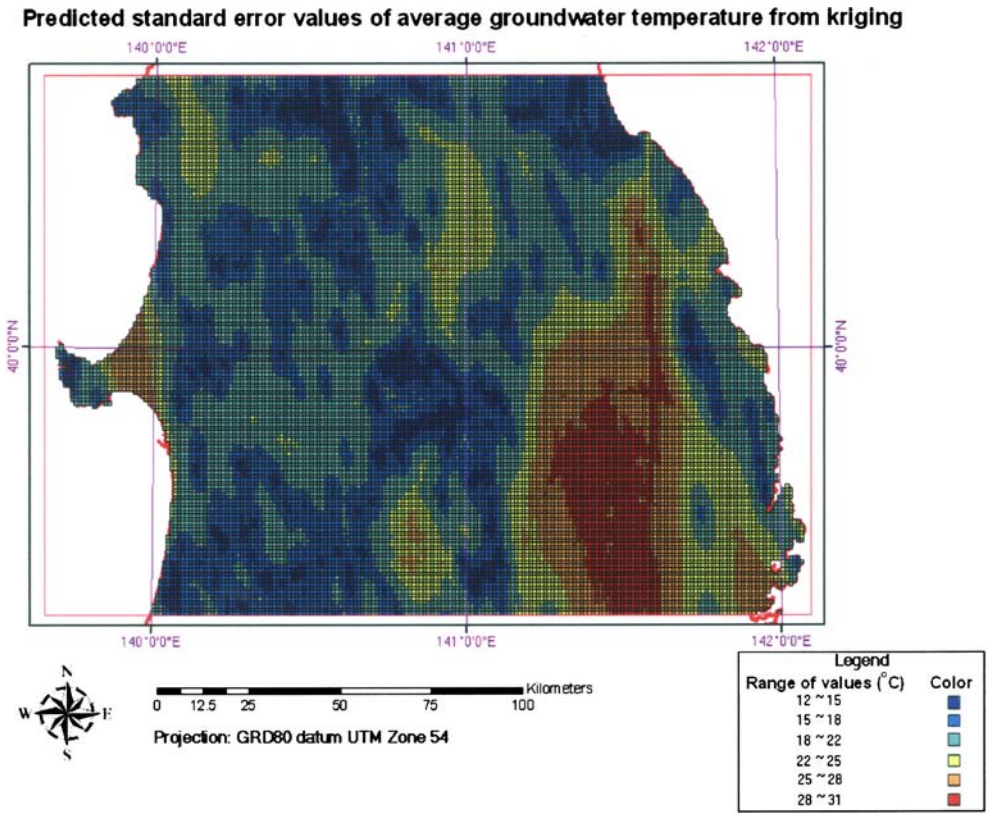


**Fig. 14** Predicted standard error values of average groundwater temperature from kriging for the study area

means the analyst or the decision maker has the flexibility to select any number of groups as he/she desires in performing the regionalized mapping. The mean values obtained for the geologic variable vector for the five groups selected are shown in Fig. 17. The same figure provides the number of cases obtained for each group. Note that in Fig. 17, the cluster numbers are arranged such that the volcanism level moves from a lowest possible volcanic disruption (Cluster 1) to a highest possibility of volcanic disruption (Cluster 5).

## Regionalized mapping of volcanism for Sengan region

### Concepts of regionalized mapping

The discussion in this section is a summary of Bohling (1997). Let us say that the number of most important geologic variables used in the CA is $v$. These $v$ variables form the variable vector $\mathbf{x}$ for a sample. Let us assume that the number of classes or groups defined at the end of the typification step is $g$. Then the geologic variable data coming from a 1 km cell center location can be assumed to come from one of the $g$ different groups, each having a specific probability density function, $f_i(\mathbf{x})$, where $i$ stands for the group number. If the probability of sampling from the $i$th group is $q_i$, then $\sum_{i=1}^{g} q_i = 1$. Note that $q_i$ is the prior probability (probability of occurrence based on prior knowledge). If $\mathbf{x}$ is known or given for the sample, then according to Bayes' theorem, the posterior probability of the sample coming from $i$th group is given by

$$p(i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum_{j=1}^{g} q_j f_j(\mathbf{x})} \tag{2}$$

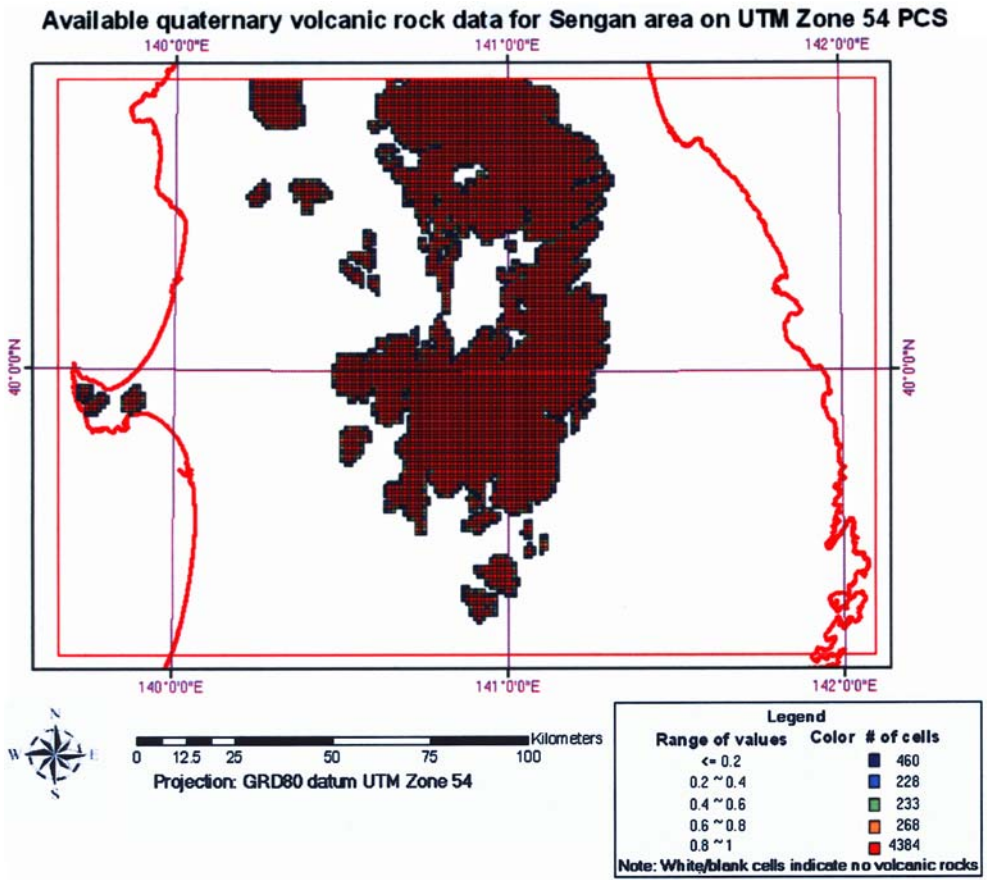


**Fig. 15** Available quaternary volcanic rock data for Sengan region on 1 km square grid using UTM zone 54 PCS

The $p(i|\mathbf{x})$ should be calculated for each group. The sample is then allocated to the group with the highest $p(i|\mathbf{x})$ value. To calculate $p(i|\mathbf{x})$ it is necessary to know all $f_i(\mathbf{x})$ and $q_i$ values. A number of parametric or non-parametric methods are available to model $f_i(\mathbf{x})$ (Mclachlan 1992; SAS 1989). However, usually the discriminant analysis assumes that the groups follow multivariate normal distributions. If the mean vector and the covariance matrix for group $i$ are denoted by $\mu_i$ and $\Sigma_i$, $f_i(\mathbf{x})$ can be given as

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp\left(-0.5d_i^2(\mathbf{x})\right) \tag{3}$$

where

$$d_i^2(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \tag{4}$$

is the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}_i$. Substituting Eqs. (3) into (2) and canceling the constant factor $(2\pi)^{-p/2}$ yields

$$p(i|\mathbf{x}) = \frac{q_i|\Sigma_i|^{-1/2}\exp\left(-0.5d_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g} q_j|\Sigma_j|^{-1/2}\exp\left(-0.5d_i^2(\mathbf{x})\right)} \tag{5}$$

$$= \frac{\exp\left(-0.5D_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g}\exp\left(-0.5D_j^2(\mathbf{x})\right)} \tag{6}$$

where

$$D_i^2(\mathbf{x}) = d_i^2(\mathbf{x}) + \ln|\Sigma_i| - 2\ln q_i \tag{7}$$

is the generalized squared distance from $\mathbf{x}$ to group $i$ following the usage of SAS (1989). Thus the sample may be allocated to the proper group either on the basis of maximum posterior proba-

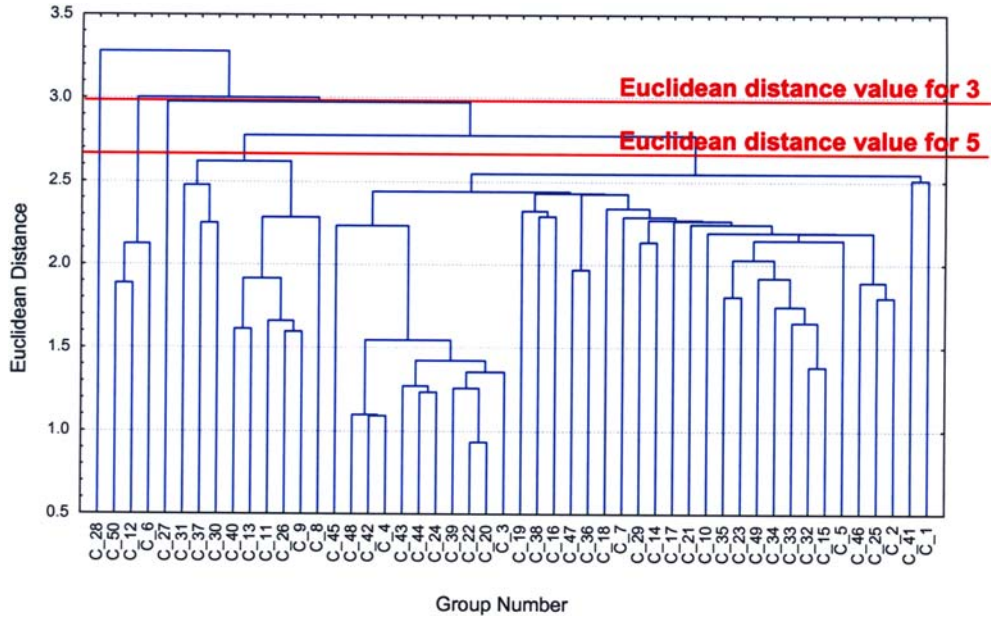


**Fig. 16** Dendrogram obtained for the final 50 groups

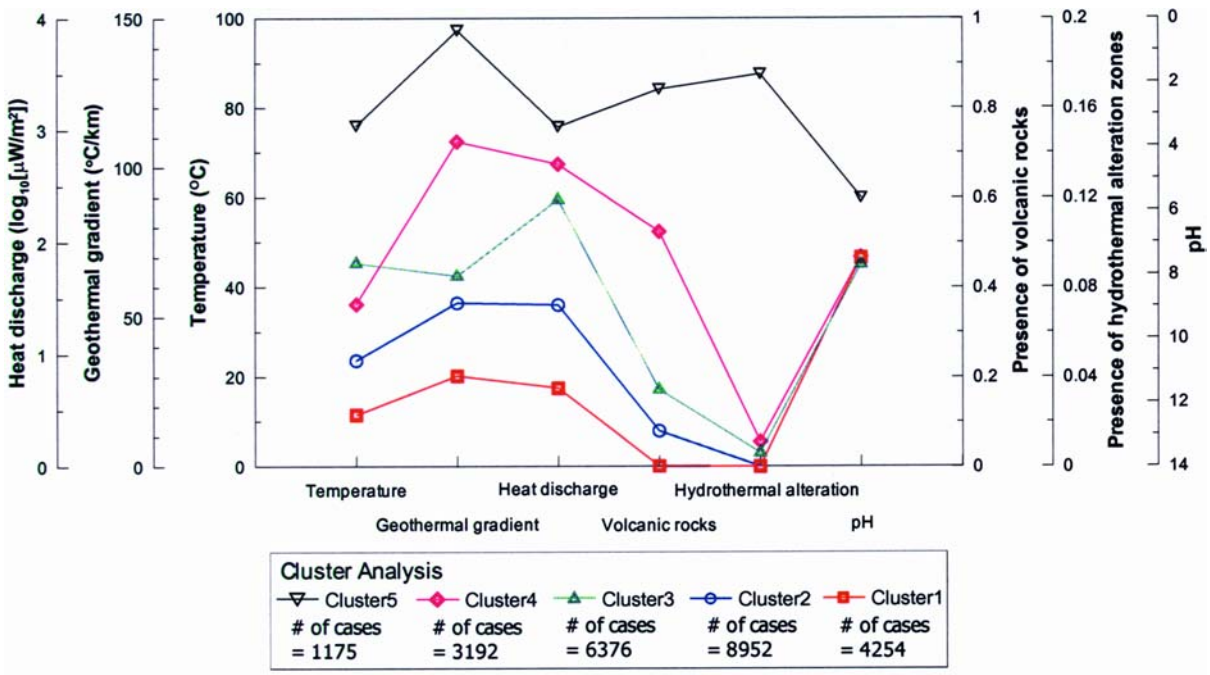


**Fig. 17** Mean variable vector for each cluster for the 5 group clustering

bility or minimum generalized squared distance, yielding equivalent results.

Results of the typification step allocate the total number of 1 km cell center locations with complete variable vectors to a finite number of distinct groups. This information can be used to estimate $q_i$ values according to Eq. (8) given below:

$$q_i = \frac{n_i}{\sum_{j=1}^{g} n_i} \tag{8}$$

where $n_i$ is the number of samples belonging to group $i$. The geologic variable data available for each group defined at the end of the typification step can be used in Eqs. (9) and (10) to estimate $\mu_i$ and $\Sigma_i$, respectively.

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \frac{\boldsymbol{1}}{\boldsymbol{n_i}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{n_i}} \mathbf{x}_{\boldsymbol{k}} \tag{9}$$

and

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} (\mathrm{x}_k - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}})(\boldsymbol{x_k} - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}})' \tag{10}$$

When the group covariance matrices are not assumed to be equal (i.e. different $\Sigma_i$ values) and $q_i$ is calculated according to Eq. (8), the analysis performed through Eqs. (6) and (7) is termed quadratic discriminant analysis.

If it can be assumed that the prior probabilities are equal, $q_i$=1/$g$, and that the groups have a common covariance matrix, $\Sigma_i$=$\Sigma$, then Eqs. (4) and (5) reduce to Eqs. (11) and (12) given below, respectively.

$$d_i^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu_i})' \Sigma^{-\boldsymbol{1}} (\mathbf{x} - \boldsymbol{\mu_i}) \tag{11}$$

$$p(i|\mathbf{x}) = \frac{\exp\left(-0.5 d_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g} \exp\left(-0.5 d_j^2(\mathrm{x})\right)} \tag{12}$$

The implication of Eqs. (11) and (12) is the sample allocation for this case can be done based only on the Mahalanobis distance. The discriminant analysis under the case of common covariance matrix is known as the linear discriminant analysis. For linear discriminant analysis, $\Sigma$ is estimated by the pooled within-groups covariance matrix given as

$$\hat{\Sigma} = \frac{1}{n - g} \sum_{i=1}^{g} (n_i - 1) \hat{\Sigma}_i \tag{13}$$

where $n$ is the total number of data. Statistical tests for equality of covariance matrices given in Anderson (1984) and McLachlan (1992) can be used to determine which discriminant method is most appropriate for the available data. Even if the results of the statistical test indicate quadratic discriminant analysis is the appropriate method, use of linear discriminant analysis has been shown to produce equally acceptable results (Bohling et al. 1990). In addition, with respect to deviations from normality, linear discriminant analysis tends to be more robust than the quadratic discriminant analysis.

## Results of regionalized mapping for the Sengan region

In Section 'Multivariate classification (typification) of volcanism for Sengan region,' a typification was performed for five groups. Both the quadratic and linear discriminant analyses were performed using these five groups. Negligible differences were found between the quadratic and linear discriminant analysis results. Values of $q_i$ calculated based on Eq. (8) and used for the quadratic discriminant analysis are given in Table 2. Probability membership maps obtained for the five groups are shown in Fig. 18a through e. The higher probability a location has in these maps, the higher confidence in that location belonging to the allocated group. The locations with lowest probabilities in each map indicate that those locations have almost a similar chance of belonging to the adjacent regionalized group having either lower or higher level of volcanism. In addition, the locations with lowest probabilities indicate demarcation boundaries among the different volcanism groups. Figure 19 shows the predicted regions for the five groups of volcanism

**Table 2** Values of $q_i$ used for quadratic discriminant analysis for 5 group clustering

| Group number | Number of data | Prior probability |
|---|---|---|
| 1 | 4254 | 0.1776 |
| 2 | 8952 | 0.3738 |
| 3 | 6376 | 0.2662 |
| 4 | 3192 | 0.1333 |
| 5 | 1175 | 0.0491 |
| Total | 23949 | 1.000 |

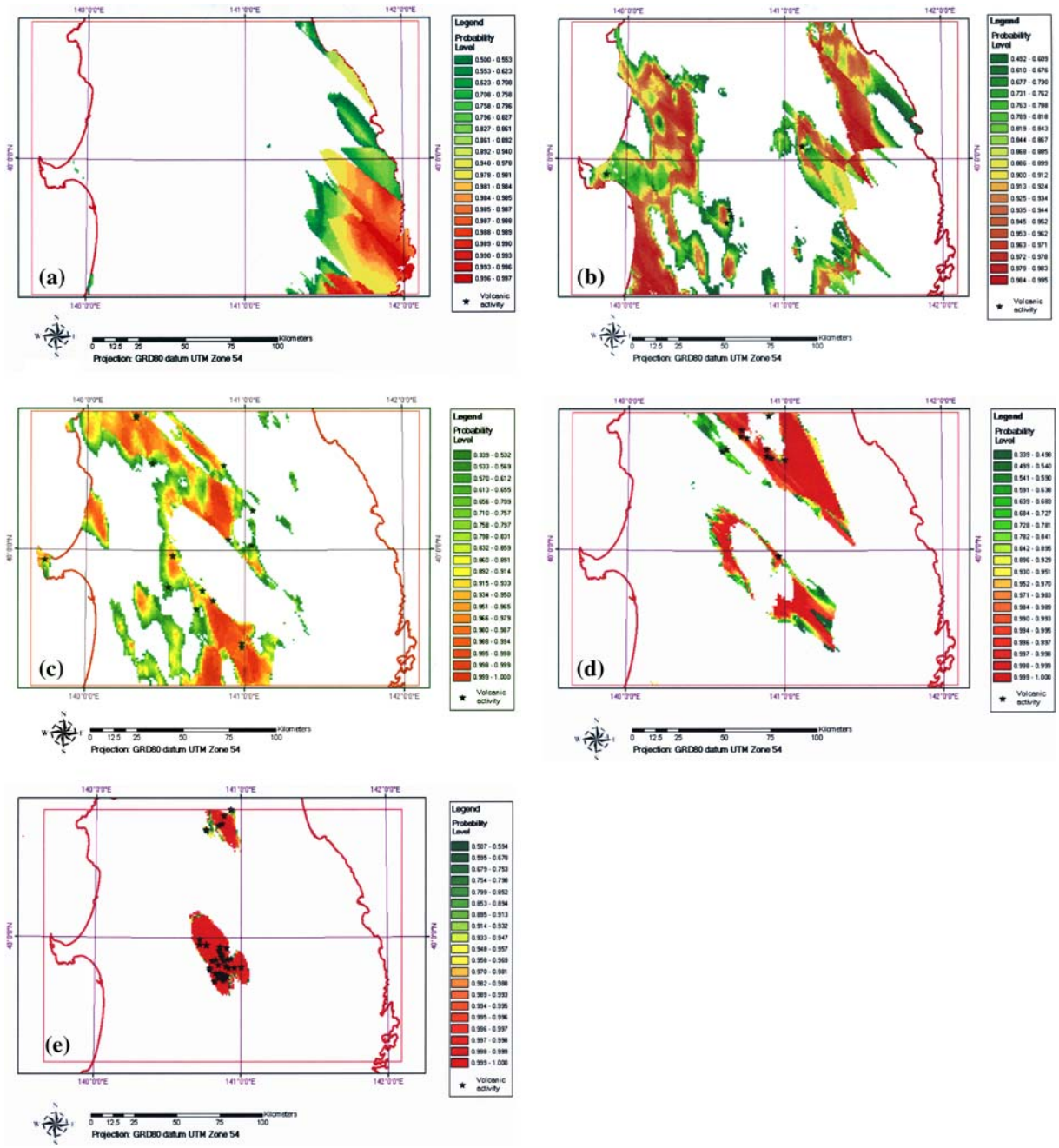


**Fig. 18** (**a**) Regionalized membership probability distribution for group 1 of 5 clusters. (**b**) Regionalized membership probability distribution for group 2 of 5 clusters. (**c**) Regionalized membership probability distribution for group 3 of 5 clusters. (**d**) Regionalized membership probability distribution for group 4 of 5 clusters. (**e**) Regionalized membership probability distribution for group 5 of 5 clusters

on one plot along with the recorded volcano data for comparison for Sengan region. In some cells more than one volcanic event appears. In such a situation, it is difficult to see more than one data point appearing in a cell. Table 3 shows the distribution of recorded volcanic data among the 5 different groups of volcanism. This table provides the volcanic data information under two different

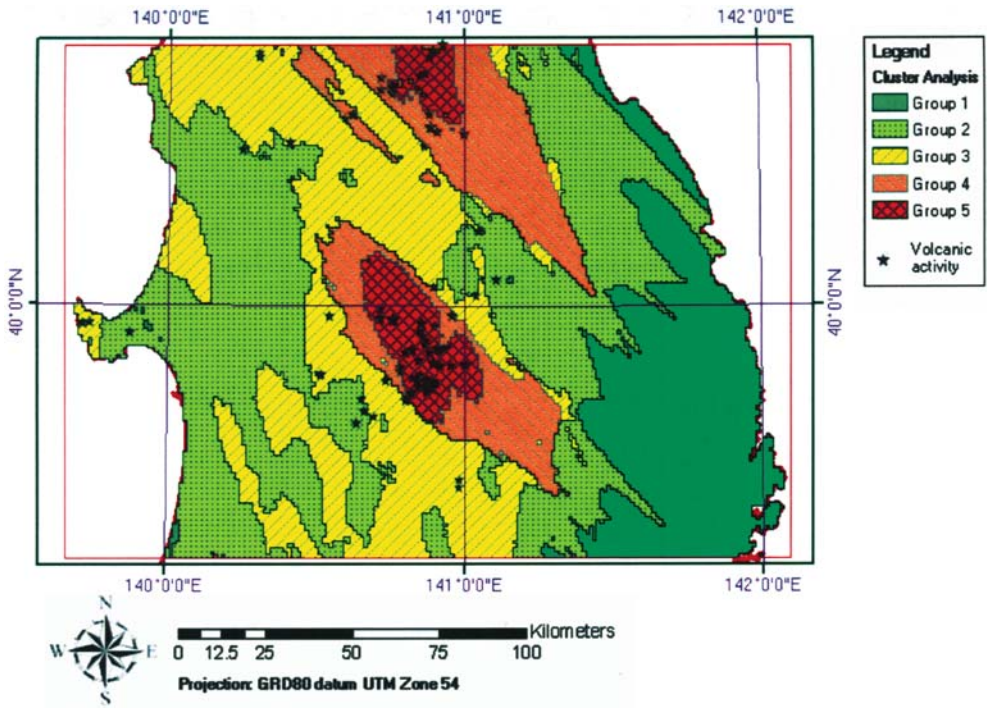


**Fig. 19** Predicted regions for the 5 groups of volcanism along with the recorded volcanoes for the Sengan region

rows: (a) number of data and (b) number of cells. The differences in values between the two rows provide the information on overlapping data that appear within certain cells that are difficult to distinguish on the shown figures. Note that none of the recorded volcanic data are located in the lowest volcanic susceptibility region (group 1). Sixty seven cells out of a total of 73 cells are located in high and moderate volcanic susceptibility regions (groups 3, 4 and 5). The number of recorded volcanoes increases as the probability of volcanism increases with group, as would be expected. Note that the volcano locations were not used in the definition of the multivariate classes or in the mapping. These locations are used solely for verification purposes. The aforementioned observations show that the regionalized mapping technique used for estimation of volcanism susceptibility has worked very well. For each regionalized volcanism group, the mean probability of a volcanic event taking place may be estimated by dividing the total number of volcanic events that have occurred in the considered regionalized group by the total monitoring time of the volcanic activities. The reciprocal of this said probability provides the return period for a volcanic activity for each regionalized group.

**Table 3** Distribution of volcanic data locations among the five different groups

| Volcanic data | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 |
|---|---|---|---|---|---|
| # of points | 0 | 11 | 18 | 24 | 60 |
| # of cells | 0 | 6 | 11 | 16 | 40 |

## Uncertainty evaluations of regionalized mapping

The entropy of classification given by Eq. (14) has been suggested in the literature (Jaynes 1957; Kitanidis 1994) to assess the uncertainty of the allocation process of the sample based on the calculated posterior probabilities.

$$H = \left( -\sum_{k=1}^{g} p_k \ln p_k \right) / \ln g \qquad (14)$$

In Eq. (14), $H \rightarrow 0$ when $p_k \rightarrow 1$ for any $k$ and $H$ reaches its maximum value of one when all the posterior probabilities are equal. Therefore, $H$ ranges between 0 and 1 with larger values indicating greater uncertainty. The entropy measure accounts for the entire set of posterior probabilities as given by Eq. (14). Mapping the entropy as calculated in Eq. (14) can be used to improve the spatial definition of group boundaries (Bohling 1997; Olea 1999).

The obtained results for entropy are shown in Fig. 20. Each recorded volcanic location along with the volcanic activity group it belongs to is also shown in the same figure. The histogram given in Fig. 21 shows the uncertainty level of the group estimations at different locations where volcanic data are available for each group. Figures 20 and 21 show that the uncertainty of the mapping estimations is relatively low on the average for volcanic data cell locations that are in the high volcanism regions (groups 4 and 5). The same two figures show that the uncertainty of the mapping estimations is relatively high on the average for volcanic data cell locations that are in the low volcanism region (group 2). Note that no volcanic data exist in the lowest volcanism region (group 1). Figures 20 and 21 also show that the uncertainty of the mapping estimations is relatively moderate on the average for volcanic data cell locations that are in the moderate volcanism region (group 3). It is recommended to collect more geologic data in the regions where the uncertainty level is high. The new data collected can be added to the old database to perform future regionalized mapping to reduce the uncertainty level of the estimations.

## Summary and conclusions

This paper provides the procedures used and results obtained for the study on hierarchical probabilistic regionalization of volcanism for Sengan region in Japan. The summary of the steps used along with the results obtained and the conclusions arrived at are given below:

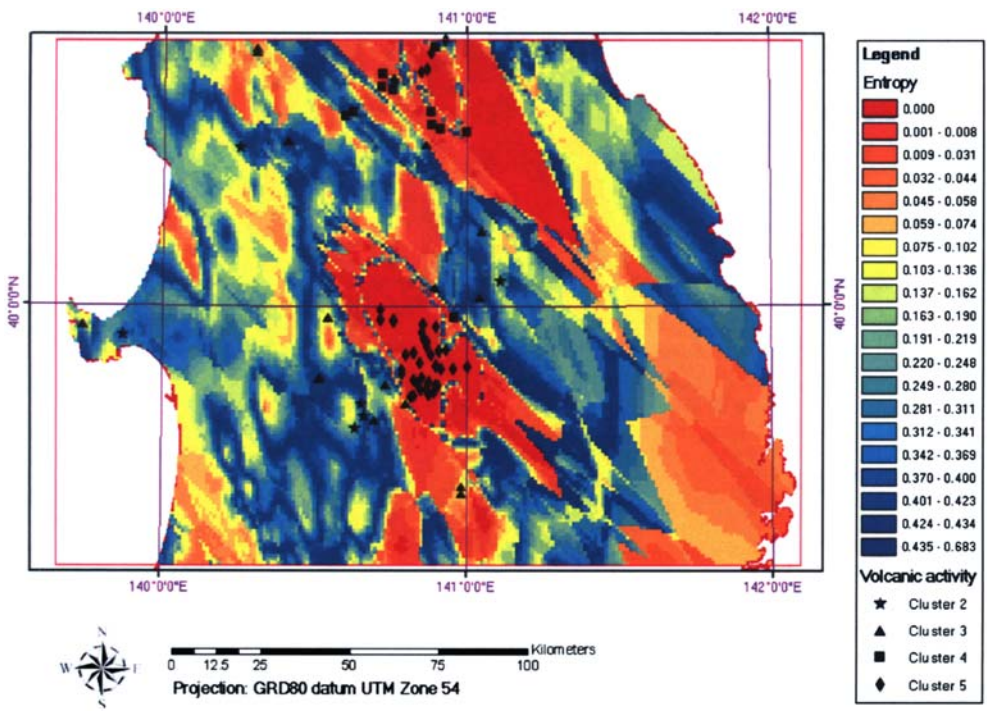


**Fig. 20** Regionalized distribution of entropy for 5 group clustering along with the available volcanism data located in different clustering groups

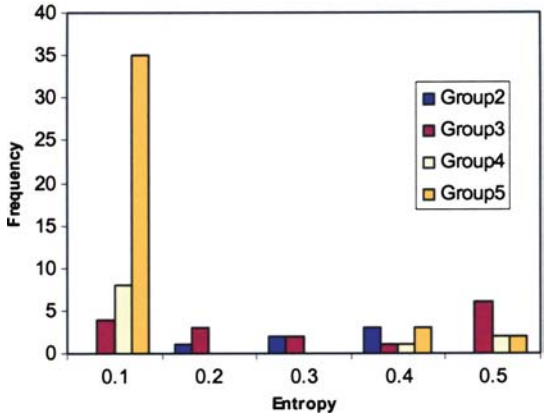


**Fig. 21** Histograms of volcanic data located in different regionalized groups and the associated uncertainty level for entropy

(1) UTM zone 54 projected coordinate system and a 1 km square regular grid system were selected to show the available volcanism and geologic data for Sengan region using ArcGIS 8.x software package.
(2) The map obtained for each geological variable was visually compared with the map of recorded volcanism to determine the geologic variables that are strongly correlated to volcanism. The variables: geothermal gradient, groundwater temperature, heat discharge, groundwater pH value, presence of volcanic activity and presence of hydrothermal alteration were labeled as the most important variables for volcanism.
(3) For each of the most important geologic variables connected with volcanism, directional variogram modeling and kriging were performed on available data to estimate values at the centers of 23949 one km square cells. These estimated values formed 23949 cases of complete variable vectors.
(4) Cluster analysis was performed on the 23949 complete variable vectors to classify them to five groups of potential volcanism spanning from lowest possible volcanism to highest possible volcanism with increasing group number.
(5) Volcanism group results obtained through cluster analysis were used with Bayes' theorem and discriminant analysis to construct maps showing the probability of group membership for each of the volcanism groups obtained in step four. These maps show good comparisons with recorded volcanism of the Sengan region. Note that no volcanic data exist in the group 1 region. The high probability areas (i.e the lowest uncertainty) within group 1 have the chance of being the no volcanism region.
(6) An entropy map was constructed to express uncertainty levels of the regionalized mapping estimations. The recorded volcanism data are also plotted on the same map to see the uncertainty level of the estimations at the locations where volcanism exists. The volcanic data cell locations that are in the high volcanism regions (groups 4 and 5) show on the average relatively low mapping estimation uncertainty. On the other hand, the volcanic data cell locations that are in the low volcanism region (group 2) show on the average relatively high mapping estimation uncertainty. The volcanic data cell locations that are in the medium volcanism region (group 3) show on the average relatively moderate mapping estimation uncertainty. It is recommended to collect more geologic data in regions where the uncertainty level is high. The new data collected can be added to the old database to perform future regionalized mapping to reduce the uncertainty level of the estimations.

**Acknowledgements** The Nuclear Waste Management Organization of Japan (NUMO) provided the data used in this study as well as the funding for this work. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94-AL-85000.

## References

Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York, 359 pp
Anderson TW (1984) An introduction to multivariate statistical analyses, 2nd edn. John Wiley & Sons, Inc., New York, 704 pp

Arnold et al. (2003) Volcanism workshop summary and status report on conceptual models of volcanic influence. Workshop held at Los Alamos National Laboratories, Sept. 2003

Backer E (1995) Computer-assisted reasoning in cluster analysis. Prentice Hall International Ltd., Hemel Hempstead, UK, 367 pp

Bohling GC (1997) GSLIB-style programs for discriminant analysis and regionalized classification. Comput Geosci 23(7):739–761

Bohling GC, Harff J, Davis JC (1990) Regionalized classification: ideas and applications. In: Bachu S (ed) Proceedings of fifth Canadian/American Conference on Hydrology, NWWA, Dublin, Ohio, pp 229–242

Davis JC (2002) Statistics and data analysis in geology, 3rd edn. John Wiley & Sons, Inc., 638 pp

Deutsch CV, Journel AG (1998) GSLIB: Geostatistical software library and user's guide, 2nd edn. Oxford University Press, New York, 340 pp

Everitt BS (1993) Cluster analysis, 3rd edn. Edward Arnold, Cambridge, UK, 170 pp

Fernanadez JAM, Vidal CB, Glahn VP (1997) Different classification of the Darss sill data set based on mixture models for compositional data. In: Proceedings of Third Annual Conference of the International Association of Mathematical Geology (IAMG), Part 1, pp 151–155

Gordon AD (1999) Classification, 2nd edn. Chapman & Hall/CRC, Boca Raton Fla., 256 pp

Harff J, Davis JC (1990) Regionalized in geology by multivariate classification. Math Geol 22(5):573–588

Harff JE, Davis JC, Eiserbeck W (1993) Prediction of hydrocarbons in sedimentary basins. Math Geol 25(7):925–936

Harff J, Davis JC, Olea RA (1991) Quantitative assessment of mineral recourses with an application to petroleum geology. Nonrenew Res 1(1):74–84

Harff J, Davis JC, Watney L, Bohling, GC, Wong JC (1989) Regionalization of western Kansas based on multivariate classification of statigraphic data from oil well. Open-File Report. 89–21, Kansas Geological Survey, Univ. of Kansas, Lawrence, Kansas, 26 pp

Harff J, Eiserbeck W, Hoth K, Springer J (1990) Computer-assisted basin analysis and regionalization aid the search for oil and gas, Geobyte, 11–14

Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York, 550 pp

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630

Johnston K, Ver Hoef JM, Krivoruchko K, Lucas N (2001) Using ArcGIS geostatistical analyst. ESRI

Journel AG, Huijbregts CJ (1978). Mining geostatistics. Academic Press, New York, 600 pp

Kitanidis PK (1994) The concept of dilution index. Water Resour Res 30(7):2011–2026

Matheron G (1971) The theory of regionalized variables and its applications. Les Cahiers du Centre Morphologie Mathematique de Fountainbleau, Ecole des Mines, Fontainebleau, France, 211 pp

McLachlan GJ (1992) Discriminant analysis and statistical pattern recognition. John Wiley & Sons, Inc., New York, 526 pp

Moline GR, Bahr JM (1995) Estimating spatial distributions of heterogeneous subsurface characteristics by regionalized classification of electrofacies. Math Geol 27(1):3–22

Olea R (1999) Geostatistics for engineers and environmental scientists. Kluwer Academic Publishers, New York, 324 pp

Pannatier Y (1996) VARIOWIN: Software for spatial data analysis in 2D. Springer, New York, 91 pp

SAS (1989) SAS/STAT User's guide, vol 1 (version 6), 4th edn. SAS Institute Inc., Cary, North Carolina

SAS (2002) SAS/STAT User's guide, version 9. SAS Institute Inc., Cary, North Carolina

Sneath PHA, Sokal RR (1973) Numerical taxonomy. W.H. Freeman & Co., San Francisco, 573 pp

StatSoft (1997) STATISTICA manual, 97 edition