SANDIA REPORT
SAND2004-4667
Unlimited Release
Printed September 2005

Hierarchical Probabilistic Regionalization of Volcanism for Sengan Region in Japan Using Multivariate Statistical Techniques and Geostatistical Interpolation Techniques

Sean A. McKenna, Pinnaduwa H.S.W. Kulatilake, Jinyong Park, and P. Balasingam

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2004-4667
Unlimited Release
Printed September 2005

Hierarchical Probabilistic Regionalization of Volcanism for Sengan Region in Japan Using Multivariate Statistical Techniques and Geostatistical Interpolation Techniques

Pinnaduwa H.S.W. Kulatilake, Jinyong Park, P. Balasingam,
Computational Rock Mechanics & Rock Mass Modeling Laboratory
Geological Engineering Program
Department of Materials Science & Engineering
The University of Arizona,
Tucson, Arizona 85721

Sean A. McKenna
Geohydrology Department
Sandia National Laboratories
PO Box 5800 MS 0735
Albuquerque, New Mexico 87185

Abstract

Sandia National Laboratories, under contract to Nuclear Waste Management Organization of Japan (NUMO), is performing research on regional classification of given sites in Japan with respect to potential volcanic disruption using multivariate statistics and geo-statistical interpolation techniques. This report provides results obtained for hierarchical probabilistic regionalization of volcanism for the Sengan region in Japan by applying multivariate statistical techniques and geostatistical interpolation techniques on the geologic data provided by NUMO.

A workshop report produced in September 2003 by Sandia National Laboratories (Arnold et al., 2003) on volcanism lists a set of most important geologic variables as well as some secondary information related to volcanism. Geologic data extracted for the Sengan region in Japan from the data provided by NUMO revealed that data are not available at the same locations for all the important geologic variables. In other words, the geologic variable vectors were found to be incomplete spatially. However, it is necessary to have complete geologic variable vectors to perform multivariate statistical analyses. As a first step towards constructing complete geologic variable vectors, the Universal Transverse Mercator (UTM) zone 54 projected coordinate system and a 1 km square regular grid system were selected. The data available for each geologic variable on a geographic coordinate system were transferred to the aforementioned grid system. Also the recorded data on volcanic activity for Sengan region were produced on the same grid system. Each geologic variable map was compared with the recorded

volcanic activity map to determine the geologic variables that are most important for volcanism. In the regionalized classification procedure, this step is known as the **variable selection step.** The following variables were determined as most important for 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 values for each variable at 23949 centers of the chosen 1 km cell grid system that represents the Sengan region. These values formed complete geologic variable vectors at each of the 23949 one km cell centers.

The cluster analysis was performed on the 23949 complete variable vectors to classify each 1 km cell center location into (a) 5 different and (b) 3 different statistically homogeneous groups with respect to volcanism spanning from lowest possible volcanism to highest possible volcanism with increasing group number. In the regionalized classification procedure, this step is known as the **typification step.** Cluster analysis results obtained on groupings were verified by performing discriminant function analysis. A discriminant analysis along with the Bayes' theorem were used to calculate the probability of each 1 km cell center belonging to each of the groups defined in the typification step. The 1 km cell center was assigned to the group that produced the highest posterior probability. These probabilities were then used to construct a probability of membership map for each selected group. These maps were then used to delineate boundaries between different groups in the physical space. The latter stated three calculations belong to the **regionalization step** of the regionalized classification procedure. The said maps showed good comparisons with the recorded volcanism data of the Sengan region. No volcanic data were found to exist in the group 1 region of the 5-group situation. The highest probability area (i.e the lowest uncertainty) within group 1 has the chance of being the no volcanism region. Two measures are suggested to assess the uncertainty of the allocation process of the 1 km cell center based on the calculated posterior probabilities: the complement of the largest probability of group membership at each location and the entropy of classification at each location. These two measures were calculated and their spatial distributions in Sengan region are shown through maps. The recorded volcanism data are also plotted on the same maps to see the uncertainty level of the estimations at the locations where volcanism exists. The two uncertainty measures lead to the same conclusions. The volcanic data cell locations that are in the high volcanism regions (group 4 and 5 of 5-group situation and group 3 of 3-group situation) showed on the average relatively low uncertainty mapping estimations. On the other hand, the volcanic data cell locations that are in the low volcanism regions (group 2 of 5-group situation and group 1 of 3-group situation) showed on the average relatively high uncertainty mapping estimations. The volcanic data cell locations that are in the medium volcanism regions (group 3 of 5-group situation and group 2 of 3-group situation) showed on the average relatively moderate uncertainty mapping estimations. 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 future regionalization mapping and reduce the uncertainty level of the existing estimations.

Table of Contents

Abstract 3
Table of Contents 5
List of Figures 6
List of Tables 9
1. INTRODUCTION 11
2. SELECTION OF A COORDINATE SYSTEM AND A GRID SYSTEM TO SHOW GEOLOGIC VARIABLE DATA FOR SENGAN REGION 15
3. AVAILABLE DATA FOR VOLCANISM AND GEOLOGIC VARIABLES IN SENGAN REGION 17
3.1 Recorded Volcanism in Sengan Region 17
3.2 Groundwater Temperature 18
3.3 Groundwater pH Value 19
3.4 Geothermal Gradient 20
3.5 Heat Discharge 21
3.6 Presence of Quaternary Volcanic Rocks 23
3.7 Presence of Hydrothermal Alteration 25
3.8 Elevation 27
3.9 Slope Magnitude 32
3.10 Seismicity 35
3.11 Gravity Anomaly 37
3.12 Magnetic Anomaly 38
3.13 Maximum Horizontal Strain 39
3.14 Geologic Variables Correlated (Important) to Volcanism 40
4. GEOSTATISTICAL ANALYSIS 42
4.1 Groundwater Temperature 43
4.2 Groundwater pH 49
4.3 Geothermal Gradient 55
4.4 Heat Discharge 62
4.5 Presence of Volcanic Rocks 68
4.6 Presence of Hydrothermal Alteration 69
5. MULTIVARIATE CLASSIFICATION (TYPIFICATION) OF VOLCANISM FOR SENGAN REGION 70
5.1 General Features of Cluster Analysis (CA) 70
5.2 Multivariate Classification of Volcanism for the Sengan Region through CA 71
5.3 Fundamentals of Discriminant Function Analysis (DFA) 81
5.3.1. An example to illustrate the fundamentals of DFA 82
5.4 Verification of the Multivariate Classification of Volcanism for Sengan Region through DFA 83
6. REGIONALIZED MAPPING OF VOLCANISM FOR SENGAN REGION 87
6.1 Concepts of Regionalized Mapping 87
6.2 Results of Regionalized Mapping for the Sengan Region 89
6.3 Uncertainty Evaluations of Regionalized Mapping 96
7. SUMMARY AND CONCLUSIONS 101
8. ACKNOWLEDGEMENTS 103
9. REFERENCES 103
Appendix A 106

List of Figures

Fig. 1.1 Schematic diagram of regionalized classification (from Bohling et al. 1990). 13
Fig. 2.1 One km square grid system on UTM zone 54 PCS for the Sengan region. 16
Fig. 3.1 Available volcanic activity data for Sengan region. 17
Fig. 3.2 Available groundwater temperature data for Sengan region 19
Fig. 3.3 Available groundwater pH data for Sengan region 20
Fig. 3.4 Available geothermal gradient data for Sengan region. 21
Fig. 3.5 Available heat discharge data for Sengan region on a raster image 22
Fig. 3.6 Areas of available heat discharge data for Sengan region on UTM zone 54 22
Fig. 3.7 Available heat discharge data for Sengan region on a 1 km square grid using UTM zone 54 PCS 23
Fig. 3.8 Digitized image of volcanic rock areas on UTM zone 54 PCS from an available map of hydrothermal alteration 24
Fig. 3.9 Polygonal coverage of available volcanic rock areas in Sengan region from two different sources 25
Fig. 3.10 Data on hydrothermal alteration zones on a raster image available from a map of hydrothermal alteration. 26
Fig. 3.11 Polygonal coverage of available hydrothermal alteration areas in Sengan region from two different sources 27
Fig. 3.12 Areas where elevation data are available on a 10 m grid system 28
Fig. 3.13 Areas where elevation data are available on 50 m and 250 m grid systems 29
Fig. 3.14 Elevation model obtained for 1 km grid cells from 50 m grid mesh. Note it does not cover the whole study area 30
Fig. 3.15 Elevation model obtained for 1 km grid cells from 250 m grid mesh. 31
Fig. 3.16 Digital Elevation Model (DEM) obtained for Sengan region 32
Fig. 3.17 Slope magnitude model obtained for 1 km grid cells from 50 m grid mesh. Note it does not cover the whole study area 33
Fig. 3.18 Slope magnitude model obtained for 1 km grid cells from 250 m grid mesh 34
Fig. 3.19 Slope magnitude model obtained for Sengan region 35
Fig. 3.20 Available data on seismic activity frequency for Sengan region on UTM zone 54 PCS 36
Fig. 3.21 Available data on seismic activity magnitude for Sengan region on UTM zone 54 PCS 37
Fig. 3.22 Available gravity anomaly data for Sengan region on UTM zone 54 PCS 38
Fig. 3.23 Available magnetic anomaly data for Sengan region on UTM zone 54 PCS 39
Fig. 3.24 Available maximum shear strain data for Sengan region on UTM zone 54 PCS. 40
Fig. 4.1 Region of the database used for calculation of directional variograms for groundwater temperature 43
Fig. 4.2 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. 44
Fig. 4.3 Variogram contour plots for groundwater temperature at different lag spacings and number of lags 44
Fig. 4.4 Omnidirectional variogram (a) and directional variograms (b) through (m) for groundwater temperature at 1 km lag spacing 46
Fig. 4.5 Variation of correlation distance with direction for groundwater temperature 47
Fig. 4.6 Estimated values of groundwater temperature from kriging for the study area. 48
Fig. 4.7 Predicted standard error values of average groundwater temperature from kriging for the study area. 49
Fig. 4.8 Region of the database used for calculation of directional variograms for groundwater pH 50

Fig. 4.9 Variogram (a) surface and (b) contour plots for groundwater pH at 1 km lag spacing and number of lags = 10. Note: In Fig. 4.9a, the black triangle shown on the color legend indicates the total variance value. 50
Fig. 4.10 Omnidirectional variogram (a) and directional variograms (b) through (m) for groundwater pH at 1 km lag spacing 52
Fig. 4.11 Variation of correlation distance with direction for groundwater pH 53
Fig. 4.12 Estimated values of groundwater pH from kriging for the study area. 54
Fig. 4.13 Predicted standard error values of average groundwater pH from kriging for the study area. 55
Fig. 4.14 Region of the database used for calculation of directional variograms for geothermal gradient. 56
Fig. 4.15 Variogram (a) surface and (b) contour plots for geothermal gradient at 3 km lag spacing and number of lags = 20. Note: In Fig. 4.15a the black triangle shown on the color legend indicates the total variance value. 56
Fig. 4.16 Omnidirectional variogram (a) and directional variograms (b) through (m) for geothermal gradient at 3 km lag spacing. 59
Fig. 4.17 Variation of correlation distance with direction for geothermal gradient 60
Fig. 4.18 Estimated values of geothermal gradient from kriging for the study area. 61
Fig. 4.19 Predicted standard error values of average geothermal gradient from kriging for the study area. 61
Fig. 4.20 Region of the database used for calculation of directional variograms for heat discharge 62
Fig. 4.21 Variogram surface plots of heat discharge at different lag spacings and number of lags. Note: the black triangle shown on the color legend indicates the total variance value. 63
Fig. 4.22 (a) Omnidirectional variogram and (b) through (m) directional variograms for heat discharge at 5 km lag spacing 65
Fig. 4.23 Variation of correlation distance with direction for heat discharge 66
Fig. 4.24 Estimated values of heat discharge from kriging for the study area. 67
Fig. 4.25 Predicted standard error values of average heat discharge from kriging for the study area. 67
Fig. 4.26 Available quaternary volcanic rock data for Sengan region on 1 km square grid using UTM zone 54 PCS 68
Fig. 4.27 Available hydrothermal alteration data for Sengan region on 1 km square grid using UTM zone 54 PCS 69
Fig. 5.1 Dendrogram obtained for the final 50 groups 72
Fig. 5.2 Mean variable vector for each cluster for the 5-group clustering 74
Fig. 5.3 Mean variable vector for each cluster for the 3-group clustering 75
Fig. 5.4 Relative frequency versus groundwater temperature plots of different clusters for 5-group situation 75
Fig. 5.5 Relative frequency versus groundwater pH plots of different clusters for 5-group situation 76
Fig. 5.6 Relative frequency versus geothermal gradient plots of different clusters for 5-group situation 77
Fig. 5.7 Relative frequency versus heat discharge plots of different clusters for 5-group situation 78
Fig. 5.8 Relative frequency versus groundwater temperature plots of different clusters for 3-group situation 79
Fig. 5.9 Relative frequency versus groundwater pH plots of different clusters for 3-group situation 79
Fig. 5.10 Relative frequency versus geothermal gradient plots of different clusters for 3-group situation 80

Fig. 5.11 Relative frequency versus heat discharge plots of different clusters for 3-group situation 80
Fig. 5.12 Scatter plot between sorting coefficient and median grain size for the samples coming from the two groups of sands 83
Fig. 5.13 Raw discriminate scores (plotted along the discriminant line) of sample belonging to the two groups of sands 83
Fig. 5.14 (a) Cluster 1 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis 85
Fig. 5.14 (b) Clusters 1 and 2 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis 86
Fig. 5.14 (c) Clusters 1, 2 and 3 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis 86
Fig. 5.15 Clusters 1 through 5 on the plot between Root 2 and Root 1 from the 5-cluster canonical analysis 87
Fig. 6.1 (a) Regionalized probability distribution for group 1 for 5-cluster analysis 92
Fig. 6.1 (b) Regionalized probability distribution for group 2 for 5-cluster analysis 92
Fig. 6.1 (c) Regionalized probability distribution for group 3 for 5-cluster analysis 93
Fig. 6.1 (d) Regionalized probability distribution for group 4 for 5-cluster analysis 93
Fig. 6.1 (e) Regionalized probability distribution for group 5 for 5-cluster analysis 94
Fig. 6.2 (a) Regionalized probability distribution for group 1 for 3-cluster analysis 94
Fig. 6.2 (b) Regionalized probability distribution for group 2 for 3-cluster analysis 95
Fig. 6.2 (c) Regionalized probability distribution for group 3 for 3-cluster analysis 95
Fig. 6.3 Regionalized distribution of the first measure of uncertainty for 3-group situation along with the available volcanic activity data located in different clustering groups 97
Fig. 6.4 Regionalized distribution of entropy (the second measure of uncertainty) for 3-group situation along with the available volcanic activity data located in different clustering groups 98
Fig. 6.5 Histograms of the frequency of volcanic data located in different regionalized groups and the associated uncertainty level (a) the first measure of uncertainty, (b) entropy for each group for 3-group situation 98
Fig. 6.6 Regionalized distribution of the first measure of uncertainty for 5-group situation along with the available volcanic activity data located in different clustering groups 99
Fig. 6.7 Regionalized distribution of entropy (the second measure of uncertainty) for 5-group situation along with the available volcanic activity data located in different clustering groups 99
Fig. 6.8 Histograms of volcanic data located in different regionalized groups and the associated uncertainty level (a) the first measure of uncertainty, (b) entropy for each group for 5-group situation 100

List of Tables

Table 2.1 Coordinates of the study area according to the UTM zone 54 PCS 16
Table 3.1 Values of basic statistics for the geologic variable data available for Sengan region. 41
Table 4.1 Values of basic statistics for the geologic variable data used to perform geostatistical analysis 42
Table 5.1 Cluster analysis results for 5-group clustering 73
Table 5.2 Cluster analysis results for 3-group clustering 74
Table 5.3 Raw coefficients for variables from canonical analysis to form Roots for 3-group clustering 84
Table 5.4 Raw coefficients for variables from canonical analysis to form Roots for 5-group clustering 85
Table 6.1a Values of q_i used for quadratic discriminant analysis for 3-group situation 91
Table 6.1b Values of q_i used for linear discriminant analysis for 3-group situation 91
Table 6.2a Values of q_i used for quadratic discriminant analysis for 5-group situation 91
Table 6.2b Values of q_i used for linear discriminant analysis for 5-group situation 91
Table 6.3 Distribution of volcanic data locations among the five different groups obtained under 5-cluster analysis 96
Table 6.4 Distribution of volcanic data locations among the three different groups obtained under 3-cluster analysis 96
Table A.1 Input data used for discriminant analysis (file SANDS.TXT from the URL www.wiley.com/college/davis) 106

1. 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. Sandia National Laboratories (SNL) has a contract with NUMO to participate in the development of spatial data analysis techniques for extrapolation of uncertain parameters with regard to the volcanic and active fault factors. The Spatial Data Analysis Techniques for Extrapolation of Uncertain Parameters research activity consists of two tasks: 1) the Discrete Feature Mapping Methodology task and 2) the Disruptive Influence Probability Mapping task. The Discrete Feature Mapping Methodology task is related to the screening of sites with regard to active faults. The Disruptive Influence Probability Mapping task is related to screening sites with regard to volcanism, based on multivariate data analysis. This second mapping task is the focus of this report. The University of Arizona (U of A) has been contracted by SNL to work on regional classification of given sites in Japan with respect to volcanism using multivariate statistical techniques and geo-statistical interpolation techniques. This report provides the procedures used and results obtained for the U of A study on hierarchical regionalization of volcanism for Sengan region in Japan.

SNL organized a workshop at Los Alamos National Laboratory (LANL) on September 17-18, 2003 to discuss models of volcanism and the geologic variables that are linked to volcanism. Separation between high volcanic activity areas and no volcanic activity areas are based on the history of volcanic activity in Japan and geologic theory and are presented in the workshop report (Arnold et al., 2003). Regionalized classification of volcanic risk in Japan resulting from application of computational procedures to available data should be consistent with the conceptual models of volcanic activity developed at the volcanism workshop. The workshop report identified the following variables as the most important variables related to volcanic activity: geothermal gradient, geothermal heat flow, groundwater temperature, presence of quaternary volcanic rocks, presence of hydrothermal alteration and groundwater pH value. In addition to these variables, there is also a 15km respect distance away from any active volcanic vent that must be observed. 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.

NUMO provided data on the aforementioned variables on several CDs to extract the relevant data for Sengan region in Japan. All of the geologic data included in CDs use a Geographic Coordinate System (GCS) to specify the locations. It seems that NUMO has used the ArcView 3.x software package to show data belonging to geologic variables on maps. 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, 1992) 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 huge differences of the available numbers of the data for different geologic variables, the third option was used to construct complete variable vectors.

Chapter 2 of this report discusses how a coordinate system and a regular grid system were selected to initiate construction of complete geologic variable vectors for Sengan region using the ArcGIS 8.x software package. Chapter 3 shows the available data on volcanic activity and the geologic variables for Sengan region using the selected coordinate system and the grid system. At the end of the same chapter a comparison is made between the map of volcanism data and the data posting of each geologic variable to select the geologic variables that are important for volcanic activity. Chapter 4 deals with the variogram modeling and kriging performed for the geologic variables to construct the complete variable vectors for Sengan region.

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 1.1 illustrates the basic idea of regionalized classification. Two variables, A and B, are measured at a number of stations distributed throughout a study area (Figure 1.1a). 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) or neural network methodology (Bishop, 1995; Schalkoff, 1997; Zaknich, 2003). This process, referred to as **typification**, might be used to identify, say, three classes of interest (Figure 1.1b). The classes identified contain observations that are simultaneously as similar as possible to other observations in the same class and as distinct as possible from observations in other classes.

Each observation can then be assigned a probability of membership in each class, 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. Because the probabilities are smooth functions of spatially continuous variables, they can be interpolated to a regular grid in physical space and contoured. Comparison of the probability maps for different classes provides estimates of the locations of the boundaries between classes. These spatial mapping steps are the essential features of regionalized classification (Figure 1.1c). This technique has 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 and 1993), gas (Harff et al. 1990 and 1993), mineral resources (Harff et al., 1991), electrofacies properties (Moline and Bahr, 1995) and grain size properties (Fernandez et al. 1997).

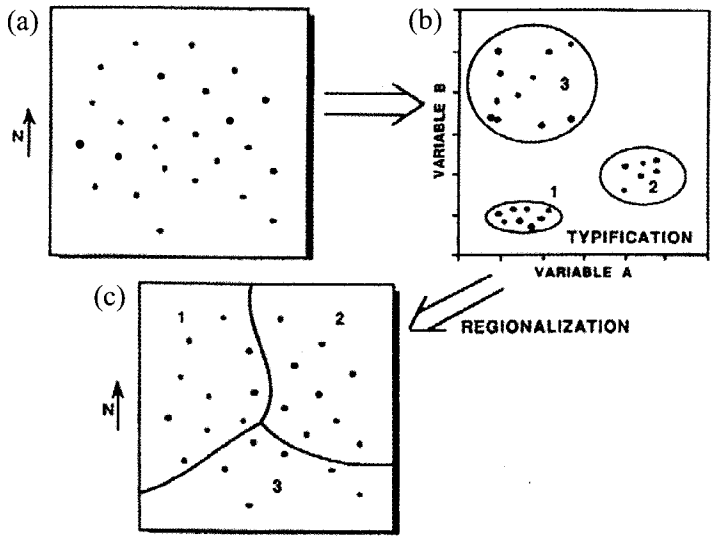


Fig. 1.1 Schematic diagram of regionalized classification (from Bohling et al. 1990).

The regionalized classification procedure used in this study 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. As mentioned before, this step is covered in Chapter 3 of this report. The constructed complete geologic variable vector data (obtained in Chapter 4) coming from selected grid point locations are classified into different statistically homogeneous groups with respect to volcanism in the **typification step**. Cluster analysis is used to perform the typification task and it is covered in Chapter 5. In the same Chapter, discriminant function analysis is used to verify the results obtained from the cluster analysis. A discriminant analysis along with the Bayes' theorem are used to calculate the probability of each 1 km cell center location belonging to each of the groups defined in the typification step. The 1 km cell center is assigned to the group that produced the highest probability. These probabilities are then

used to construct a probability of membership map with respect to volcanism for each selected group. These maps are then used to delineate boundaries between different groups in the physical space. The latter three calculations belong to the **regionalization step** and they are covered for the Sengan region in Chapter 6. Two measures are suggested to assess the uncertainty of the allocation process of the grid point locations based on the calculated posterior probabilities. These are the complement of the maximum posterior probability of class membership and the classification entropy. These measures are calculated and their spatial distribution in the Sengan region is shown in Chapter 6. Maps obtained in Chapter 6 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. A brief summary and conclusions are given in Chapter 7.

2. SELECTION OF A COORDINATE SYSTEM AND A GRID SYSTEM TO SHOW GEOLOGIC VARIABLE DATA FOR SENGAN REGION

Two main types of coordinate systems are available to show the global locations of geologic data. The first type available is a geographic coordinate system (GCS). In a GCS, the distances are not the same for the same degree increment of latitude and longitude. In addition, distances change with latitude for the same increment of latitude as the position moves along a longitude. In a GCS, grid spacing will be given in terms of degrees, minutes and seconds. The GCS changes with the chosen datum. GCS-Tokyo is a GCS available for Japan.

The second type available is a projected coordinate system (PCS). In a PCS, grid spacing can be given in terms of km. Many projection methods are available in ArcGIS 8.x to convert from a GCS to a PCS. In ArcGIS 8.x, Transverse Mercator (TM) projection method is used to convert GCS-Tokyo to 19 different projected coordinate grid systems (or zones) for Japan. Each one differs by the chosen datum. The mid point of the Sengan region has a longitude close to 141 degrees and a latitude close to 40 degrees. The Datum for Zone 10 has a central longitude of 140.833 degrees and an origin of latitude of 40 degrees. Out of all the 19 zones, zone 10 seems to be the most appropriate one to use for the Sengan region. However, if the goal is to cover the northern area of Japan including the Sengan region, it would be more appropriate to use the UTM (Universal Transverse Mercator) projection method to convert from a GCS to a PCS system. The UTM method is more common in Japan relative to the TM method. 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

All of the geologic data included in the CDs, provided by NUMO, use a GCS to specify the locations. It seems that NUMO has used the ArcView 3.x software package to show data belonging to geologic variables on maps. **Data of all the geologic variables considered in this research have been converted from a GCS to a PCS using UTM zone 54 with the datum GRD 80.**

The Sengan geothermal region is defined from longitude 140.50 to 141.25 and from latitude 40.1667 to 39.667 in the directory entitled "Geological map of Sengan" available in the Sengan No. 1 CD. This is roughly a 75×75 km area. On the other hand, according to the directory entitled "Elevation 250m mesh for Sengan area 0105" given in the Sengan No. 2 CD, the Sengan region lies between longitudes 139.667-142.0844 degrees and latitudes 39.333-40.667 degrees using Greenwich as the prime longitude. The latter stated Sengan region covers 209 km in the E-W direction and 149km in N-S direction. This study uses the latter mentioned extended area to represent Sengan region. Table 2.1 shows the starting and ending 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. 2.1. 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.

Table 2.1 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

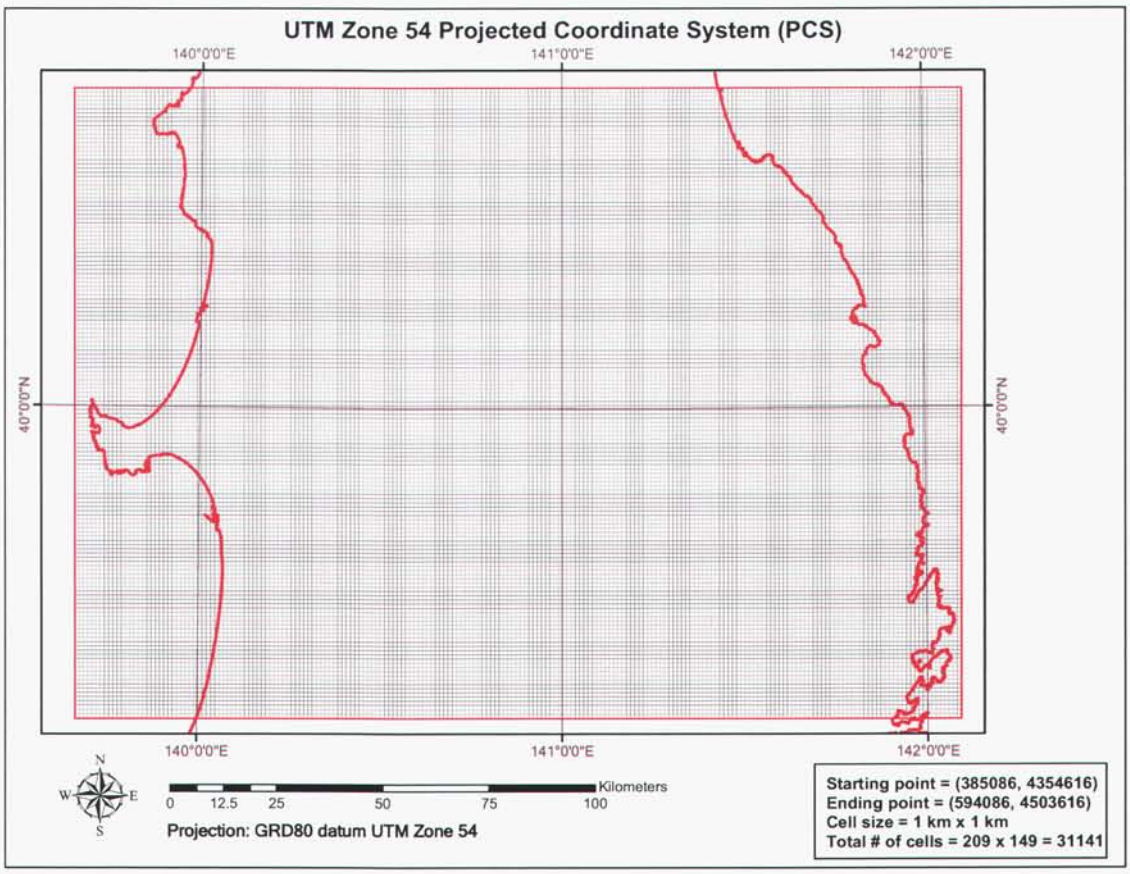


Fig. 2.1 One km square grid system on UTM zone 54 PCS for the Sengan region.

3. AVAILABLE DATA FOR VOLCANISM AND GEOLOGIC VARIABLES IN SENGAN REGION

3.1 Recorded Volcanism in Sengan Region

The CD entitled '**ITM3-1**'\0401Quaternary volcanoes\Edifice by topography.shp, Edifice by vent.shp & Volcano center by topography.shp provides information available on volcanism for the study area. The available data can be separated into three groups as follows: points classified as 'edifice by vent' = 80 values; points classified as 'edifice by topography' = 5 values; points classified as 'center by topography' = 30 values. Fig. 3.1 shows the locations for the aforesaid volcanism in the study area.

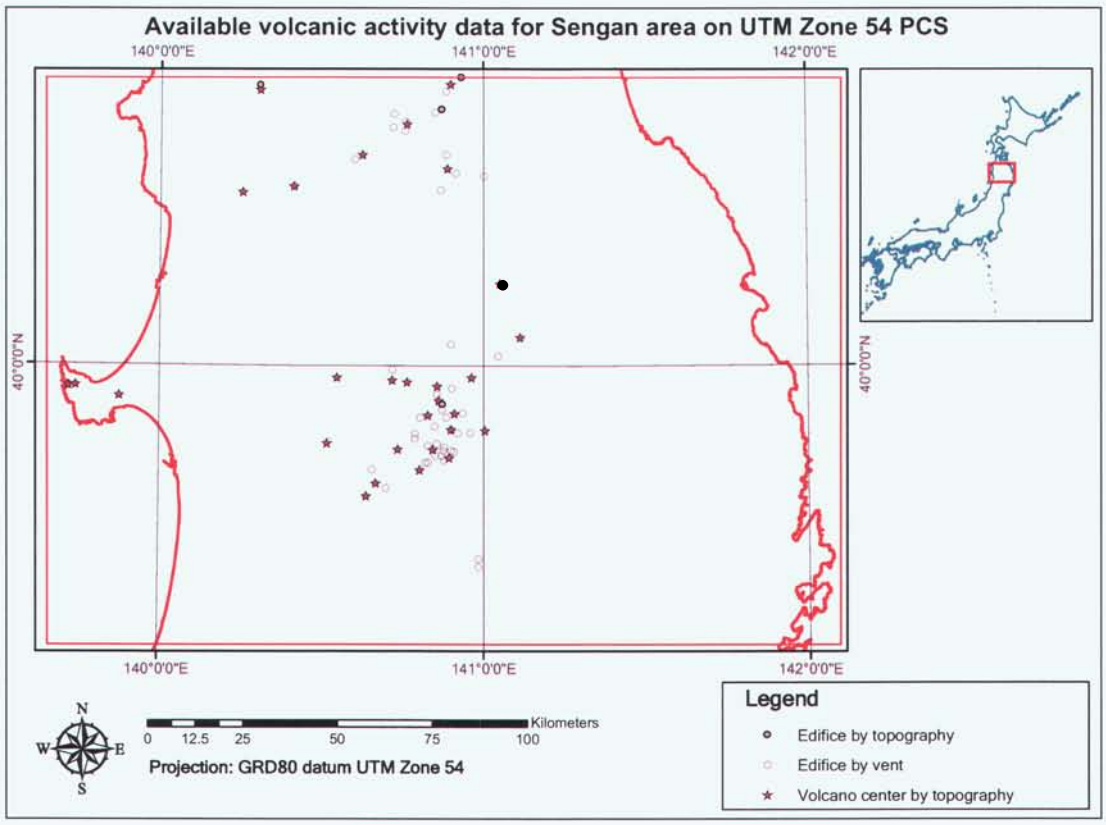


Fig. 3.1 Available volcanic activity data for Sengan region.

3.2 Groundwater Temperature

Groundwater temperature data were obtained through the following sources: groundwater databases; hotspring databases; fumaroles and geothermal wells. The following groundwater databases were used to obtain a portion of the groundwater temperature data:

- CD title '**ITM3-3**'\0840GWDB for Aomori, Akita and Yamagata Prefecture/

  Aomori_Pref.shp, Akita_Prep.shp and Yamagata_Pref.shp

- CD title '**Sengan No2**'\JNC GWDB for Akita and Iwate Pref\Iwate_Pref.shp

In the above 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. The following hotspring database was used to obtain a part of the groundwater temperature data:

- CD title '**ITM3-2**'\0413Distribution map and catalogue of hot

  spring\HotspringDB.shp

The following two tables were used to extract groundwater temperature data from the 'fumaroles' in Akita area:

- CD title '**Geothermal resource map of Tohoku and**

  Kyushu'\TABLE\FMD1AKI.txt and FMD2AKI.txt

Many repeated data points were found in these data sets. After careful comparison between data sets, averages were calculated to represent the repeated data points. The following two tables were used to extract groundwater temperature (vapor temperature) data from the 'geothermal wells' in Akita area:

- CD title '**Geothermal resource map of Tohoku and**

  Kyushu'\TABLE\WLD1AKI.txt and WLD2AKI.txt

All the aforementioned groundwater temperature data were combined to get a final database for the groundwater temperature variable. Fig. 3.2 shows the constructed data base available for groundwater temperature on UTM zone 54 PCS. A few 1 km square cells had more than one data point and many cells did not have a single data point. In Fig. 3.2 a color scheme is used to separate the groundwater temperature data into five arbitrary groups. 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 no data.

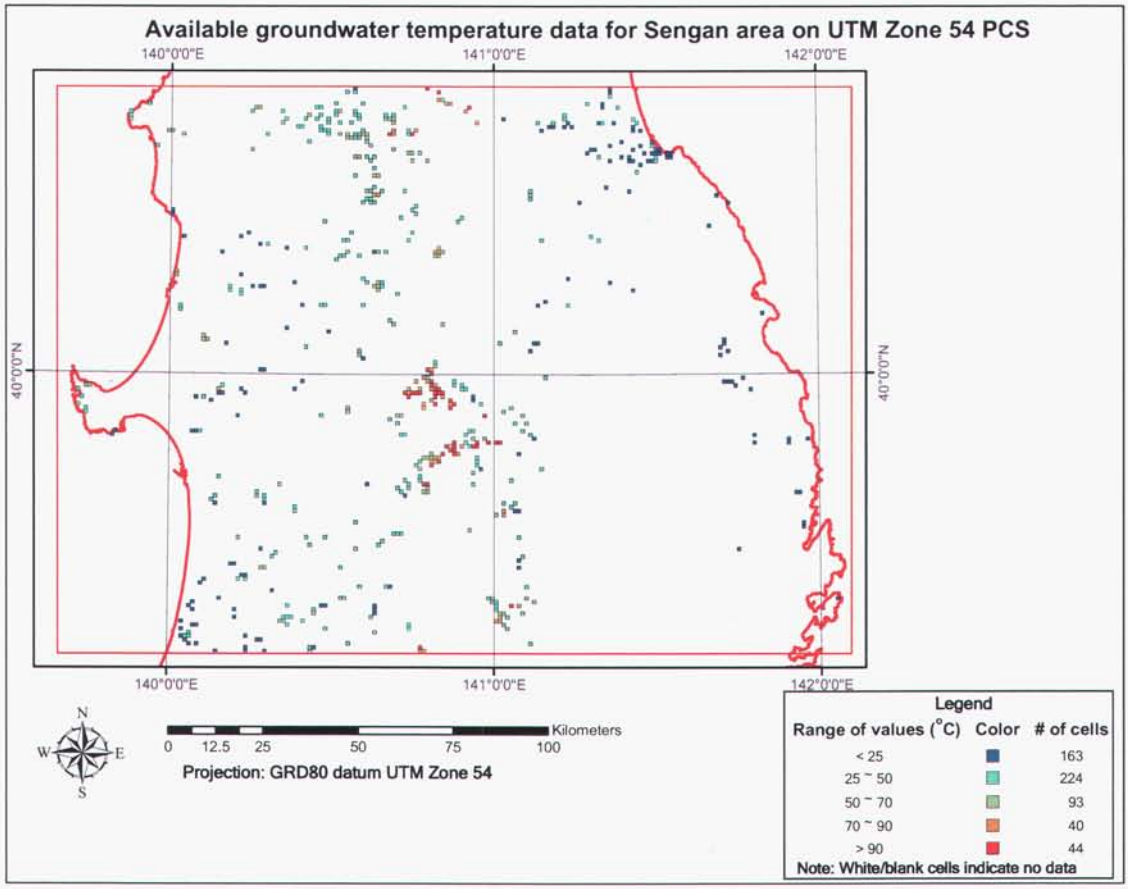


Fig. 3.2 Available groundwater temperature data for Sengan region

3.3 Groundwater pH Value

The following groundwater databases provide the point location pH values for the study area:

- CD title '**ITM3-3**'\0840GWDB for Aomori, Akita and Yamagata Prefecture/ Aomori_Pref.shp, Akita_Prep.shp and Yamagata_Pref.shp
- CD title '**Sengan No2**'\JNC GWDB for Akita and Iwate Pref\Iwate_Pref.shp

Note that even though temperature values are available from the groundwater databases, hotspring databases, fumaroles and geothermal wells, pH values are only available from the groundwater databases. Fig. 3.3 shows the available pH values for the study area on UTM zone 54 PCS. A few 1 km square cells had more than one data point and many cells did not have a single data point. In Fig. 3.3 a color scheme is used to separate the groundwater pH data into five arbitrary groups. 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 no data.

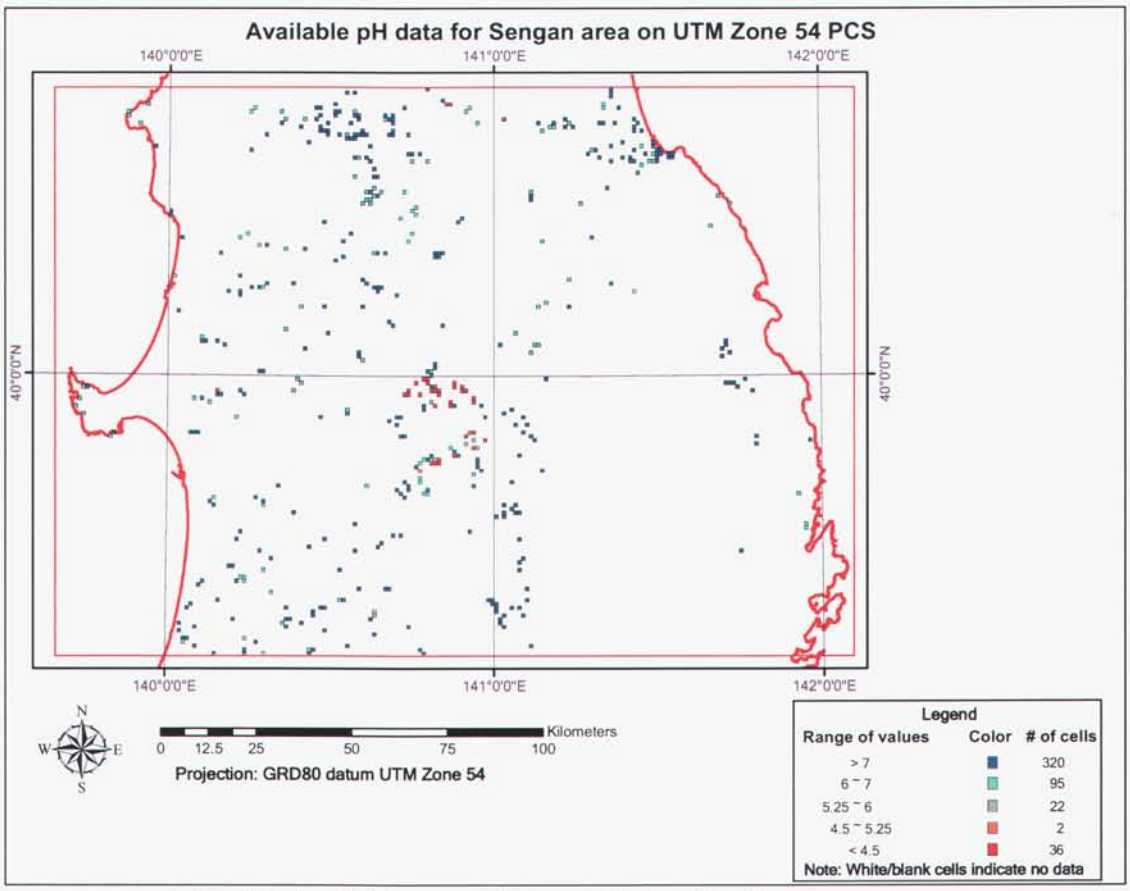


Fig. 3.3 Available groundwater pH data for Sengan region

3.4 Geothermal Gradient

The following database provides the point geothermal gradient values for the study area:

- CD title '**Sengan No2**'\geothermal gradient data0408\0408geoth_point.shp

Note that the unit of geothermal gradient is C°/km. Fig. 3.4 shows the available geothermal gradient values for the study area on UTM zone 54 PCS. A few 1 km square cells had more than one data point and many cells did not have a single data point. In Fig. 3.4 a color scheme is used to separate the geothermal gradient data into five arbitrary groups. Note 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.

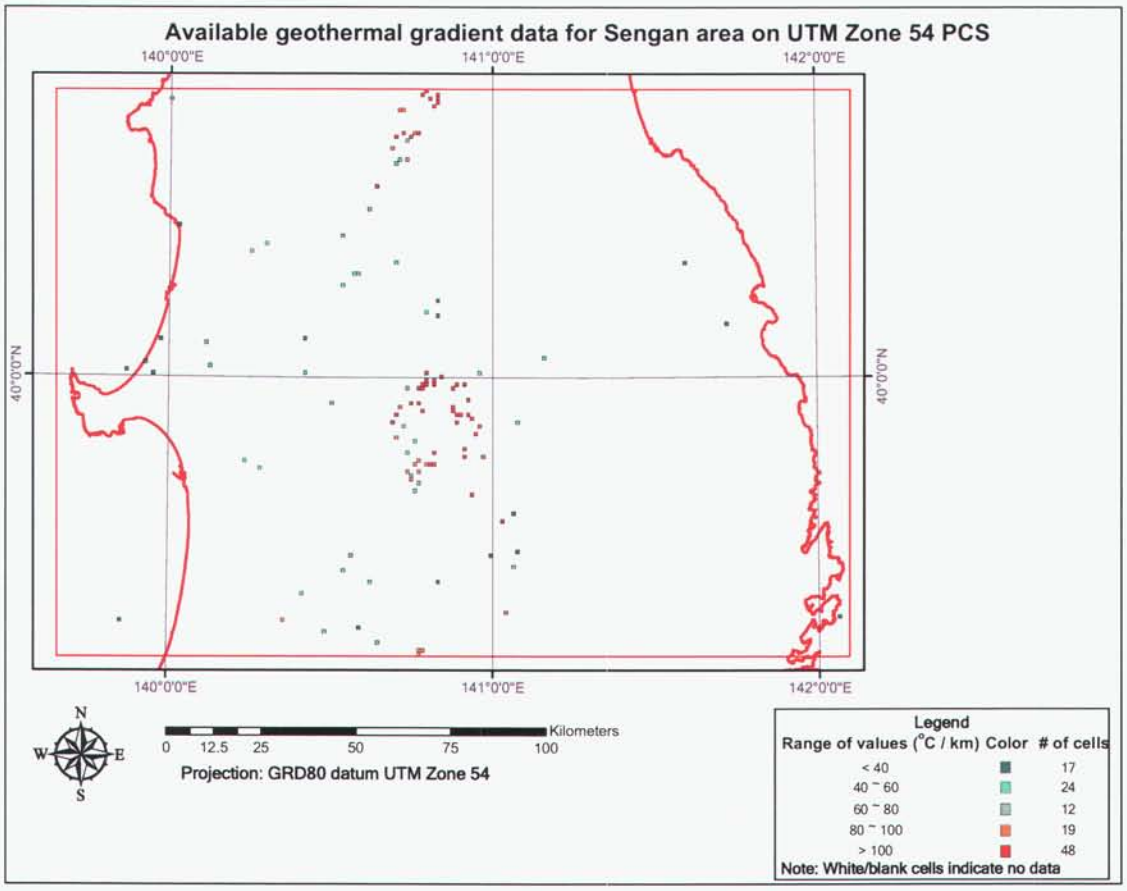


Fig. 3.4 Available geothermal gradient data for Sengan region.

3.5 Heat Discharge

Heat discharge data by hot springs are available for the study area through the following database:

- CD title '**Sengan No2**'\ Distribution map of heat discharge 0410tif \U0410G.tif

This database provides a raster image having an area mesh of heat discharge as shown in Fig. 3.5. The numbers given in the map are the logarithms of the heat discharge. The unit of heat discharge is $\log_{10}(\mu Wm^{-2})$.

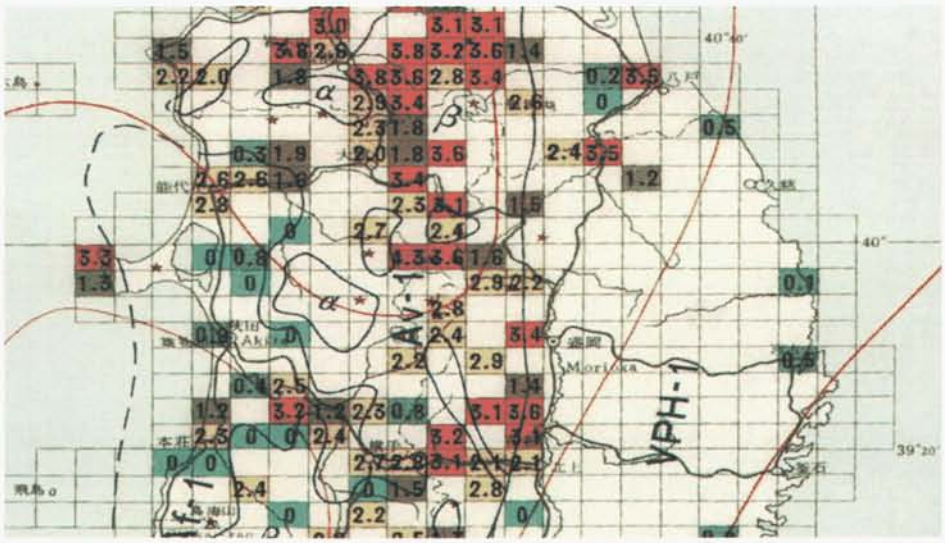


Fig. 3.5 Available heat discharge data for Sengan region on a raster image

Each rectangle of Fig. 3.5 was digitized, and then converted to UTM zone 54 PCS coordinates (Fig. 3.6). The 1 km square grid system was superimposed on the top of the digitized heat discharge map. Then heat discharge value for each 1 km square was specified by the value given to the rectangle that covered the majority area of the 1 km cell. The obtained values are shown in Fig. 3.7. In Fig. 3.7 a color scheme is used to separate the heat discharge data into five arbitrary groups. Note that the higher the heat discharge value, the higher the chance of volcanism. Also note that there is a large area in the southeastern region of the study area with no data.

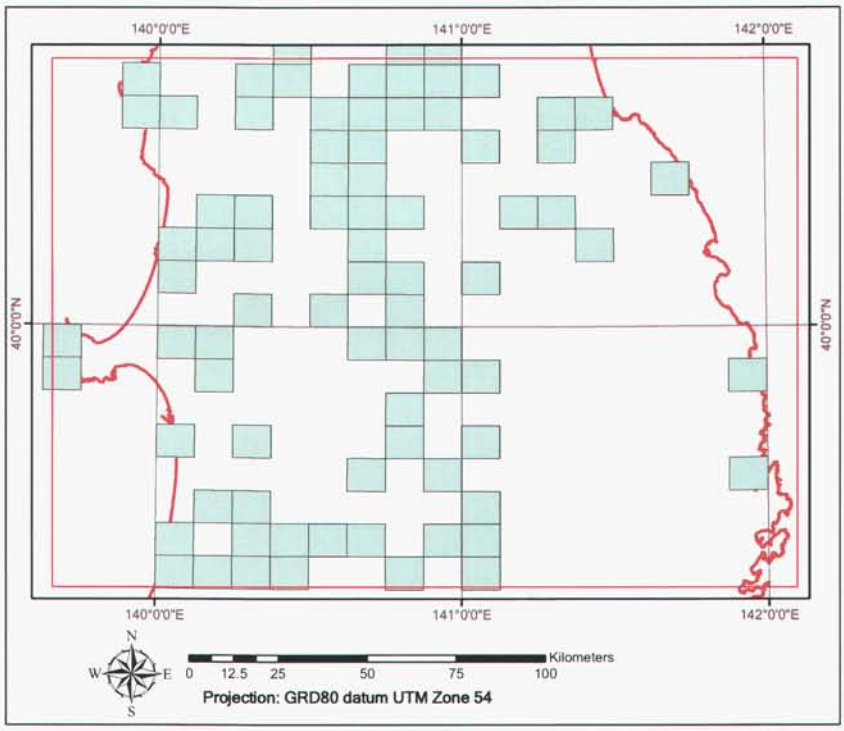


Fig. 3.6 Areas of available heat discharge data for Sengan region on UTM zone 54

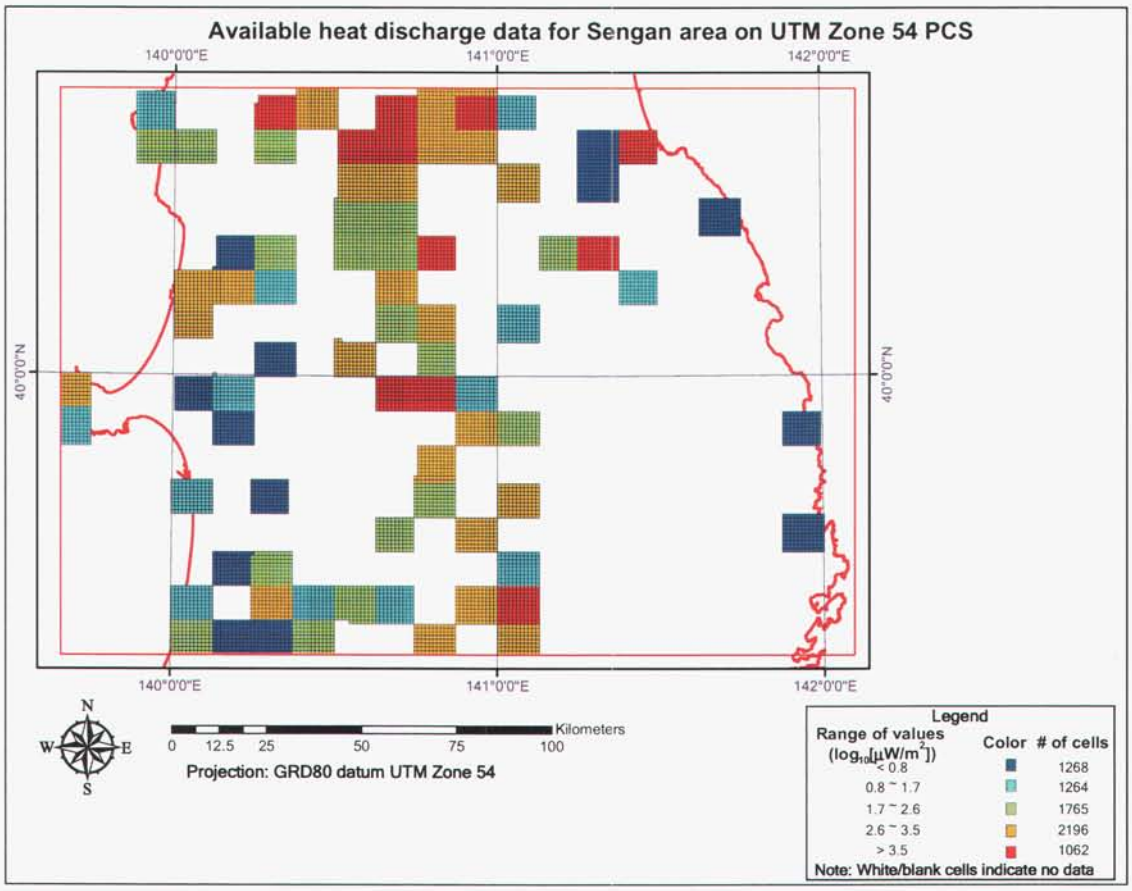


Fig. 3.7 Available heat discharge data for Sengan region on a 1 km square grid using UTM zone 54 PCS

3.6 Presence of Quaternary Volcanic Rocks

A raster image of the distribution of quaternary volcanic rocks is available from the following database:

- CD title '**ITM3-2**'\ 0603Distribution map of hydrothermal alteration\ U0603G.tif'

The volcanic rock areas from this raster image were digitized and then converted to a UTM zone 54 PCS to represent volcanic rocks in a map (Fig. 3.8).

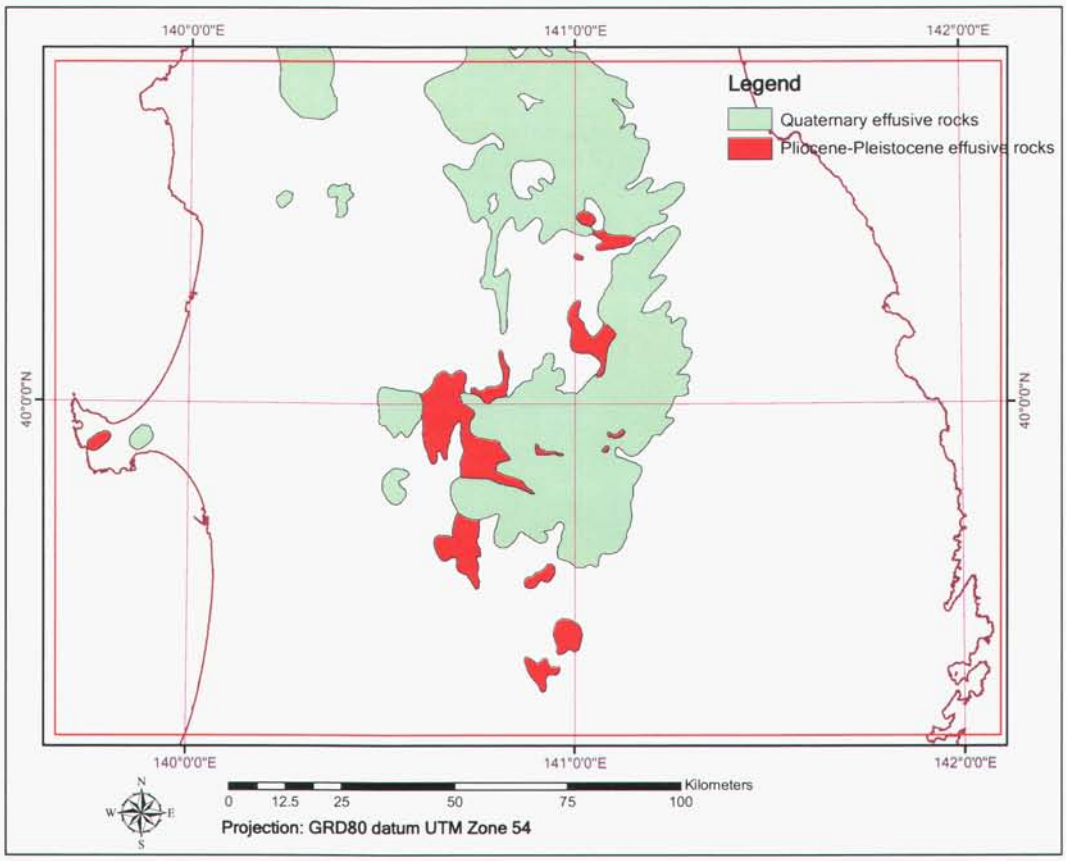


Fig. 3.8 Digitized image of volcanic rock areas on UTM zone 54 PCS from an available map of hydrothermal alteration

The following geologic database provides a polygonal coverage of presence of volcanic rocks for the study area:

- CD title '**Geothermal resource map of Tohoku and Kyushu**'\ ARC\ ARCDATA\ THK\ SHAPE\ GEOTHK.shp

Both maps were combined and then a polygonal coverage was created (Fig.3.9).

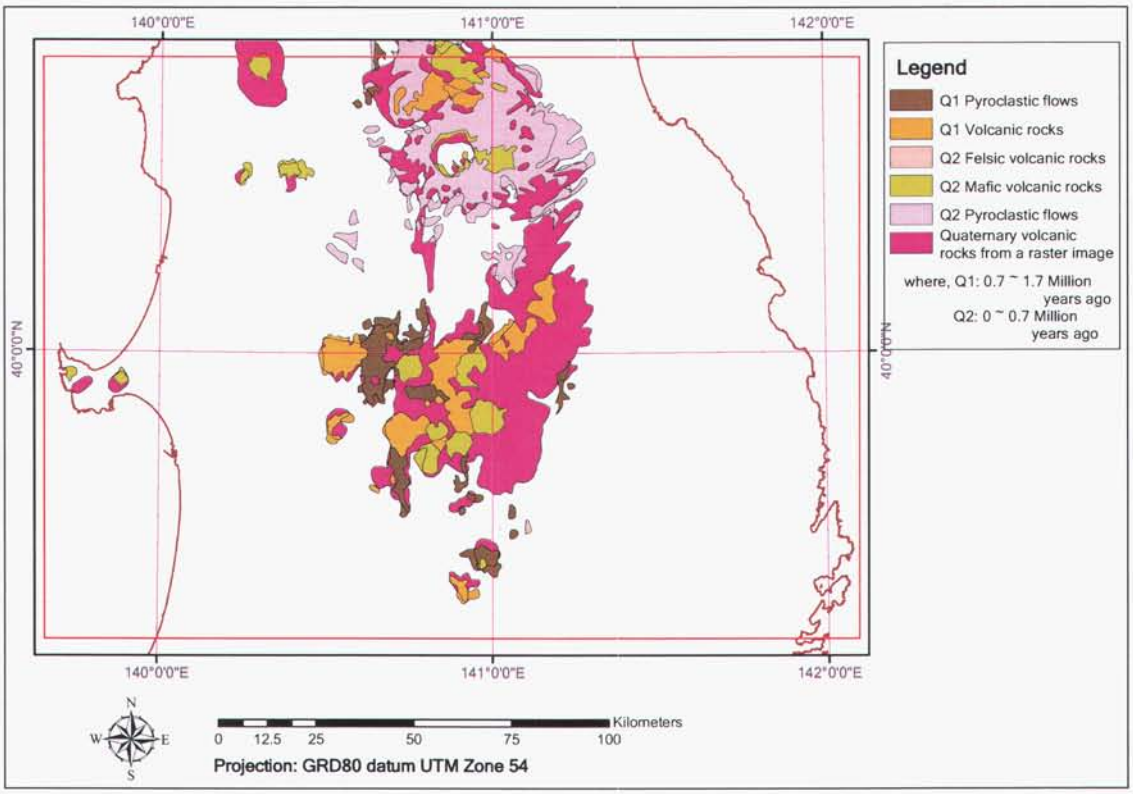


Fig. 3.9 Polygonal coverage of available volcanic rock areas in Sengan region from two different sources

3.7 Presence of Hydrothermal Alteration

A raster image of the distribution of hydrothermal alteration zones is available from the following database:

- CD title '**ITM3-2**'\ 0603Distribution map of hydrothermal alteration\ Distribution map of heat discharge 0410tif \

The hydrothermal alteration areas from this raster image were digitized and then converted to a UTM zone 54 PCS to represent hydrothermal alteration in a map (Fig. 3.10).

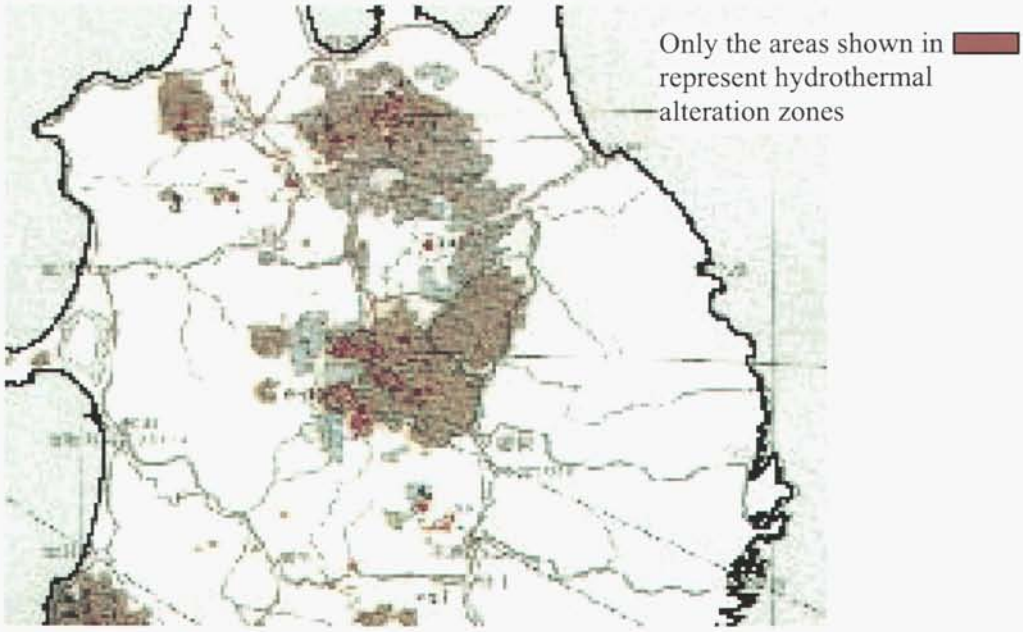


Fig. 3.10 Data on hydrothermal alteration zones on a raster image available from a map of hydrothermal alteration.

The following hydrothermal alteration database provides a polygonal coverage of presence of hydrothermal alteration for the study area:

- CD title '**Geothermal resource map of Tohoku and Kyushu**'\ ARCDATA\ THK\ SHAPE\ ALTTHK.shp

Both maps were combined and then a polygonal coverage was created (Fig. 3.11).

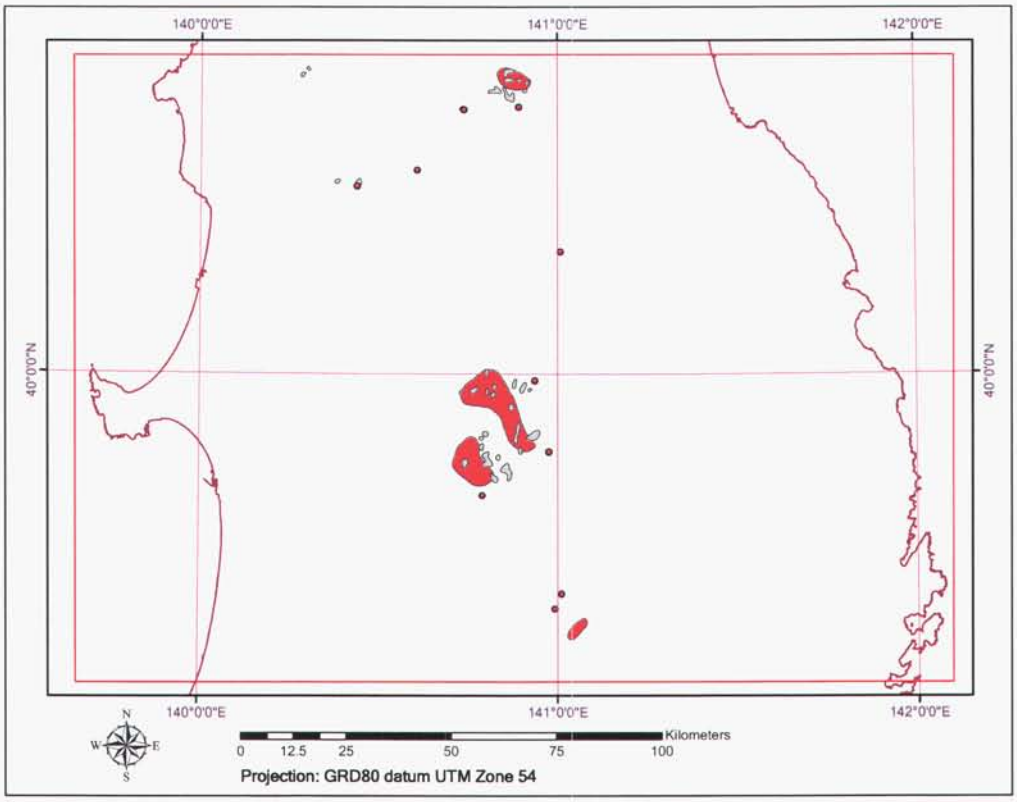


Fig. 3.11 Polygonal coverage of available hydrothermal alteration areas in Sengan region from two different sources

3.8 Elevation

Elevation data are available in three different grid sizes: 10 m, 50 m and 250 m. However, the 10 m grid mesh covers only Mt. Iwate and Mt. Akitakomagatake (Fig. 3.12). The 50 m grid mesh available with one of the CDs provided does not cover the whole study area even though it covers most of the study area (Fig. 3.13). The 250 m grid mesh covers the whole study area, but provides elevation values at a lower resolution compared to the 50 m grid mesh. Thus the 50 m grid system was used to estimate the elevation values at centers of 1 km square grid cells that fell within the study area covered by the 50 m grid mesh and the 250 m grid system was used for the remaining portion of the study area for which the 50 m grid values were not available. Note that the cell sizes of 250 m and 50 m grid meshes are respectively, 7.5" longitude × 11.25" latitude and 1.5" longitude × 2.25" latitude based on a geographic coordinate system. The following databases provide the elevation values for 50 m and 250 m grid meshes:

- 50 m grid mesh:
- CD title '**Sengan No3**'\ Elevation 50m mesh for Sengan area0106\ 5940, 5941, 6040 & 6041 (Note that the numbers indicate particular areas)
- 250 m grid mesh:
- CD title '**ITM3-1**'\ 0105Elevation 250m mesh\ Mesh Code5839, Mesh Code5840, Mesh Code5841, Mesh Code5939, Mesh Code5942, Mesh Code6039, Mesh Code6139, Mesh Code6140 & Mesh Code6141

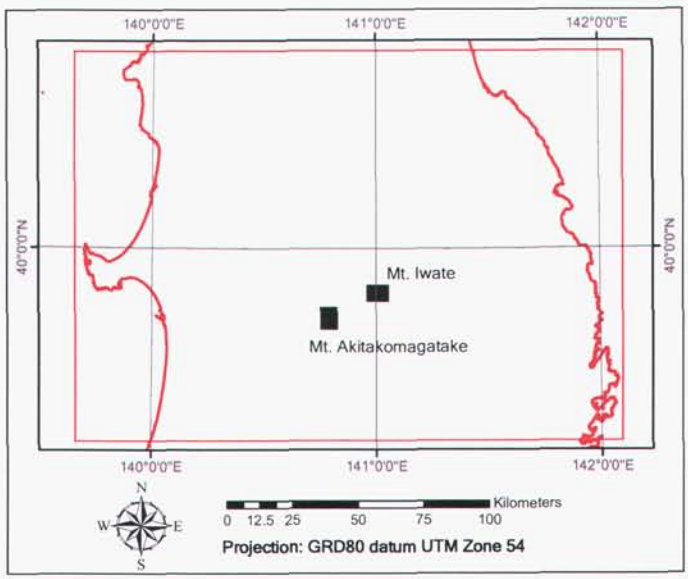


Fig. 3.12 Areas where elevation data are available on a 10 m grid system

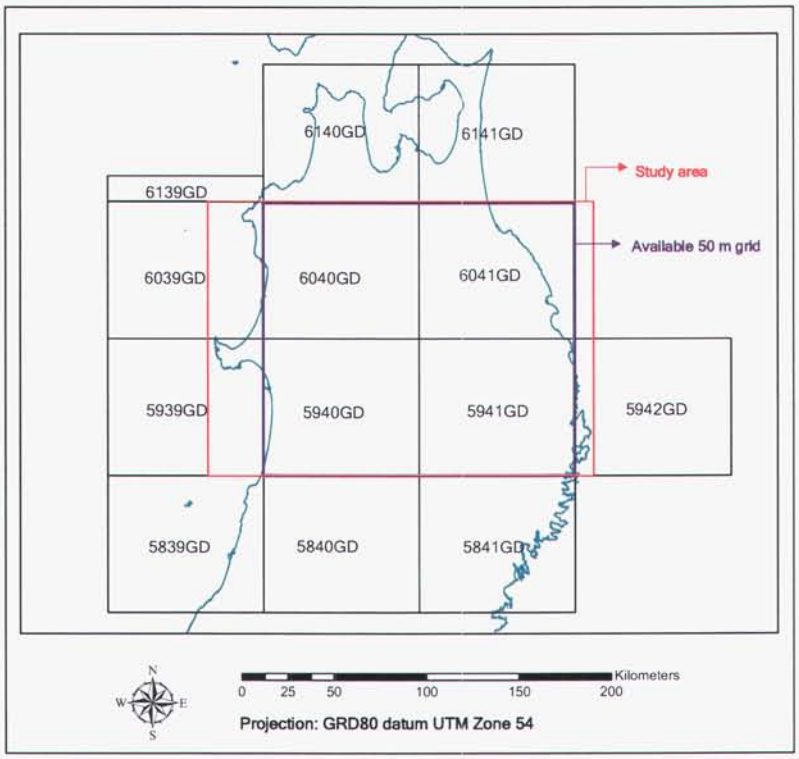


Fig. 3.13 Areas where elevation data are available on 50 m and 250 m grid systems
The 50 m grid mesh was converted to a UTM coordinate system. Note that the 50 m grid mesh did not have exactly 50 m × 50 m grid cells. Using the Inverse Distance Weighted (IDW) technique (McCoy, 2001) [1], a new 50 m grid mesh, which has correct 50 m × 50 m grid cells, was created. To estimate the elevation value at the center of each 50 m grid cell, four nearest points were used. The elevation value at the center of each 1 km square grid cell was calculated by averaging the elevation values at the closest nine 50 m cells from the center of the 1 km cell. Fig. 3.14 shows the elevation map obtained from this re-mapping and averaging process.

[1] IDW estimates cell values by averaging the values of sample data points in the vicinity of each cell. The closer a point is to the center of the cell being estimated, the more influence, or weight, it has in the averaging process. This method assumes that the variable being mapped decreases in influence with distance from its sampled location. The weight applied was inversely proportional to squared distance.

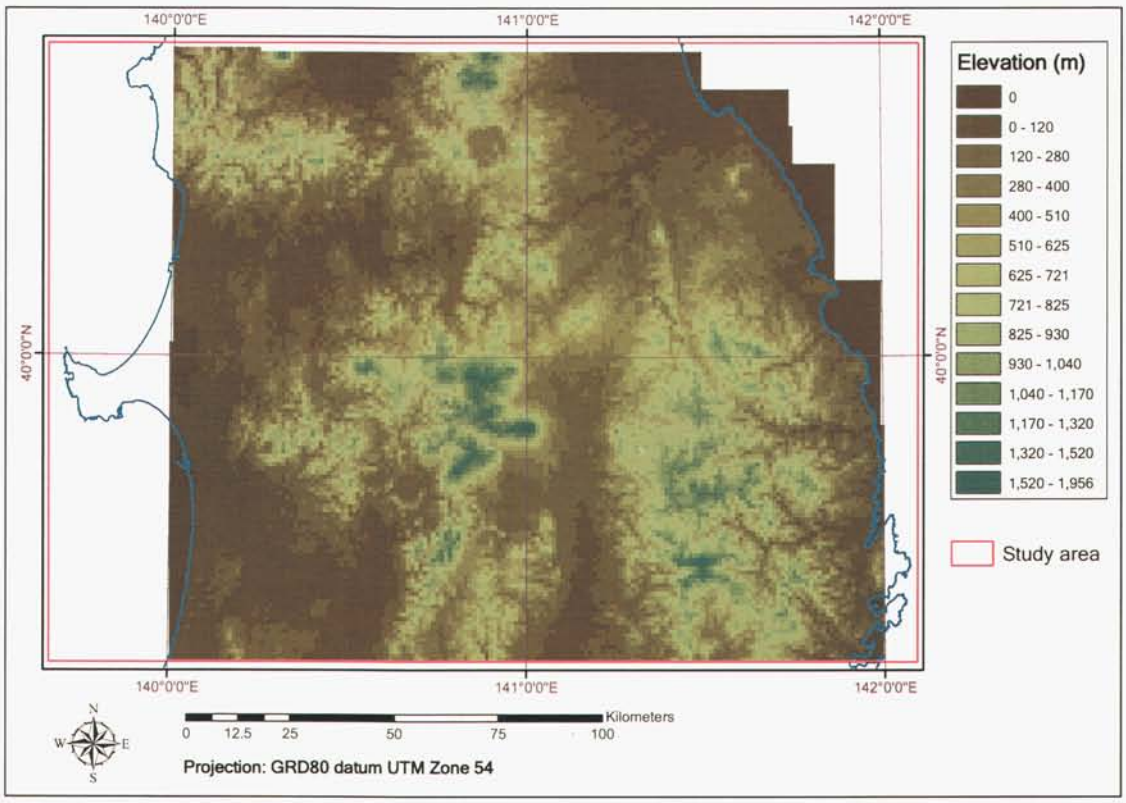


Fig. 3.14 Elevation model obtained for 1 km grid cells from 50 m grid mesh. Note it does not cover the whole study area

The 250 m grid mesh was used to estimate the elevation values at the centers of 1 km square cells that fell within the remaining part of the study area. The 250 m grid mesh was converted to a UTM coordinate system. Note that the 250 m grid mesh did not have exactly 250 m × 250 m grid cells. Using the IDW technique, a new 250 m grid mesh, which has correct 250 m × 250 m grid cells, was created. To estimate the elevation value at the center of each 250 m grid cell, four nearest points were used (Fig. 3.15).

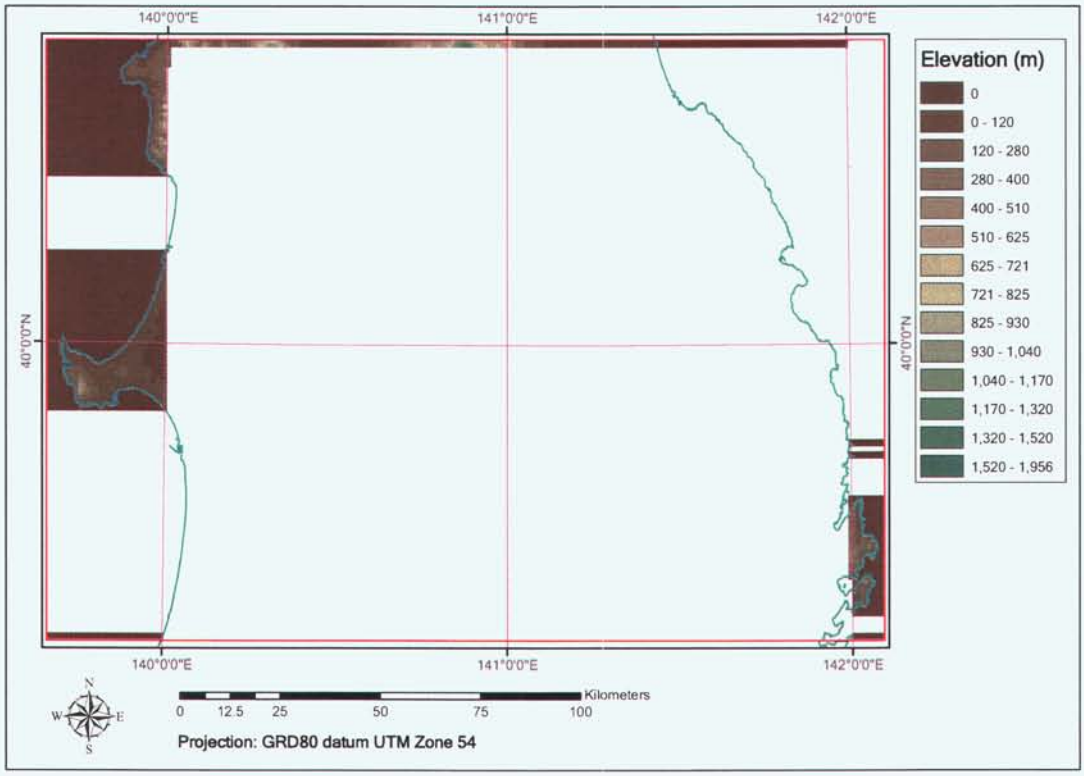


Fig. 3.15 Elevation model obtained for 1 km grid cells from 250 m grid mesh.

Finally, 1 km grid meshes from 50 m and 250 m grid meshes were combined to obtain the digital elevation model shown in Fig. 3.16.

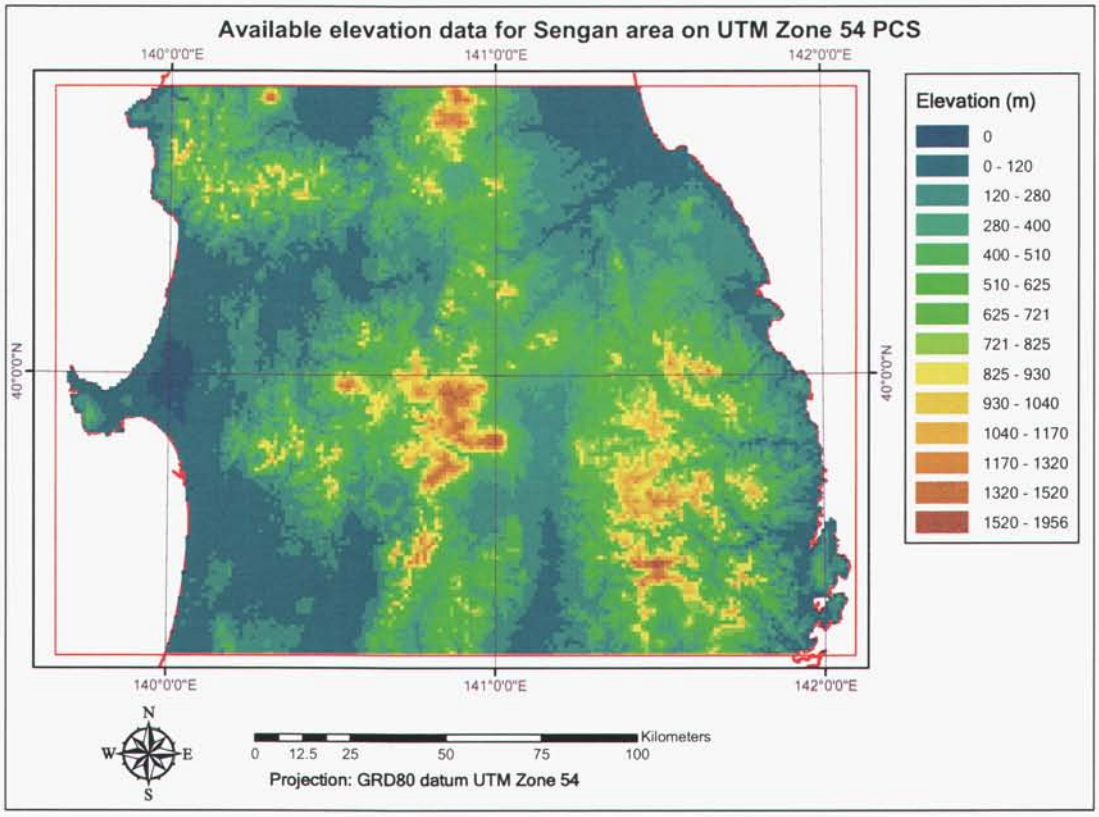


Fig. 3.16 Digital Elevation Model (DEM) obtained for Sengan region

3.9 Slope Magnitude

The new 50 m and 250 m grid meshes created for elevation modeling were used to calculate the slope magnitudes at the centers of 1 km square cells of the study region. Slope magnitude at the center of each 50 m cell was calculated using a Slope[2] function in ArcMap Spatial Analyst (McCoy, 2001). The slope magnitude value at the center of each 1 km square grid cell was calculated by averaging the slope magnitude values at the closest nine 50 m cells from the center of the 1 km cell. Fig. 3.17 shows the obtained slope magnitude map from the aforesaid process.

[2] The Slope function calculates the maximum rate of change between each cell and its neighbors – the maximum change in elevation over distance between the cell and its eight neighbors.

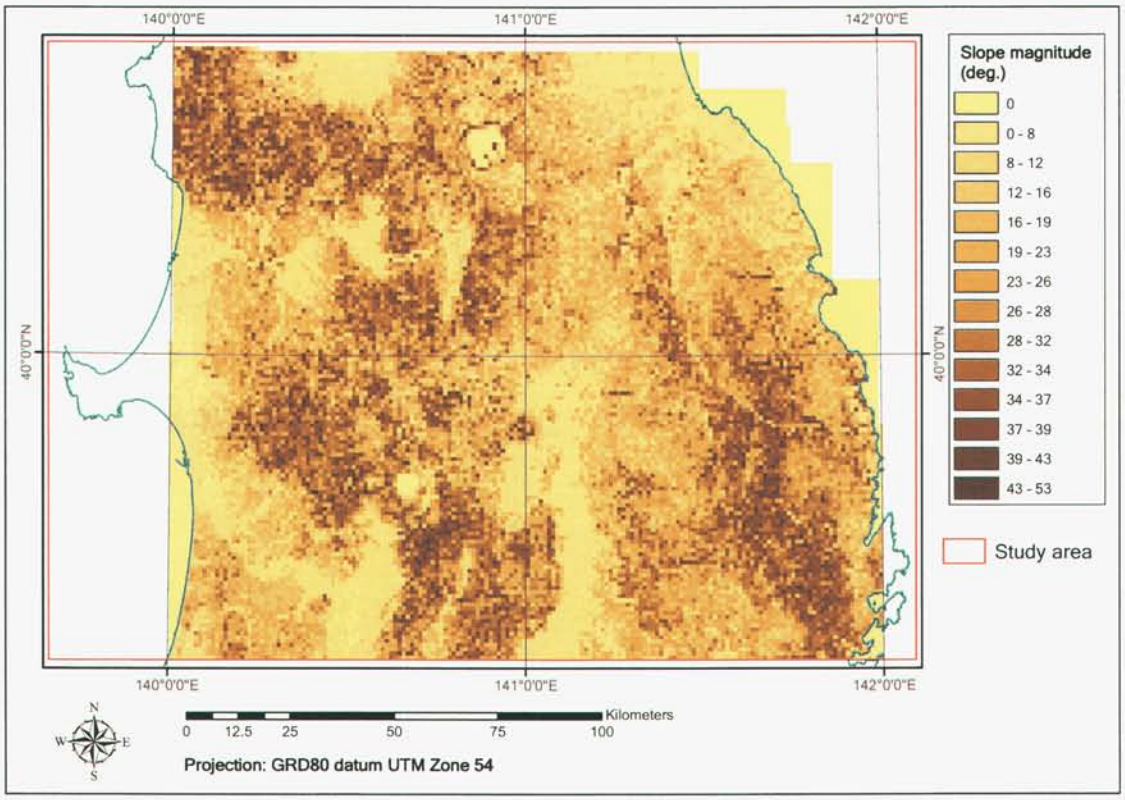


Fig. 3.17 Slope magnitude model obtained for 1 km grid cells from 50 m grid mesh. Note it does not cover the whole study area

The new 250 m grid mesh was used along with the same slope function in ArcMap Spatial Analyst to estimate the slope magnitude values at the centers of 250m square cells that fell within the remaining part of the study area. The slope magnitude value at the center of each 1 km square grid cell that fell within the remaining part of the study area was calculated by averaging the slope magnitude values at the closest nine 250 m cells from the center of the 1 km cell. Fig. 3.18 shows the obtained slope magnitude map from this calculation.

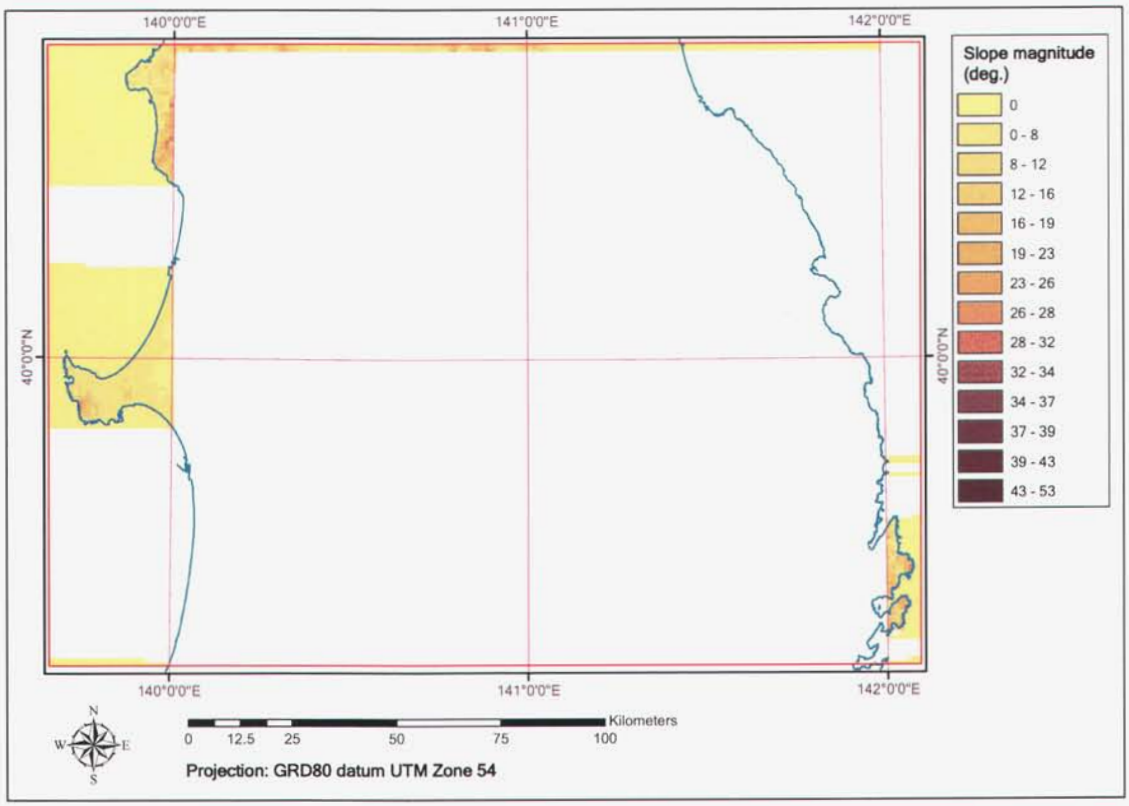


Fig. 3.18 Slope magnitude model obtained for 1 km grid cells from 250 m grid mesh

Finally, 1 km grid meshes from both the 50 m and 250 m grid meshes were combined to obtain the digital slope magnitude model shown in Fig. 3.19.

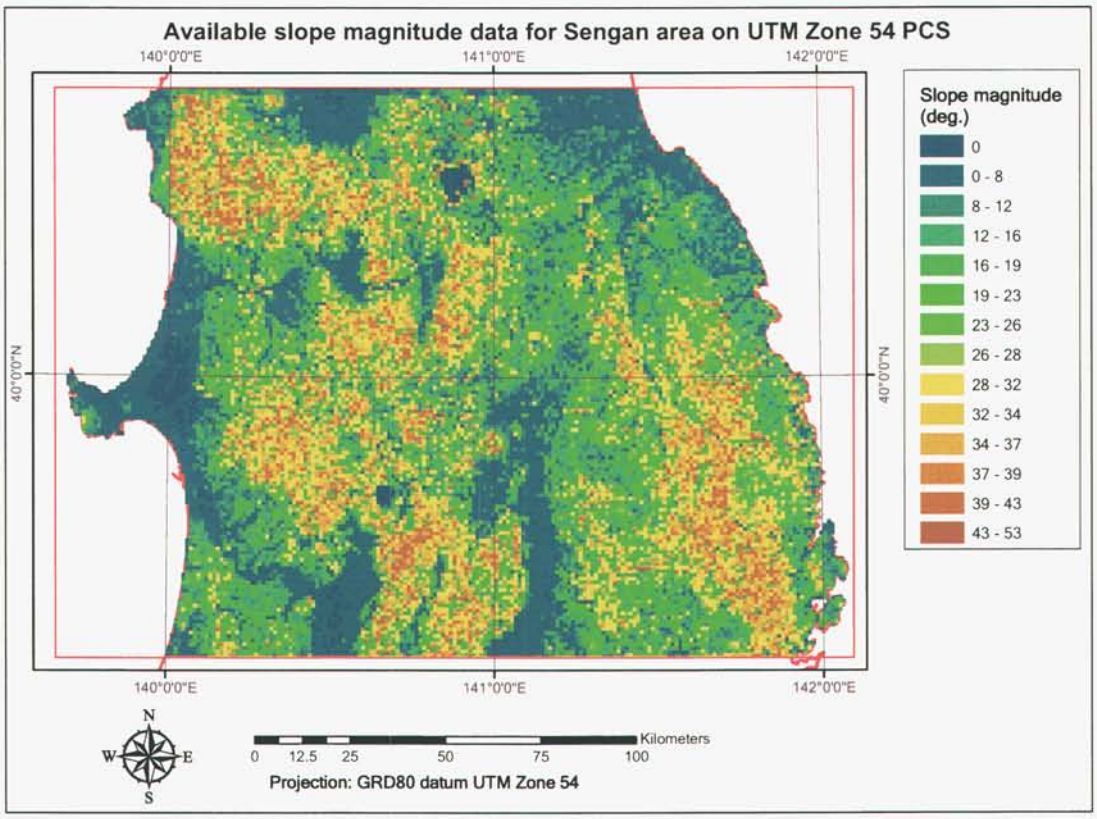


Fig. 3.19 Slope magnitude model obtained for Sengan region

3.10 Seismicity

The available point data coverage on seismicity for the study area was obtained from the following database:

- CD title '**ITM3-1**'\ 0323Hypocenter data 1926-2000 by Japan Meteorological Agency\ Hypocenter1926_2000.shp

The available seismic data for all magnitudes and depths during the period 1926 to 2000 were used to calculate the number of seismic occurrences per each 1 km square cell area. The obtained results are shown in Fig. 3.20.

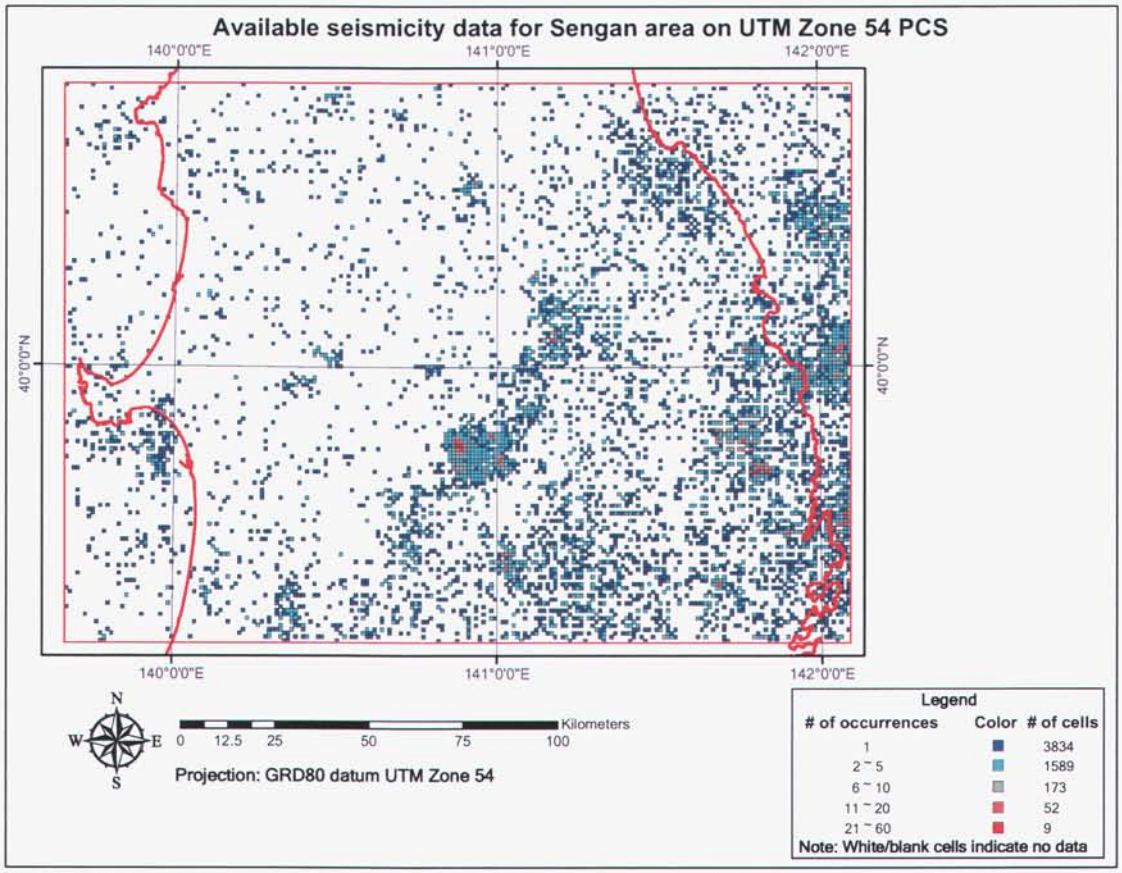


Fig. 3.20 Available data on seismic activity frequency for Sengan region on UTM zone 54 PCS

Also the maximum magnitude of seismicity for each 1 km square cell was extracted from the available data. The obtained results are shown in Fig. 3.21.

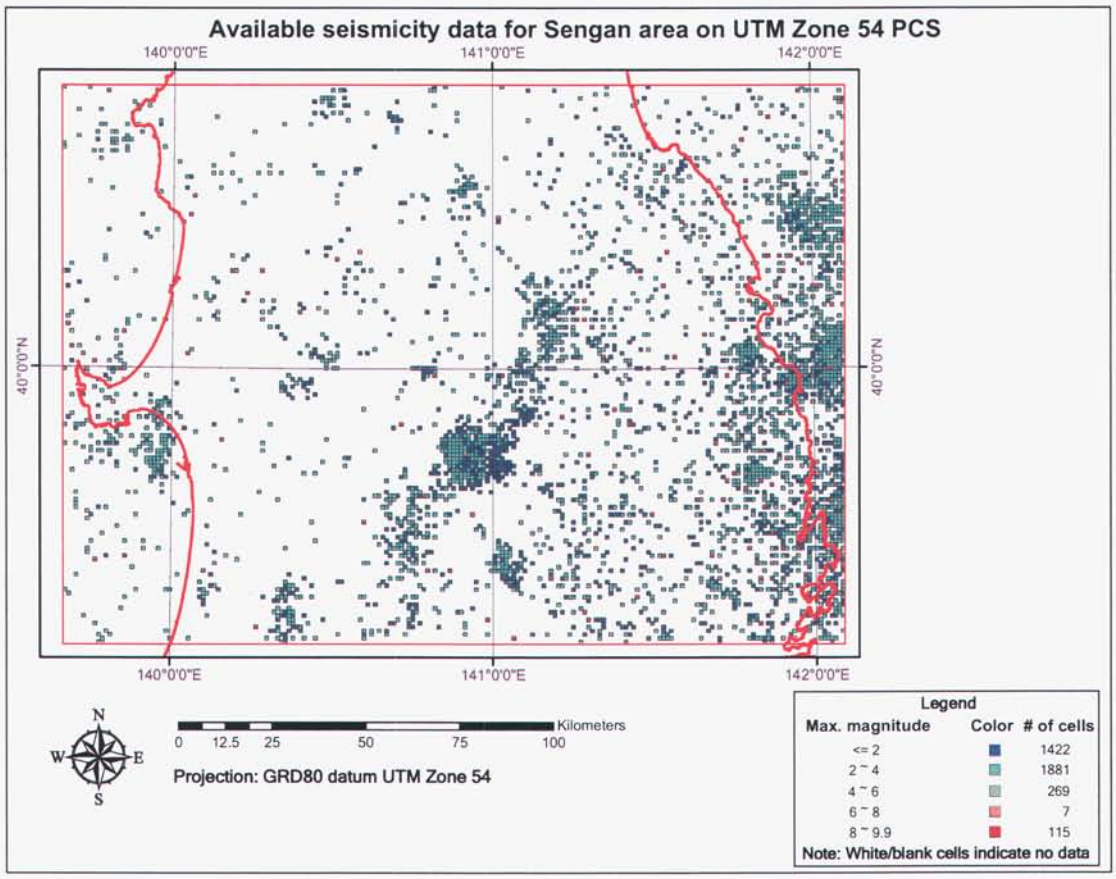


Fig. 3.21 Available data on seismic activity magnitude for Sengan region on UTM zone 54 PCS

3.11 Gravity Anomaly

The available point data coverage on gravity anomaly for the study area was obtained from the following database:

- CD title '**ITM3-3**'\ 0810Gravity CD-ROM of Japan\ 0810gsj.shp, 0810mmaj.shx & 0810nedo.shp

The Bouguer gravity anomaly (density = 2.67) was used to represent the gravitational anomaly for the study area. Note the unit used for the gravity anomaly is mGal. Fig. 3.22 shows the gravity anomaly map obtained on UTM zone 54 PCS for the study area.

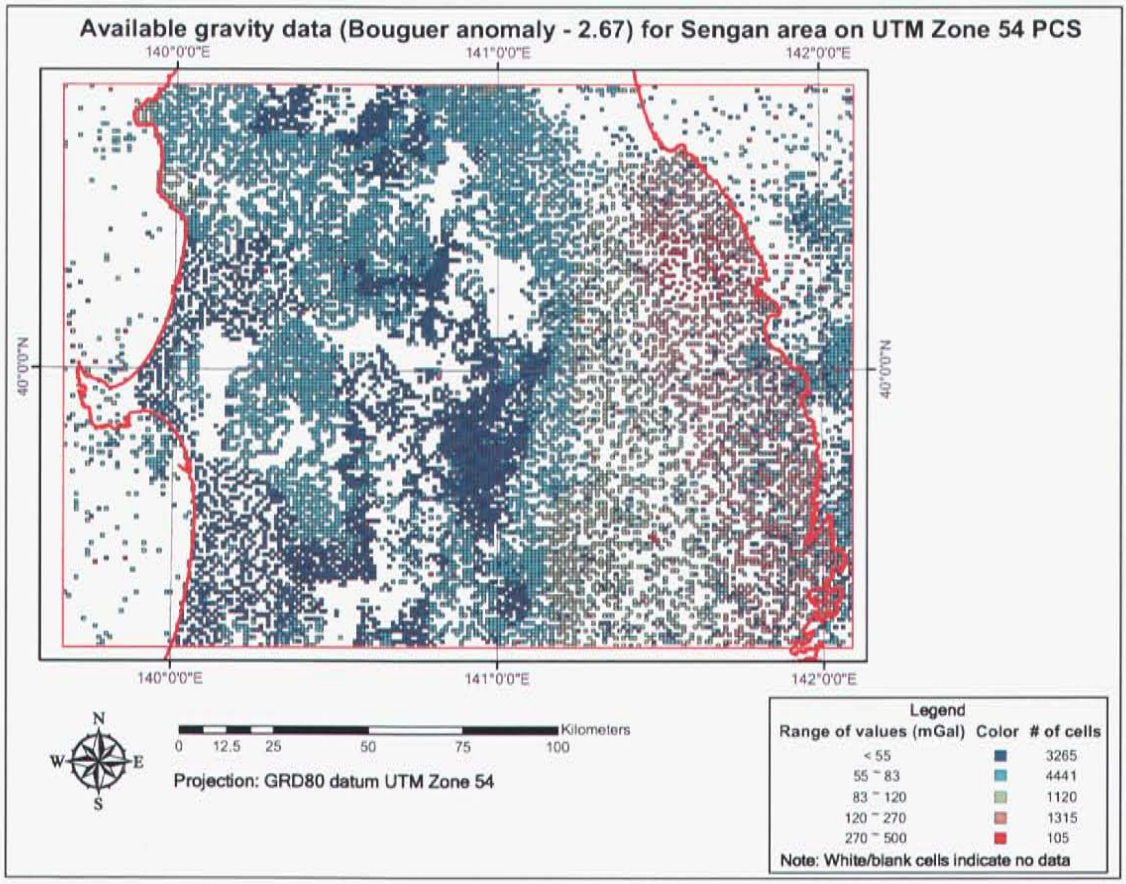


Fig. 3.22 Available gravity anomaly data for Sengan region on UTM zone 54 PCS

3.12 Magnetic Anomaly

The available point data coverage on magnetic anomaly for the study area was obtained from the following database:

- CD title '**ITM3-3**'\ 0814Magnetic anomaly map of east Asia\ Uea_point.shp

The magnetic anomaly value (unit = 0.1nT) is used in this report to represent Magnetic anomaly. Fig. 3.23 shows the magnetic anomaly map obtained on UTM zone 54 PCS for the study area.

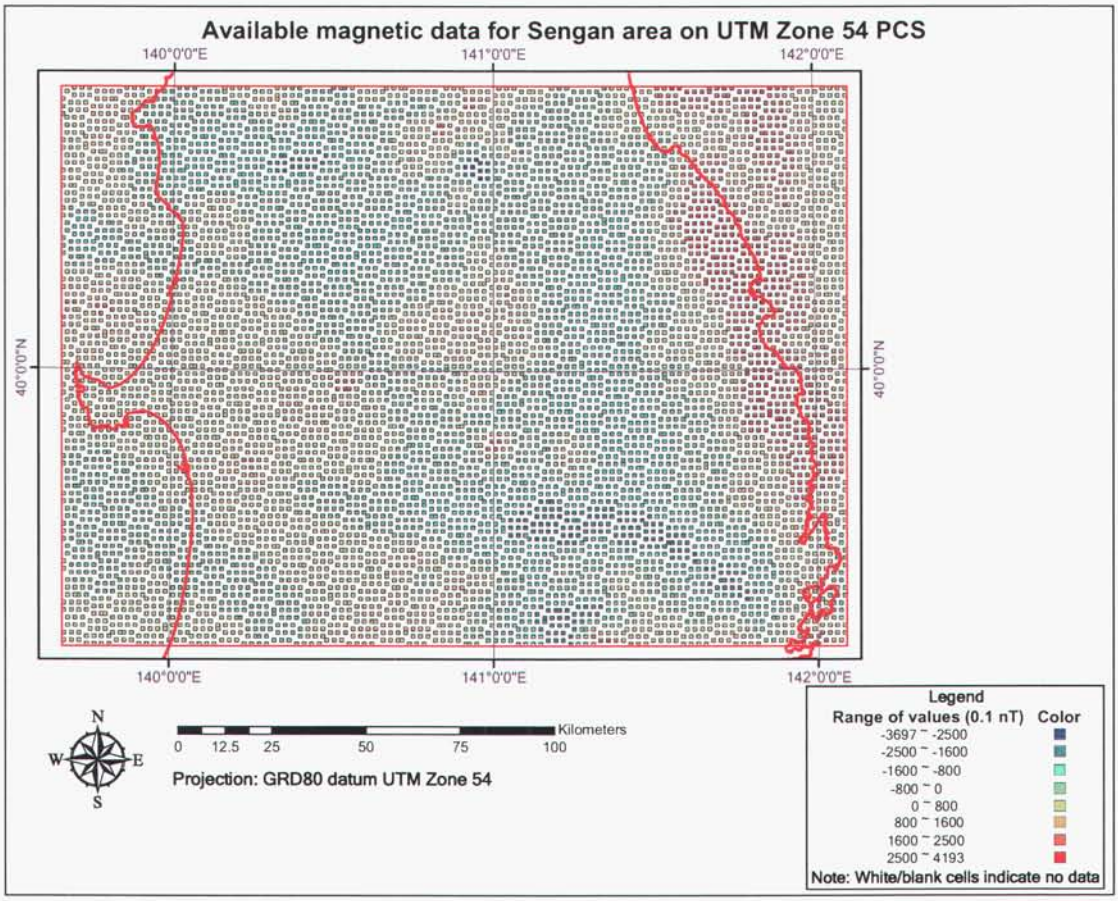


Fig. 3.23 Available magnetic anomaly data for Sengan region on UTM zone 54 PCS

3.13 Maximum Horizontal Strain

The available point data coverage on maximum horizontal strain for the study area was obtained from the following database:

- CD title '**ITM3-3**'\ 0840Horizonatal strain in Japan_1994-1883\

  max_shear_UTM_study_area.shp

Fig. 3.24 shows the maximum horizontal strain map obtained on UTM zone 54 PCS for the study area.

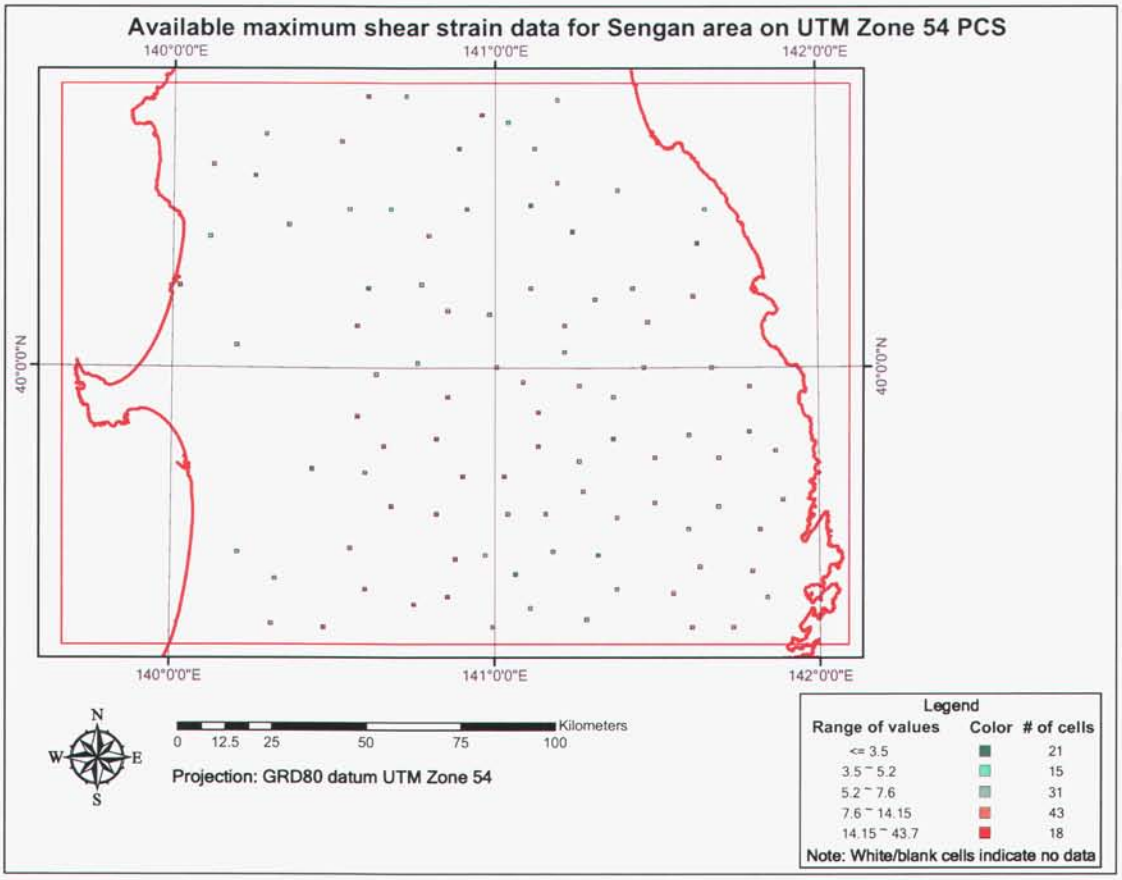


Fig. 3.24 Available maximum shear strain data for Sengan region on UTM zone 54 PCS.

3.14 Geologic Variables Correlated (Important) to Volcanism

The map obtained for the volcanic activity in the study area (Fig. 3.1) was visually compared to the map obtained for each geologic variable to determine whether the considered geologic variable is correlated to volcanic activity. **This comparison clearly revealed that the following geologic variables are correlated (or important) to volcanic activity: geothermal gradient, groundwater temperature, heat discharge, presence of volcanic rocks, presence of hydrothermal alteration and groundwater pH value.** The rest of the geologic variables were found to be either poorly or weakly correlated (or unimportant) to volcanic activity when examined across the study area. Some of these variables may show locally strong correlation with volcanism, but not when viewed across the entire study area. Therefore, the geostatistical and multivariate statistical analyses reported below were conducted only for the geologic variables that were found to be important to volcanic activity. Table 3.1 provides the values of basic statistical parameters for the geologic variable data available for Sengan region.

Table 3.1 Values of basic statistics for the geologic variable data available for Sengan region.

	# of data	Min.	Max.	Mean	Median	Variance	Coefficient of variation
Temperature (°C)	727	2	227	44.9464	43	743.3043	0.6066
pH	594	1.0	10.2	7.1867	7.6	2.6271	0.2255
Geothermal gradient (C°/km)	127	11.0	279.0	101.4016	91.0	3566.8295	0.5890
Heat discharge] ($\log_{10}$ [μWm^{-2})])	308*	0.0	4.3	2.1338	2.3	1.4206	0.5586

* Heat discharge data were available on a raster image having a cell size of latitude 5″ and longitude 5″. The value at each cell was replaced by four data points having the same value.

4. 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. For some cells on the map, data are available for the considered geologic variable, and for some other cells in the map, data are lacking for the considered variable. Cell locations for which data are available for one geologic variable vary with another geologic variable. Therefore, when all of the aforementioned 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, variogram contours and directional variograms. Variogram modeling then can be performed to capture the essential properties depicted by the directional variograms. The obtained 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 aforementioned 6 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, 1999). 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.

Table 4.1 Values of basic statistics for the geologic variable data used to perform geostatistical analysis

	# of data	Min.	Max.	Mean	Median	Variance	Coefficient of variation
Temperature (°C)	999	2	227	43.3203	42	725.5887	0.6218
pH	905	1.0	10.2	7.2703	7.6	2.4549	0.2155
Geothermal gradient (C°/km)	191	8.0	334.0	101.0524	85.0	3970.2393	0.6235
Heat discharge] ($\log_{10}$ [μWm^{-2})])	452*	0.0	4.3	2.0673	2.3	1.4610	0.5847

* Heat discharge data were available on a raster image having a cell size of latitude 5″ and longitude 5″. The value at each cell was replaced by four data points having the same value.

Table 4.1 provides the values of basic statistics for the geologic variable data used to perform geostatistical analysis. 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 aforementioned 6 geologic variables. The results obtained for each geologic variable is given below in this section under a separate heading for each variable.

4.1 Groundwater Temperature

Figure 4.1 shows the extended database region used to perform variogram calculations and kriging for the groundwater temperature in the study area. 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 and contours shown in Figs. 4.2 and 4.3, respectively.

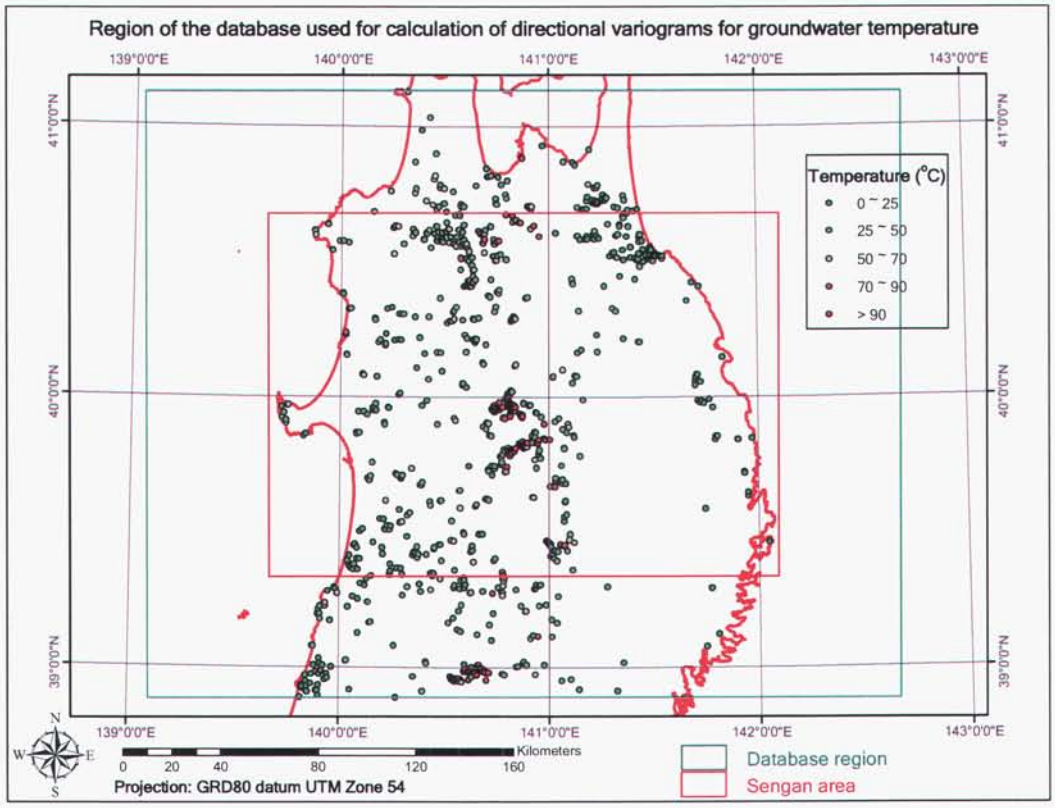


Fig. 4.1 Region of the database used for calculation of directional variograms for groundwater temperature

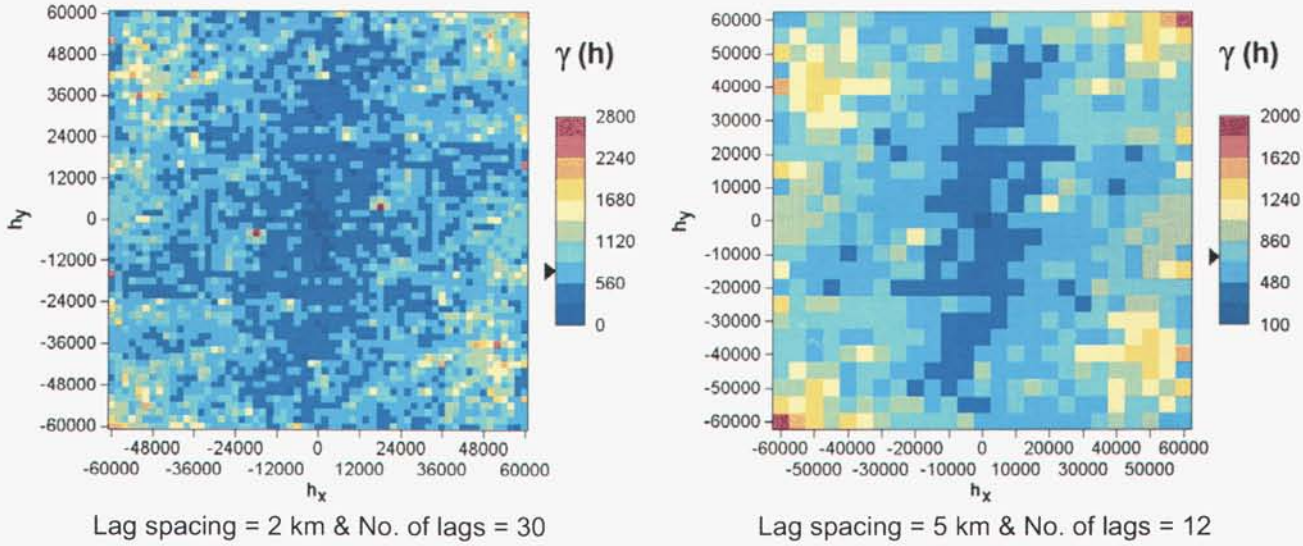


Fig. 4.2 Variogram surface plots of groundwater temperature at different lag spacings and number of lags. Note: the black triangle shown on the color legend indicates the total variance value.

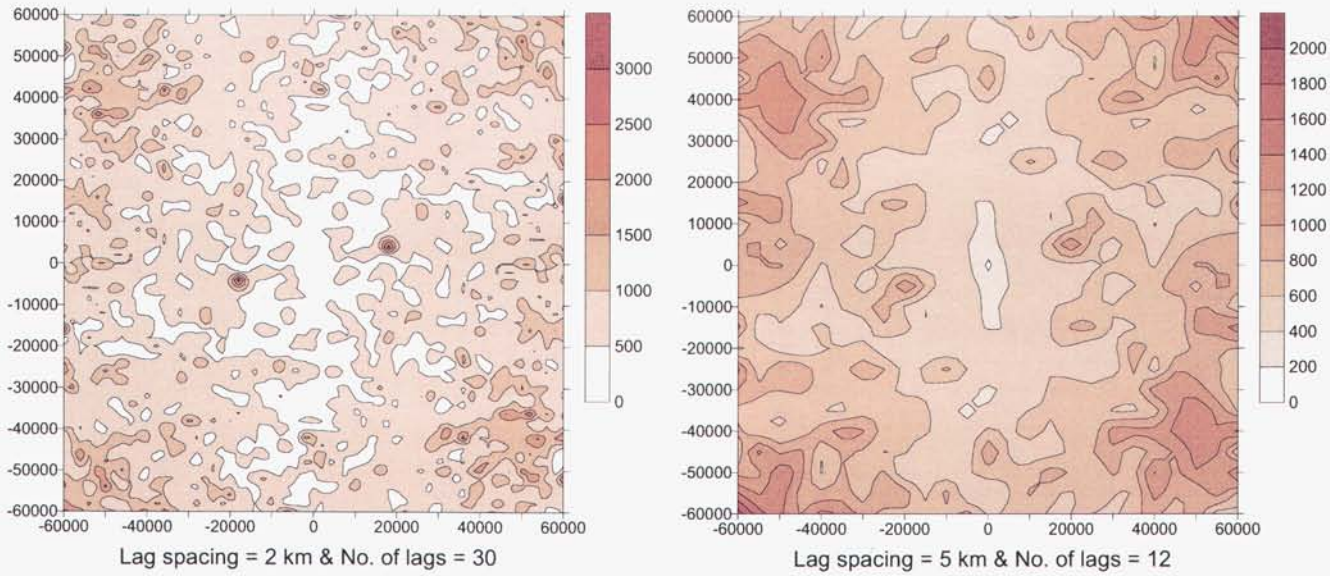
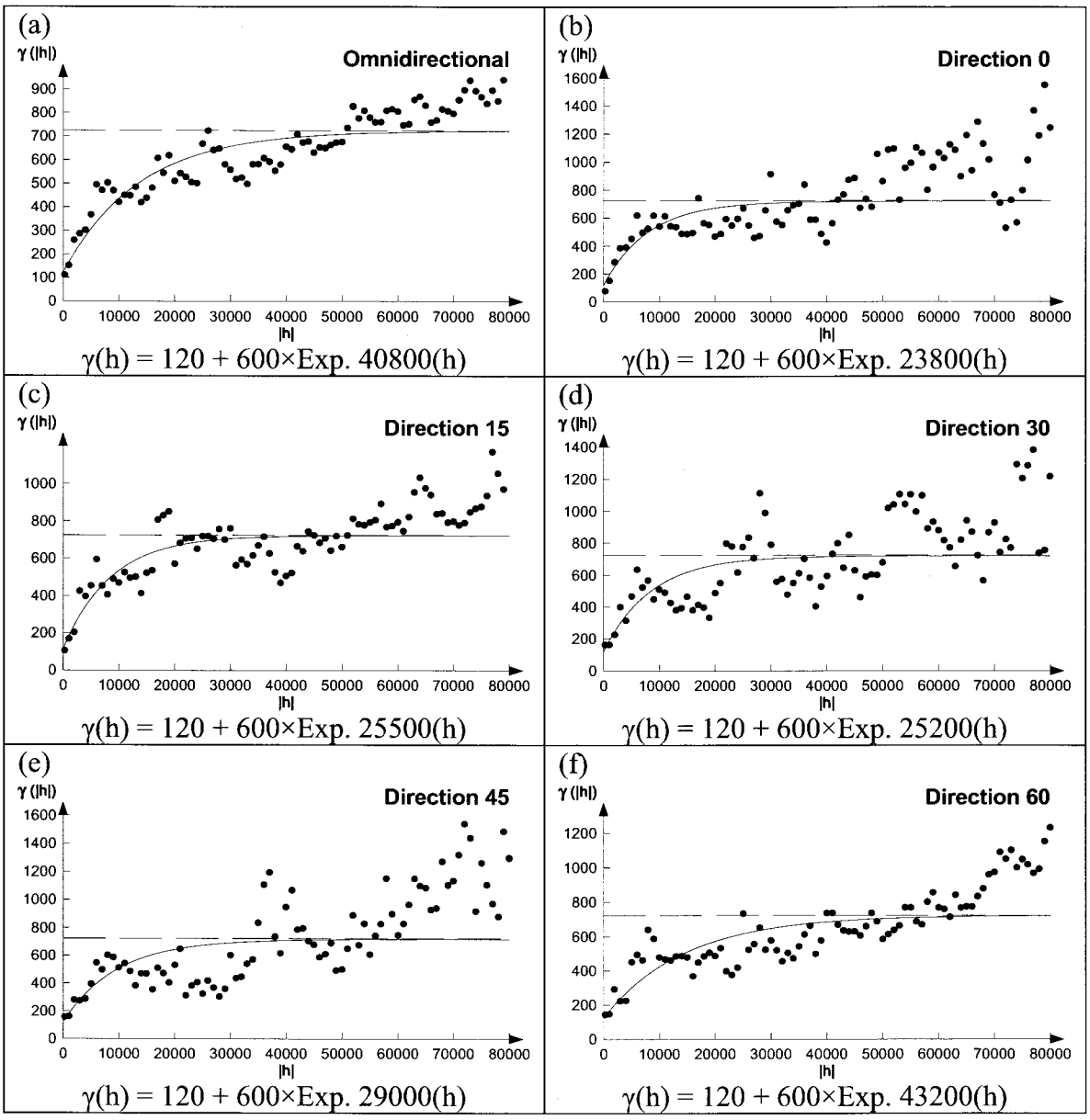


Fig. 4.3 Variogram contour plots for groundwater temperature at different lag spacings and number of lags

Variogram surface and contour plots show that the major axis direction for correlation distance is around N 5 to 15° W. Figure 4.4a 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° counter-clockwise starting at 0° (East) using a lag spacing of 1 km (Figs. 4.4b

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 the goodness of fit test results. The obtained variogram models for directional variograms are shown in Figs. 4.4b 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.



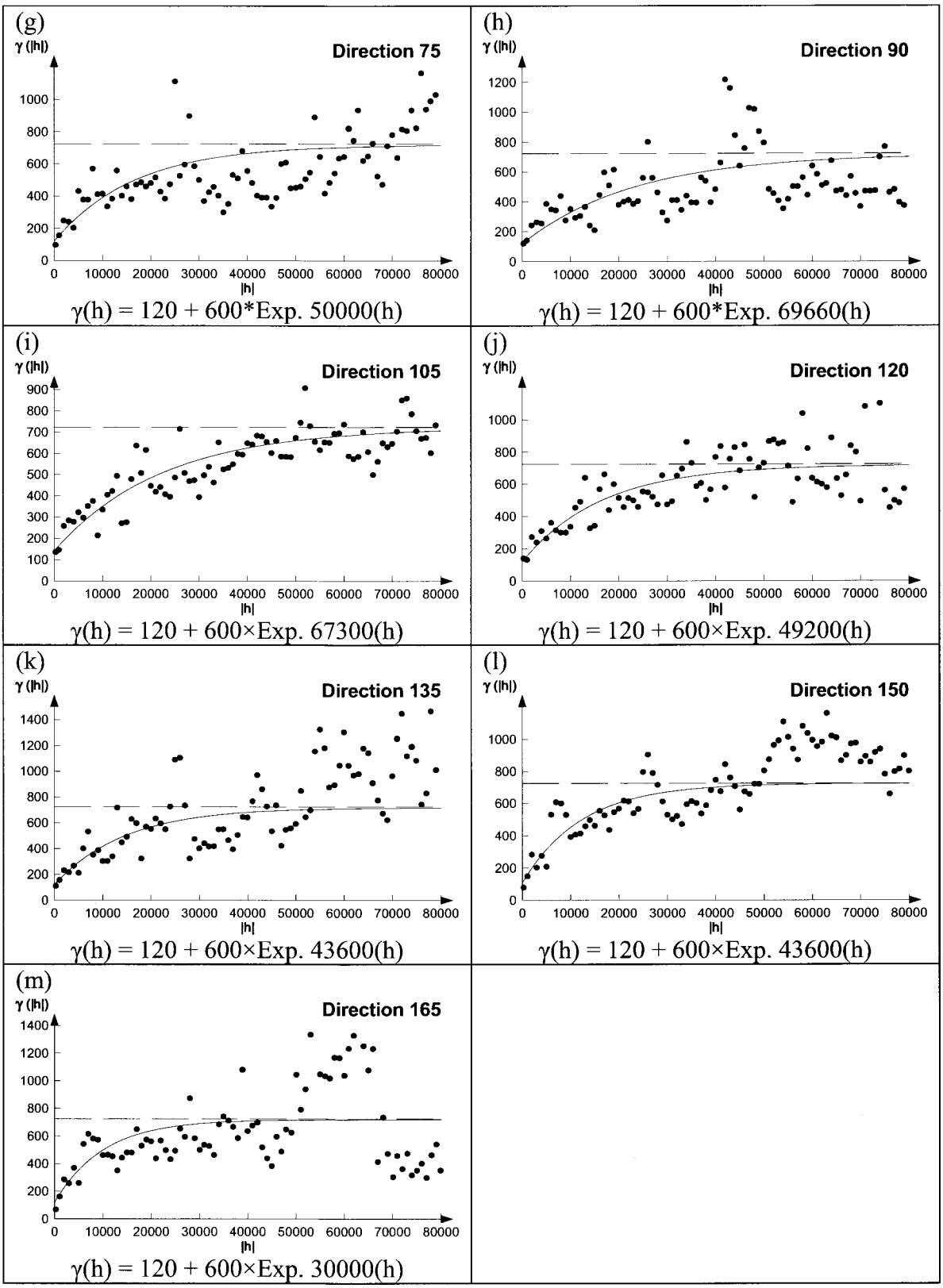


Fig. 4.4 Omnidirectional variogram (a) and directional variograms (b) through (m) for groundwater temperature at 1 km lag spacing

The correlation distance obtained for every 15-degree is plotted on a polar coordinate system in Fig. 4.5. 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.

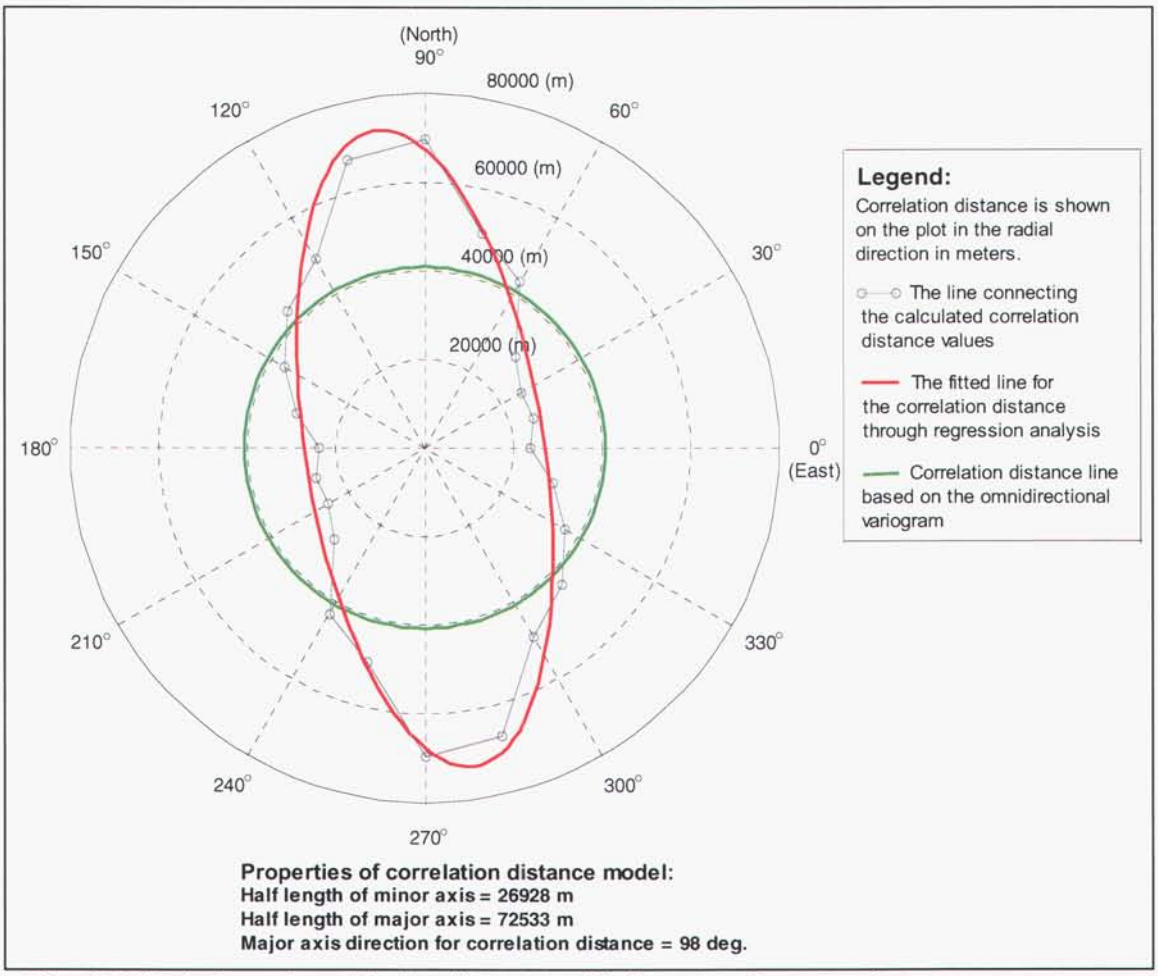


Fig. 4.5 Variation of correlation distance with direction for groundwater temperature

Properties of the aforementioned variogram model were used to perform kriging using ArcGIS Geostatistical Analyst (Johnston et al, 2001). Fig. 4.6 shows the map obtained for the predicted values of groundwater temperature.

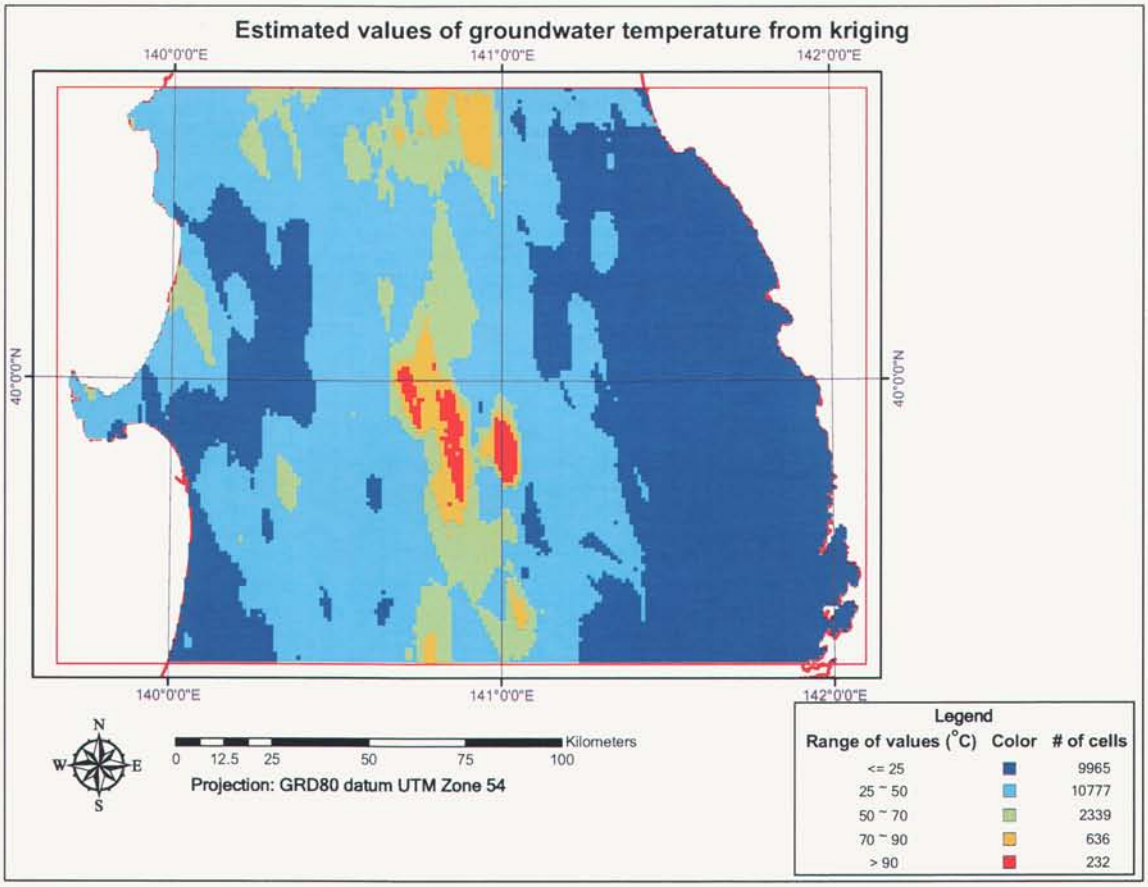


Fig. 4.6 Estimated values of groundwater temperature from kriging for the study area.

The map obtained for the standard error (the square root of the kriging variance) of the prediction is shown in Fig. 4.7. Note that the standard error (uncertainty) increases in areas with a lack of data.

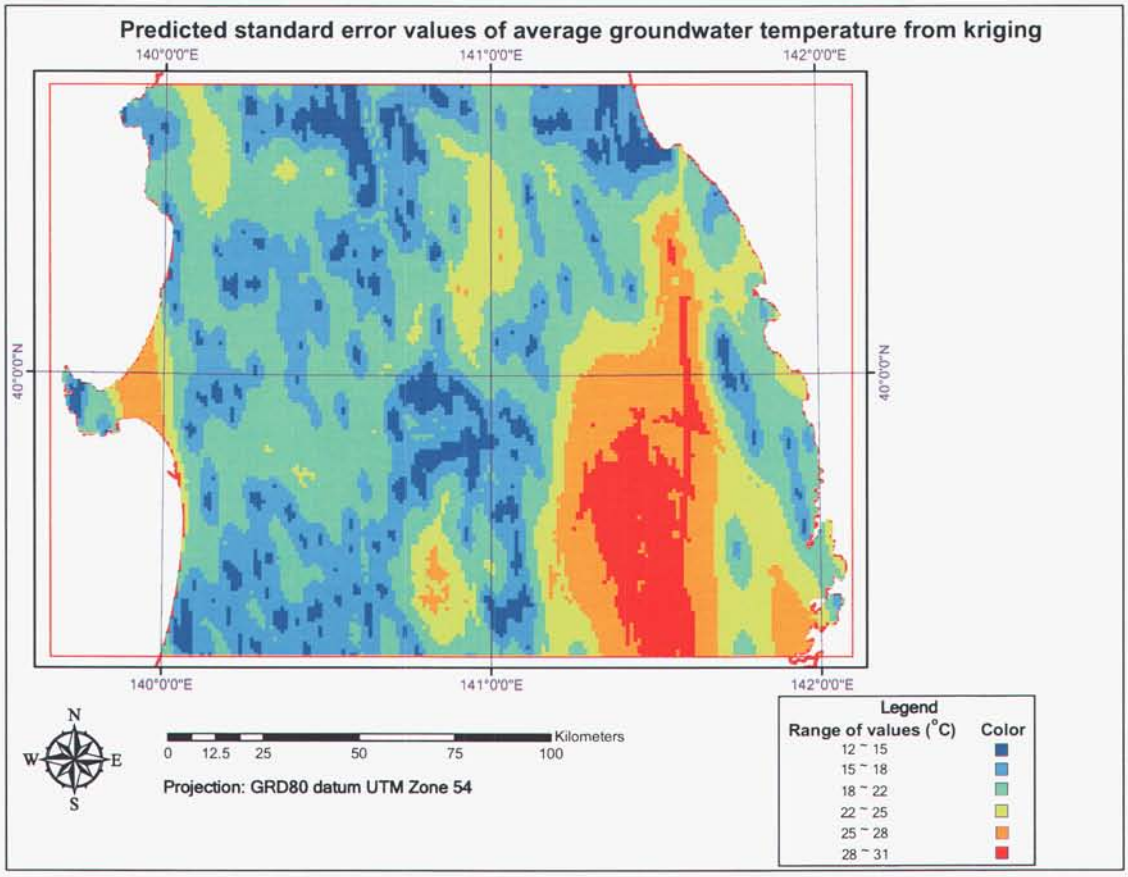


Fig. 4.7 Predicted standard error values of average groundwater temperature from kriging for the study area.

4.2 Groundwater pH

Figure 4.8 shows the extended database region used to perform variogram calculations and kriging for the groundwater pH in the study area. In the extended data base region groundwater pH data were available for 905 one km square cells. These data were used to obtain the variogram surface and contours shown in Figs. 4.9a and b, respectively.

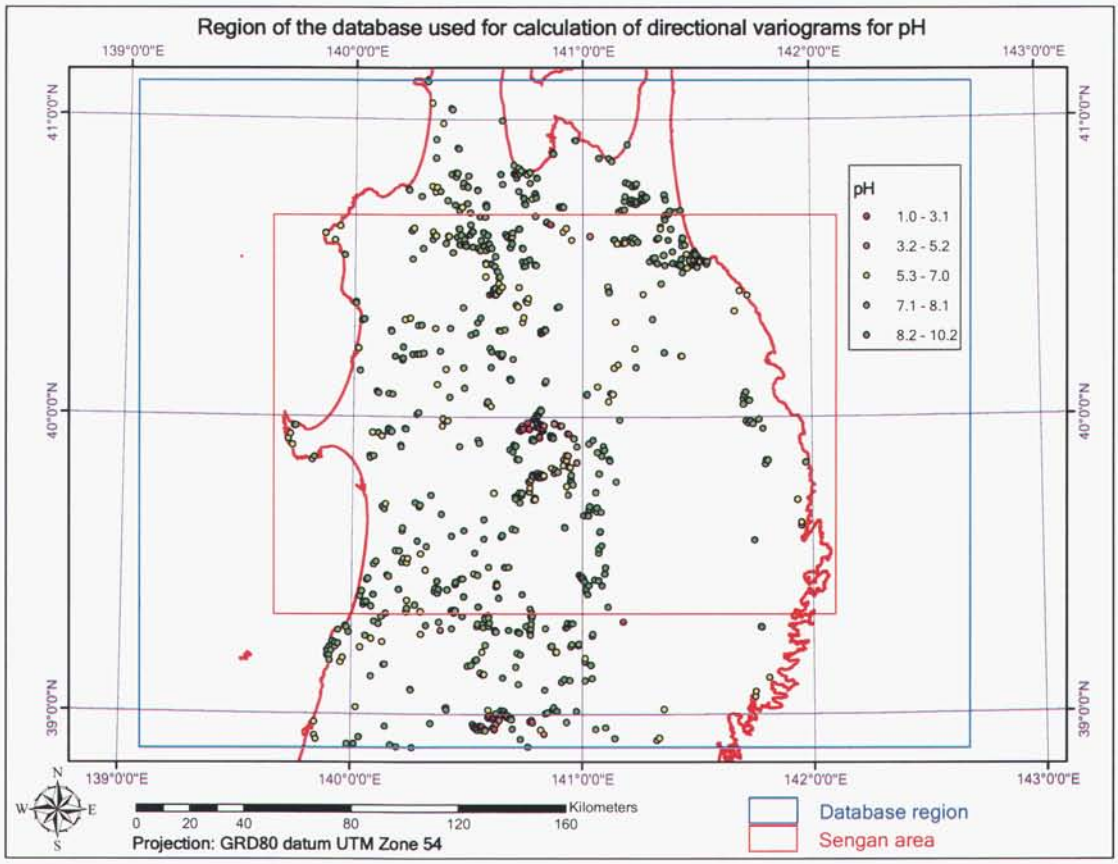
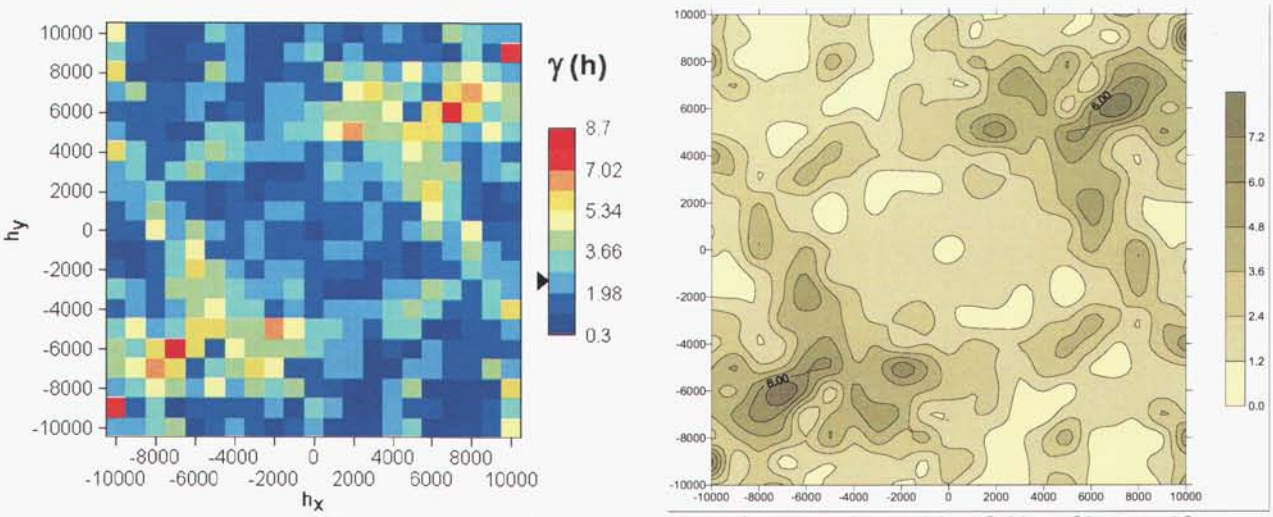


Fig. 4.8 Region of the database used for calculation of directional variograms for groundwater pH

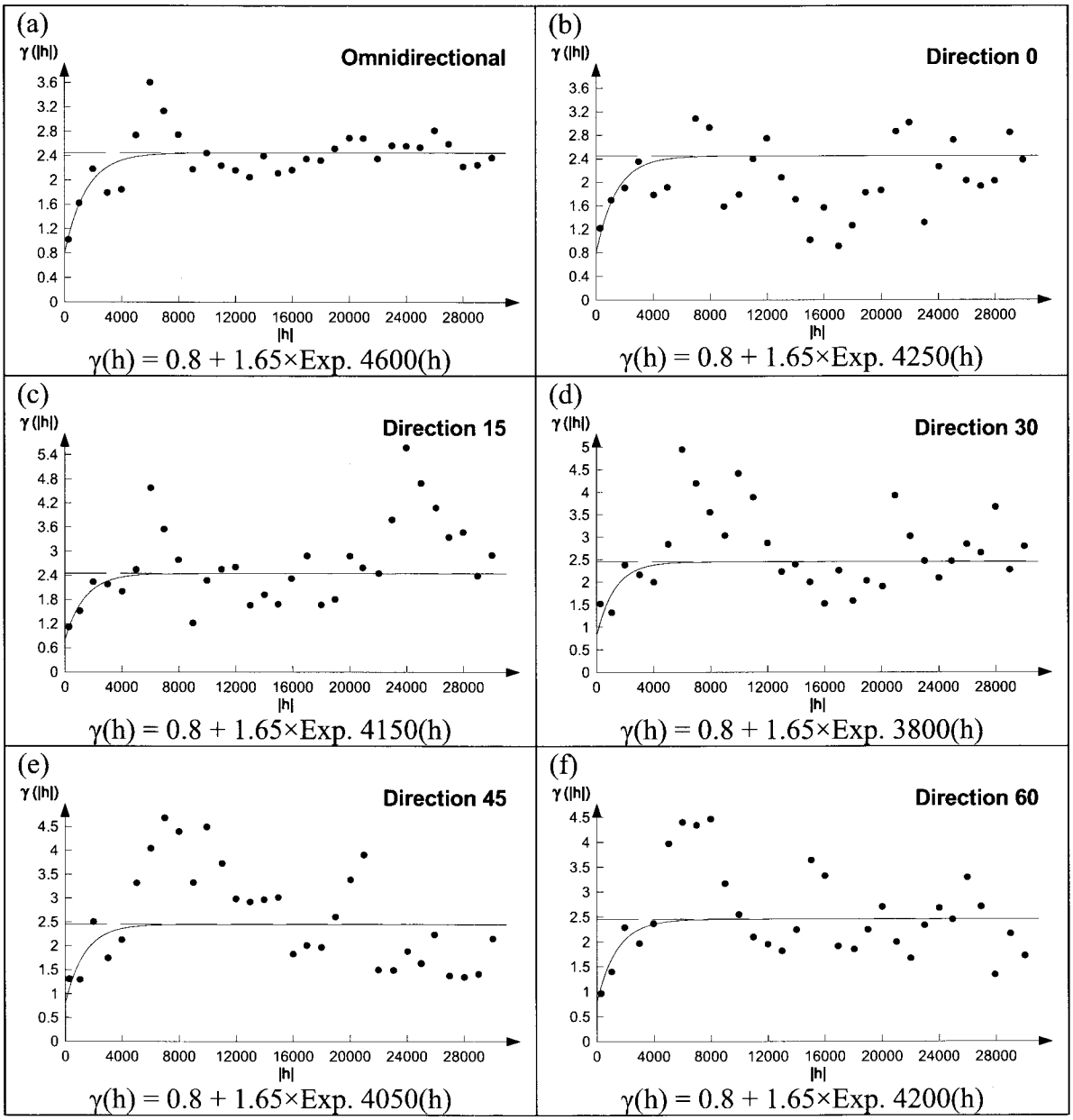


Lag spacing = 1 km & No. of lags = 10
(a)

Lag spacing = 1 km & No. of lags = 10
(b)

Fig. 4.9 Variogram (a) surface and (b) contour plots for groundwater pH at 1 km lag spacing and number of lags = 10. Note: In Fig. 4.9a, the black triangle shown on the color legend indicates the total variance value.

Variogram surface and contour plots show that the major axis direction for correlation distance is around N 30 to 40° W. Figure 4.10a shows the variogram (1 km lag spacing) calculated for groundwater pH assuming isotropic spatial variation (omni directional) in the considered region. Directional variograms were calculated at every 15° counter-clockwise starting at 0° (East) using a lag spacing of 1 km (Figs. 4.10b through m). The same directional parameters were used as for the groundwater temperature in calculating the directional variograms. 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 the goodness of fit test results. The obtained variogram models for directional variograms are shown in Figs. 4.10b through m. Note that the nugget and partial sill values for the variograms were found to be 0.8 and 1.65, respectively. Correlation distances were found to vary with the direction.



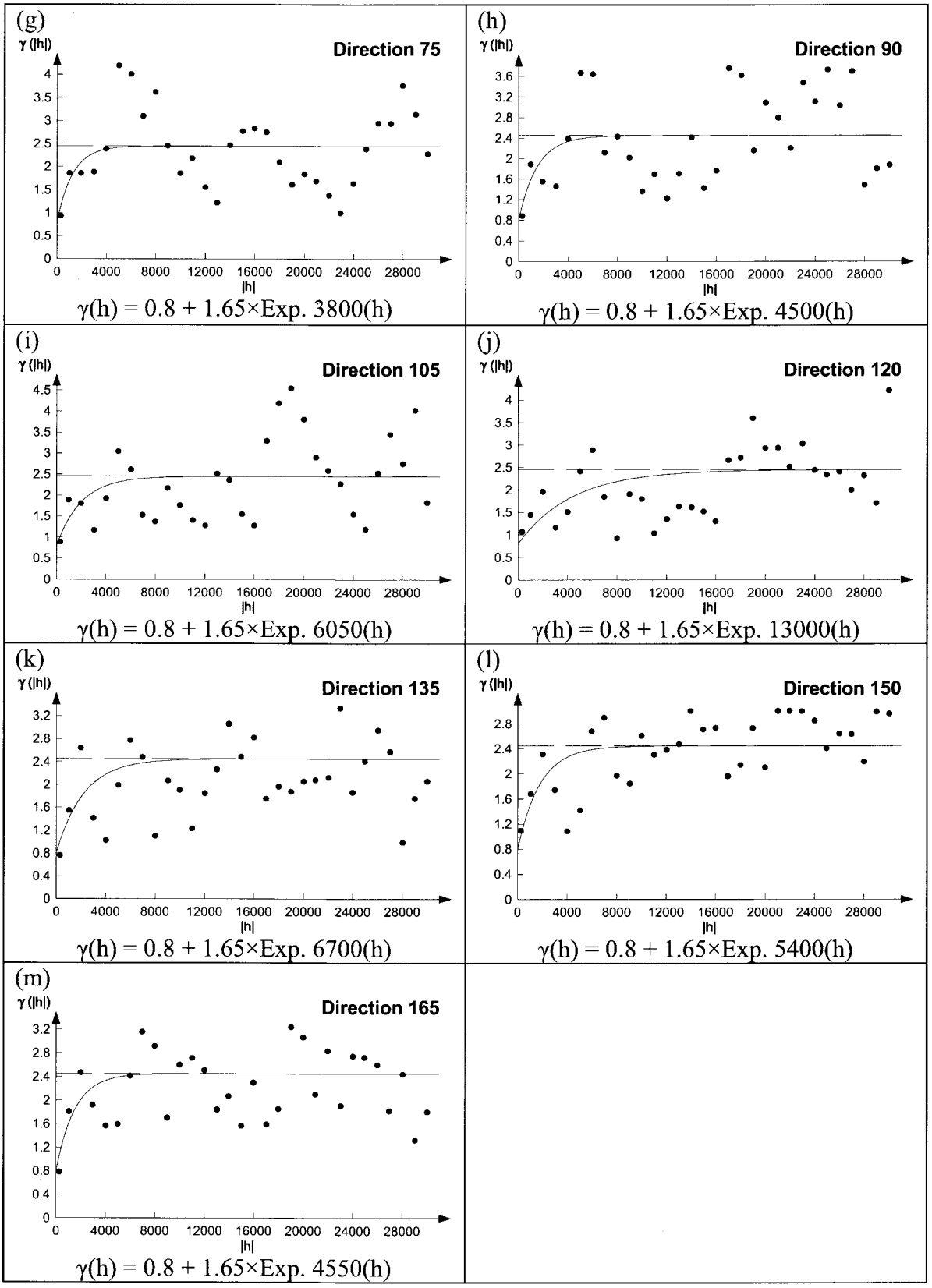


Fig. 4.10 Omnidirectional variogram (a) and directional variograms (b) through (m) for groundwater pH at 1 km lag spacing

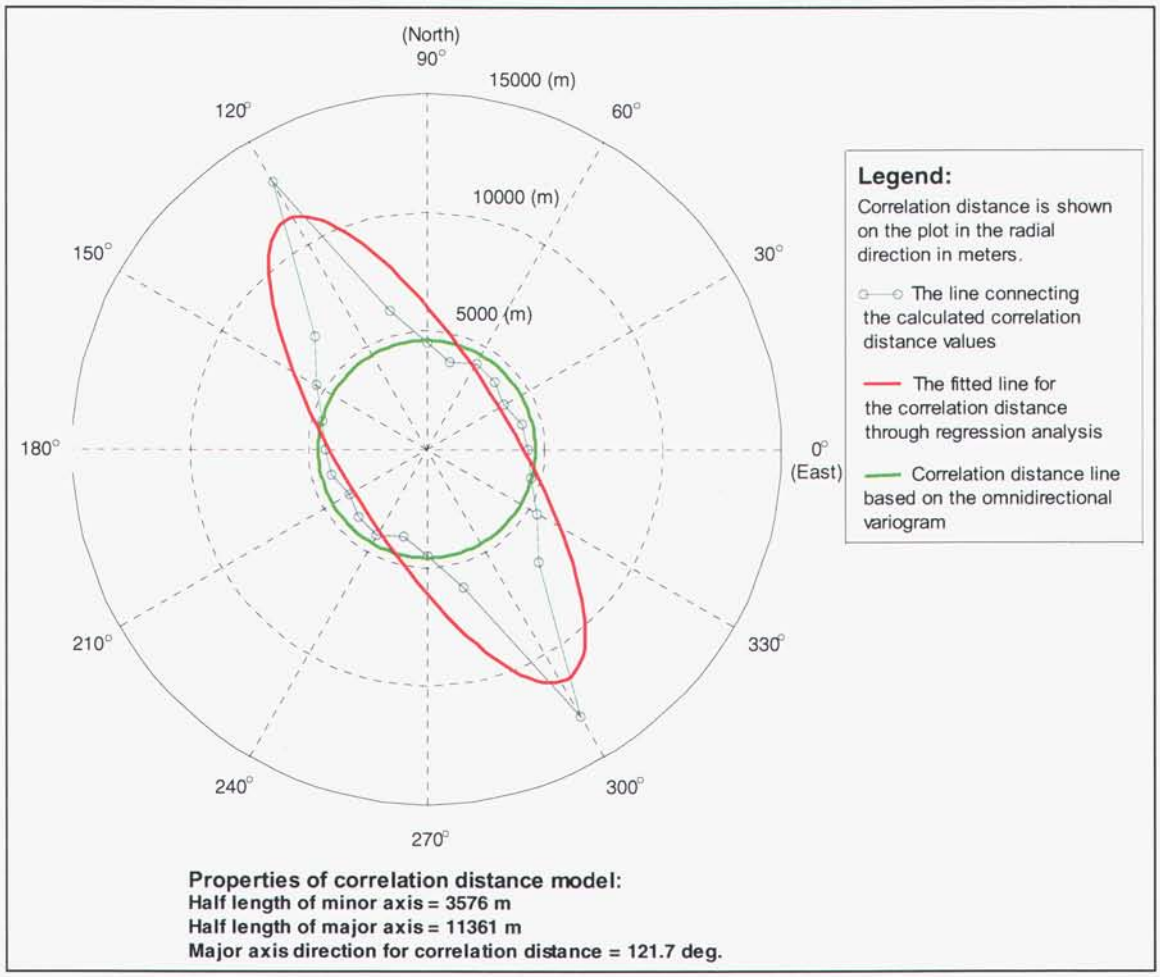


Fig. 4.11 Variation of correlation distance with direction for groundwater pH

The correlation distance obtained for every 15-degree is plotted on a polar coordinate system in Fig. 4.11. 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 pH. Major axis direction, and semi-major and semi-minor axis lengths obtained for the correlation distance model are also given in the figure. These lengths are considerably less than those calculated for the groundwater temperature.

Properties of the aforementioned variogram model were used to perform kriging using ArcGIS Geostatistical Analyst (Johnston et al, 2001). Fig. 4.12 shows the map obtained for the predicted values. The map obtained for the standard error (the square root of the kriging variance) of the prediction is shown in Fig. 4.13. Note that the standard error (uncertainty) increases in areas with a lack of data.

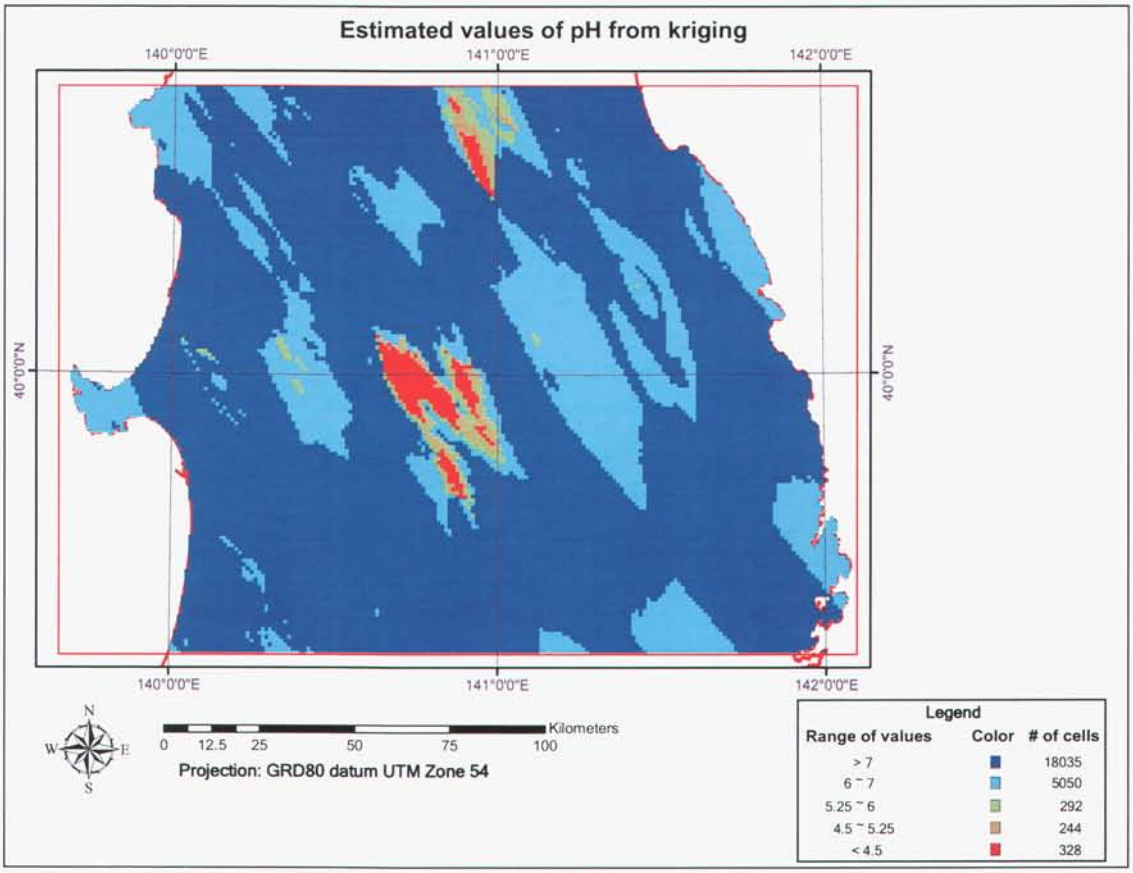


Fig. 4.12 Estimated values of groundwater pH from kriging for the study area.

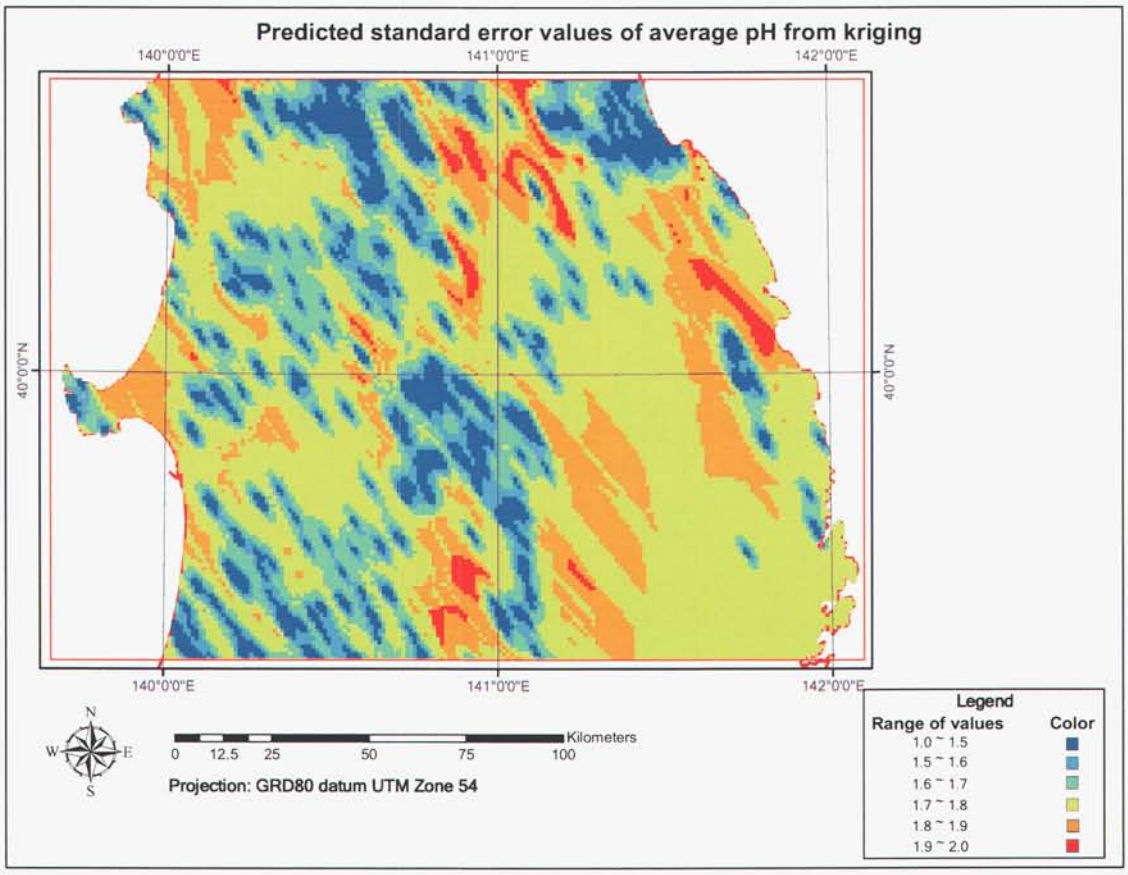


Fig. 4.13 Predicted standard error values of average groundwater pH from kriging for the study area.

4.3 Geothermal Gradient

Figure 4.14 shows the extended database region used to perform variogram and kriging modeling for the geothermal gradient in the study area. In the extended database region geothermal gradient data were available for 191 one km square cells. These data were used to obtain the variogram surface and contours shown in Figs. 4.15a and b, respectively. The significantly smaller number of data for the geothermal gradient relative to groundwater temperature and pH have an obvious impact on the calculation of the variogram surface maps (Figure 4.15).

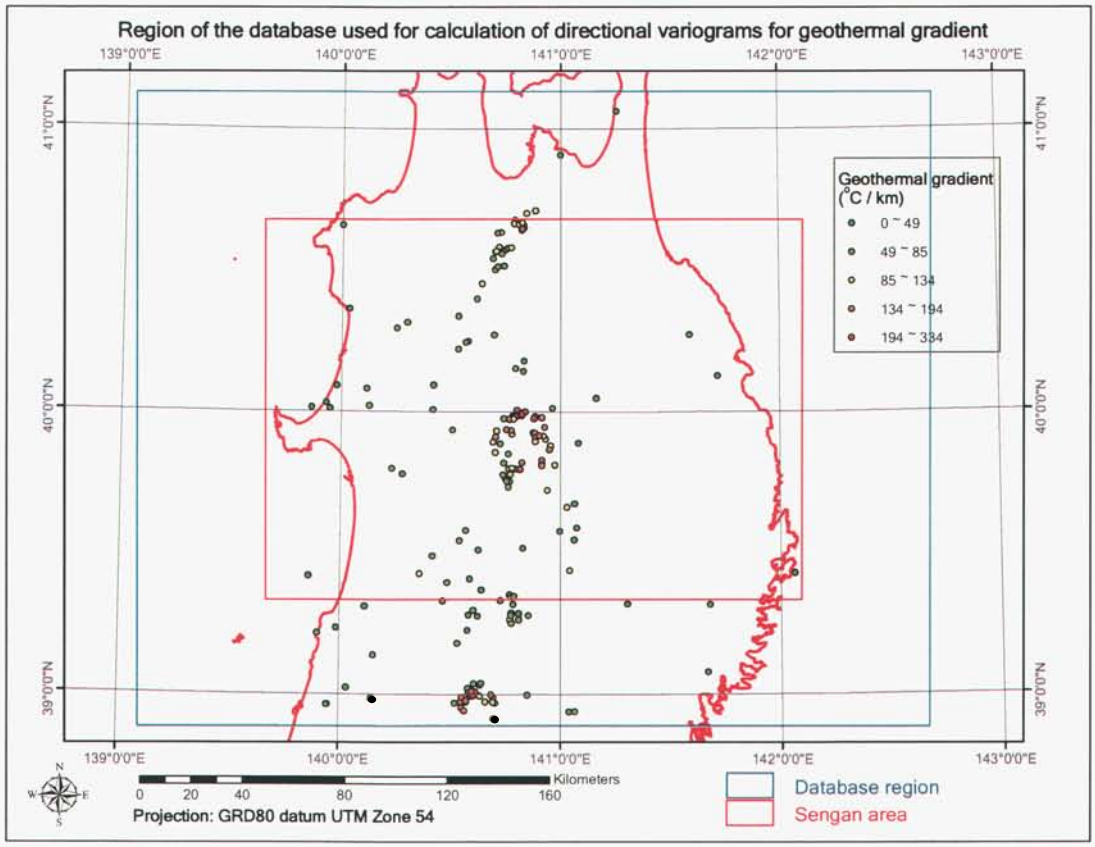


Fig. 4.14 Region of the database used for calculation of directional variograms for geothermal gradient.

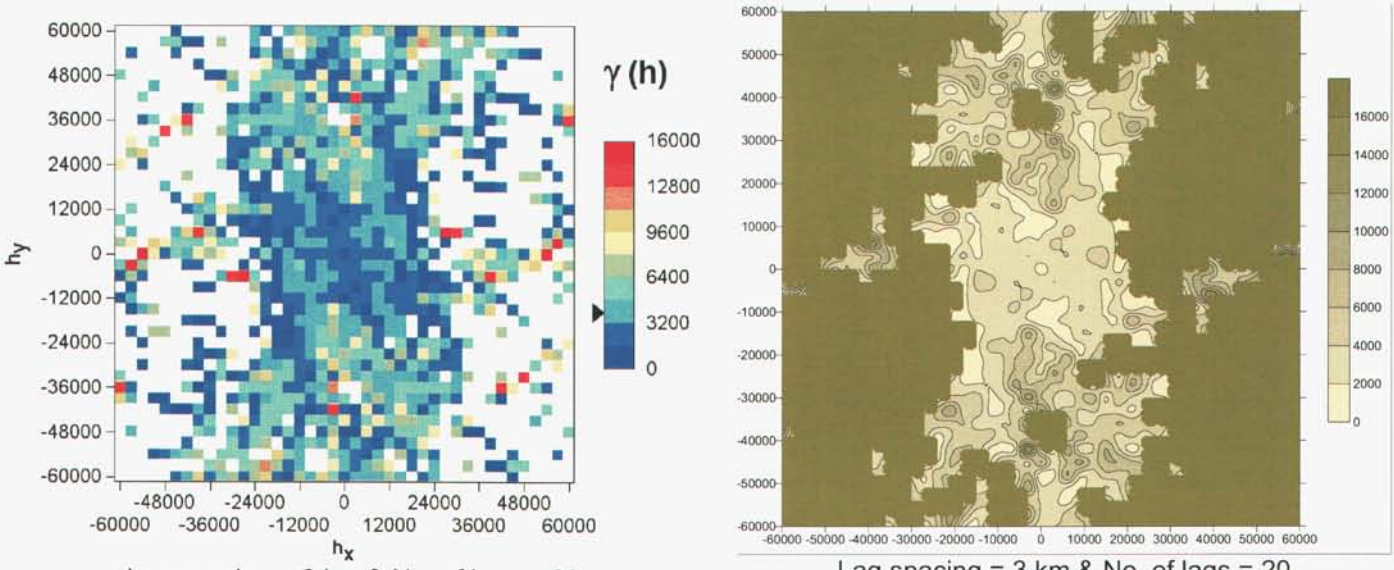
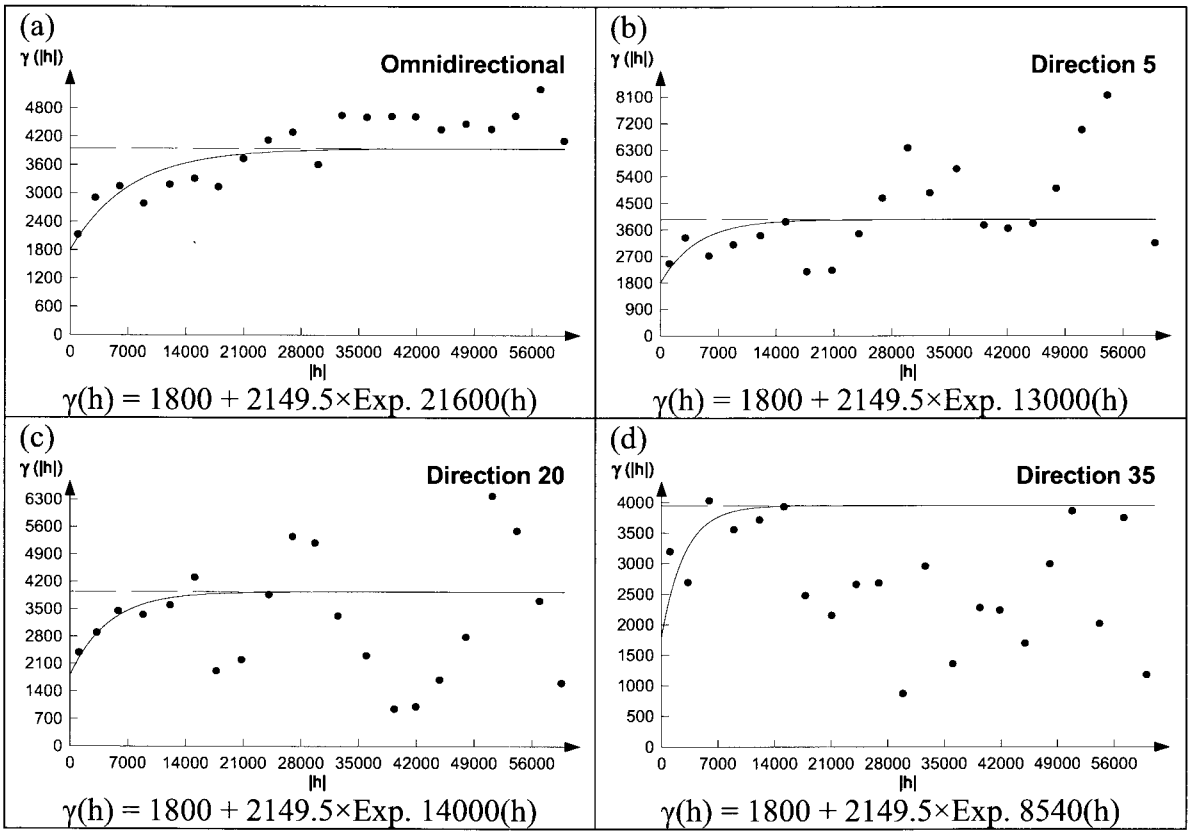
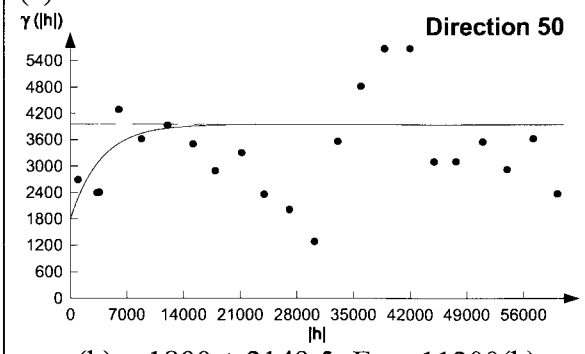


Fig. 4.15 Variogram (a) surface and (b) contour plots for geothermal gradient at 3 km lag spacing and number of lags = 20. Note: In Fig. 4.15a the black triangle shown on the color legend indicates the total variance value.

Variogram surface and contour plots show that the major axis direction for correlation distance is most probably around N 10 to 30° W. Fig. 4.16a shows the variogram (3 km lag spacing) calculated for geothermal gradient assuming isotropic spatial variation (omni directional) in the considered region. Directional variograms were calculated at every 15° anti-clockwise starting at 5° (East) using a lag spacing of 3 km (Figs. 4.16b through m). The same directional parameters were used as for the groundwater temperature in calculating the directional variograms. 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 the goodness of fit test results. The obtained variogram models for directional variograms are shown in Figs. 4.16b through m. Note that the nugget and partial sill values for the variograms were found to be 1800 and 2149.5, respectively. Correlation distances were found to vary with the direction.

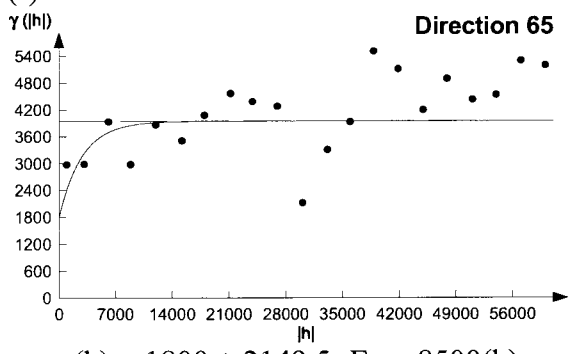


(e)



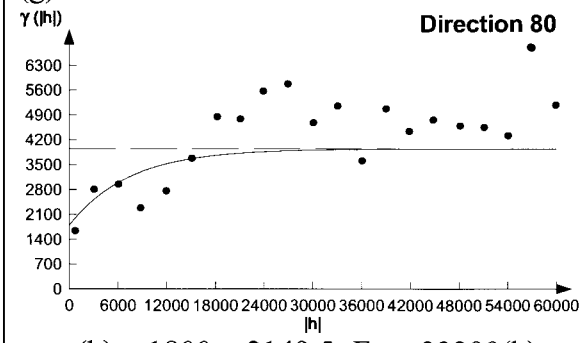
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 11200(h)$

(f)



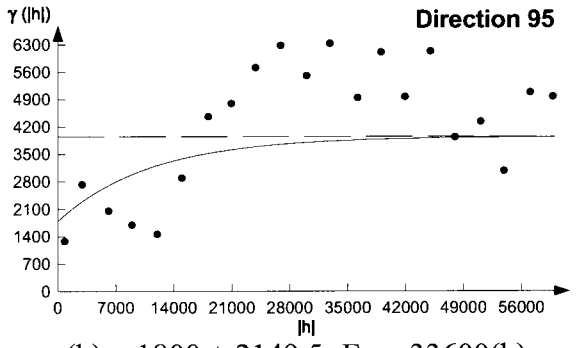
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 8500(h)$

(g)



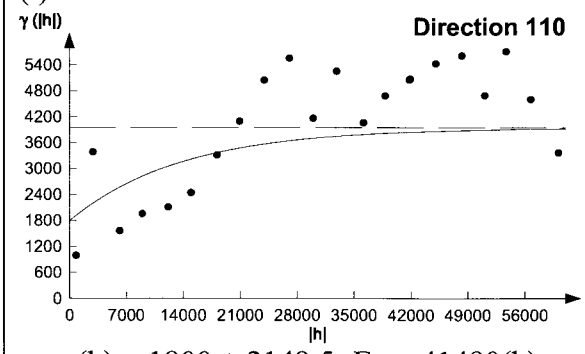
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 22200(h)$

(h)



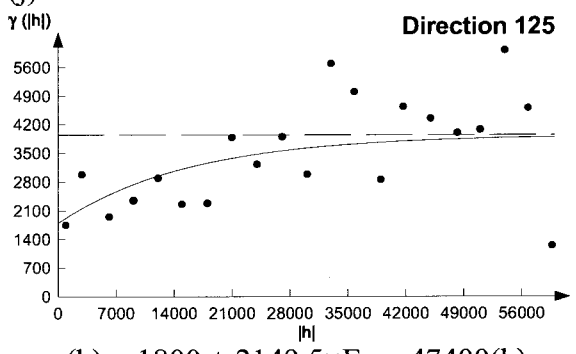
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 33600(h)$

(i)



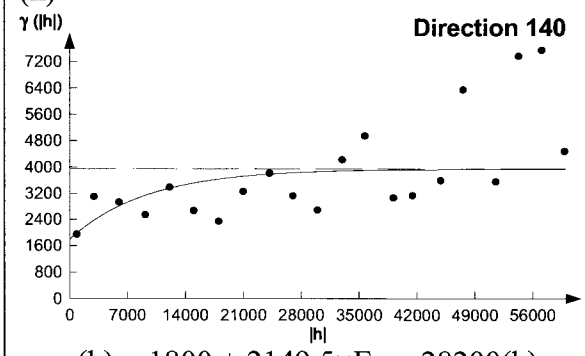
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 41480(h)$

(j)



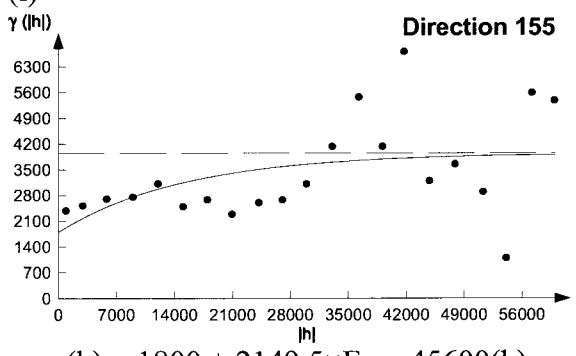
$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 47400(h)$

(k)



$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 28200(h)$

(l)



$\gamma(h) = 1800 + 2149.5 \times \text{Exp. } 45600(h)$

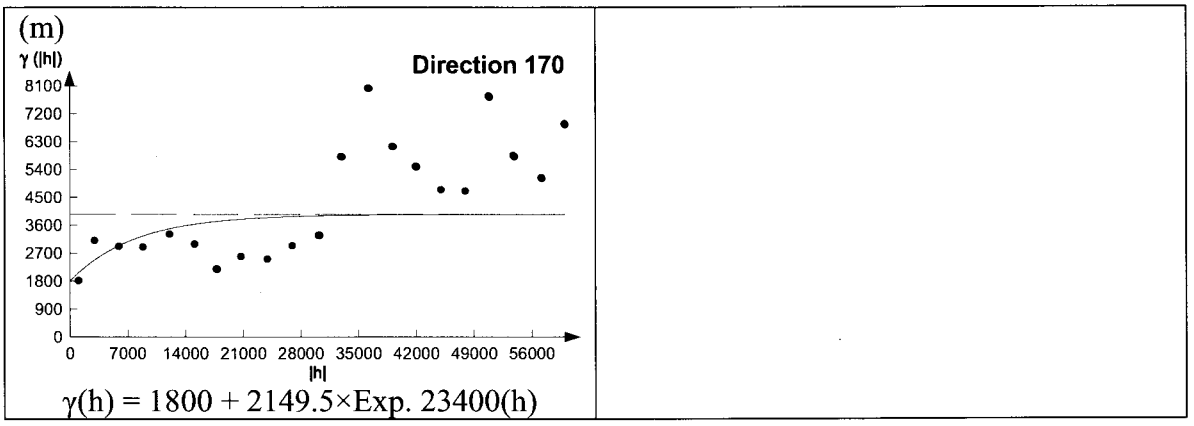


Fig. 4.16 Omnidirectional variogram (a) and directional variograms (b) through (m) for geothermal gradient at 3 km lag spacing.

The correlation distance obtained for every 15-degree is plotted on a polar coordinate system in Fig. 4.17. The least square elliptical fit obtained for the directional correlation distance is also shown in the same figure. Figure 4.17 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 geothermal gradient. Major axis direction, and semi-major and semi-minor axis lengths obtained for the correlation distance model are also given in the figure.

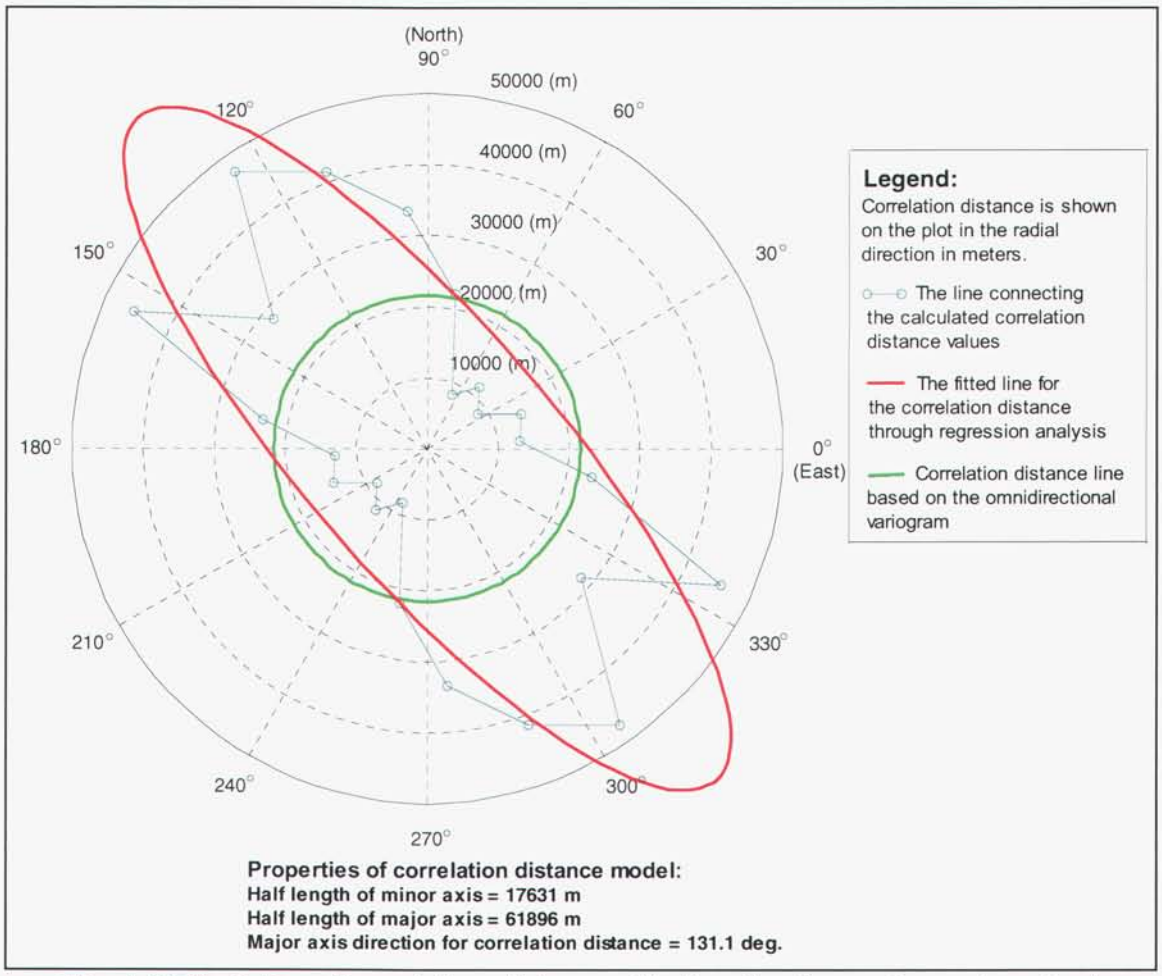


Fig. 4.17 Variation of correlation distance with direction for geothermal gradient

Properties of the aforementioned variogram model were used to perform kriging using ArcGIS Geostatistical Analyst (Johnston et al, 2001). Fig. 4.18 shows the map obtained for the predicted values. The map obtained for the standard error (the square root of the kriging variance) of the prediction is shown in Fig. 4.19. Note that the standard error (uncertainty) increases in areas with a lack of data.

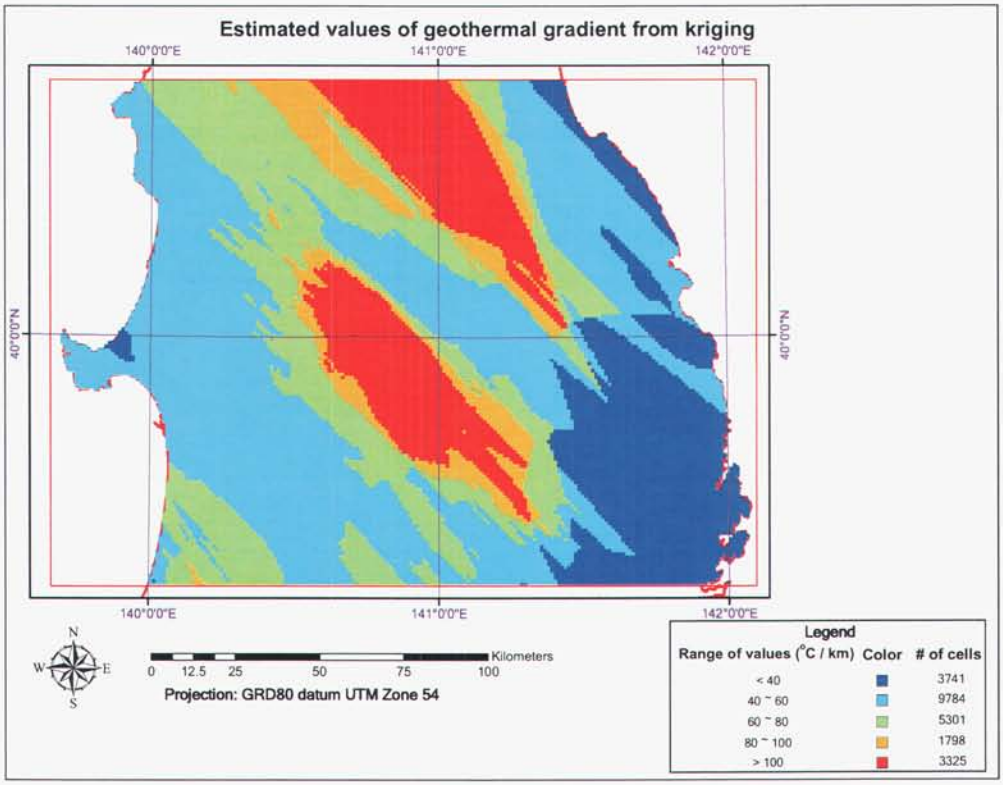


Fig. 4.18 Estimated values of geothermal gradient from kriging for the study area.

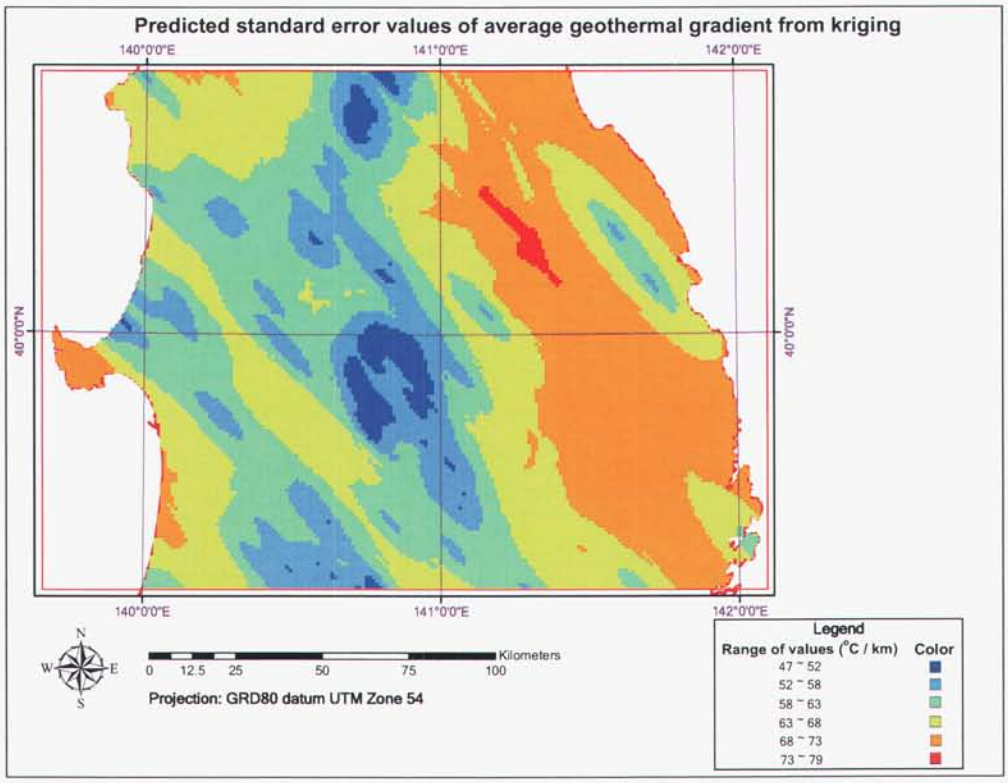


Fig. 4.19 Predicted standard error values of average geothermal gradient from kriging for the study area.

4.4 Heat Discharge

Figure 4.20 shows the extended database region used to perform variogram and kriging modeling for the heat discharge in the study area. In the extended data base region heat discharge data were available for 113 rectangular (almost) cells of size approximately 15 km x 12 km. These data were used to obtain the variogram surfaces shown in Figs. 4.21a and b. For the geostatistical analysis, four points in each rectangle have been specified. Note that those four points within a rectangle are assigned the same heat discharge value.

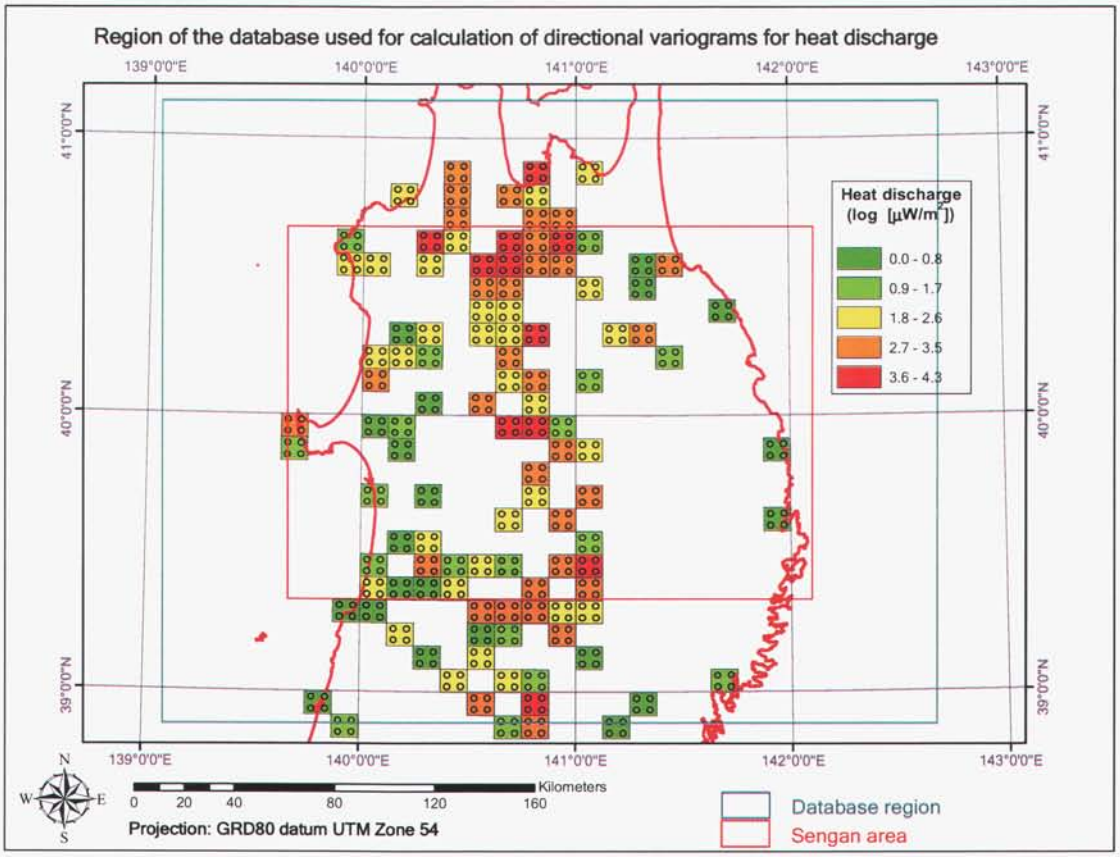
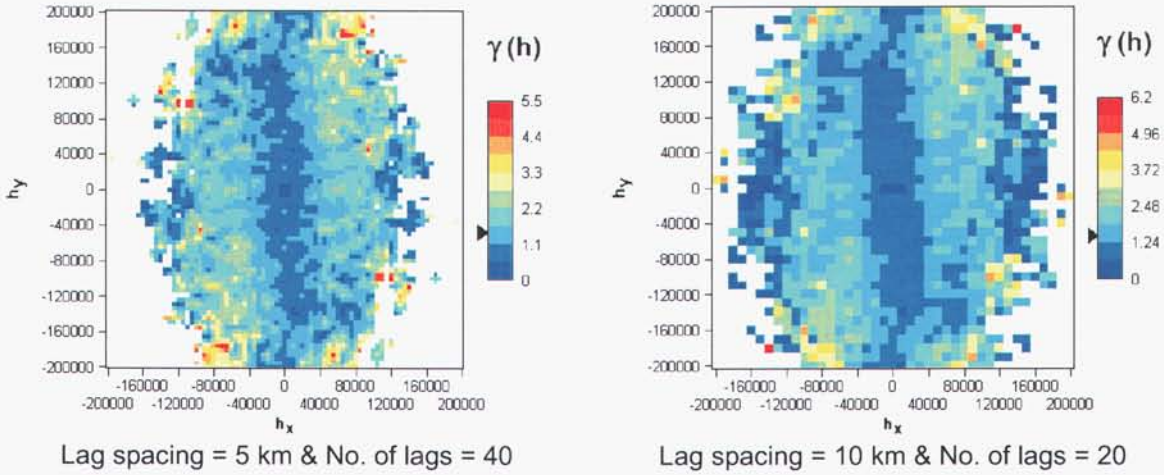


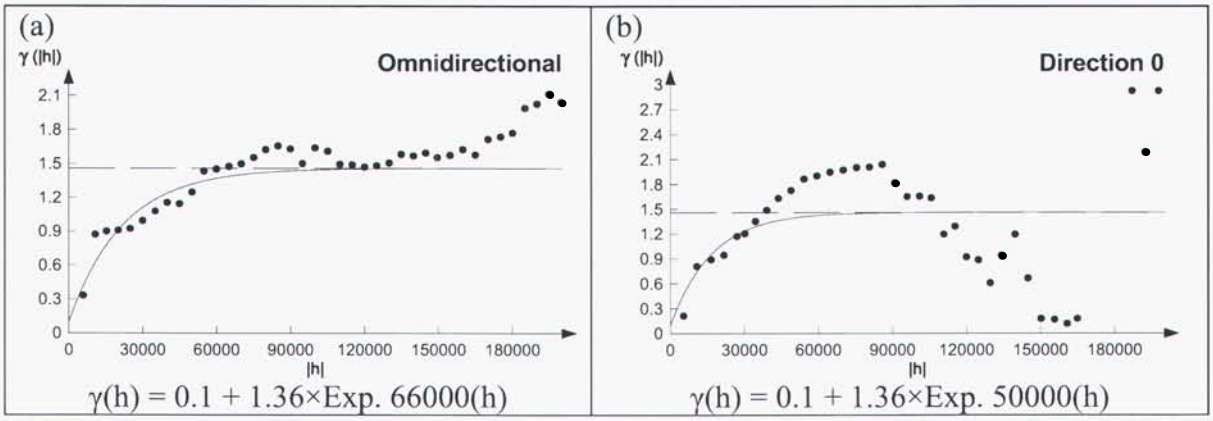
Fig. 4.20 Region of the database used for calculation of directional variograms for heat discharge



Lag spacing = 5 km & No. of lags = 40 Lag spacing = 10 km & No. of lags = 20

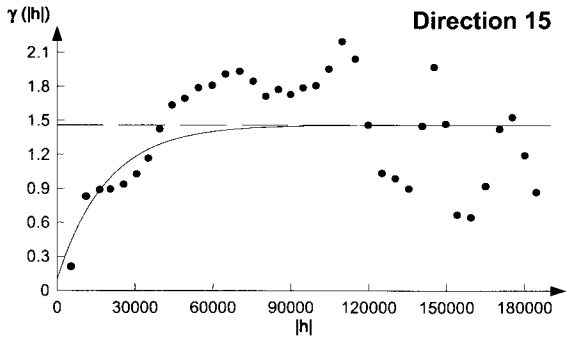
Fig. 4.21 Variogram surface plots of heat discharge at different lag spacings and number of lags. Note: the black triangle shown on the color legend indicates the total variance value.

Variogram surface plots show that the major axis direction for correlation distance is most probably around North. Fig. 4.22a shows the variogram (5 km lag spacing) calculated for heat discharge assuming isotropic spatial variation (omni directional) in the considered region. Directional variograms were calculated at every 15° anti-clockwise starting at 0° (East) using a lag spacing of 5 km (Figs. 4.22b through m). The same directional parameters were used as for the groundwater temperature in calculating the directional variograms. 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 the goodness of fit test results. The obtained variogram models for directional variograms are shown in Figs. 4.22b through m. Note that the nugget and partial sill values for the variograms were found to be 0.1 and 1.36, respectively. Correlation distances were found to vary with the direction. The relatively small nugget value in the variograms is due to the assignment of four equal values of heat discharge within each of the 15×12 km cells.



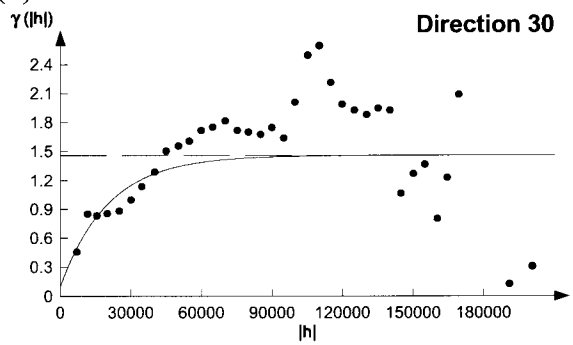
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 66000(h)$ $\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 50000(h)$

(c)



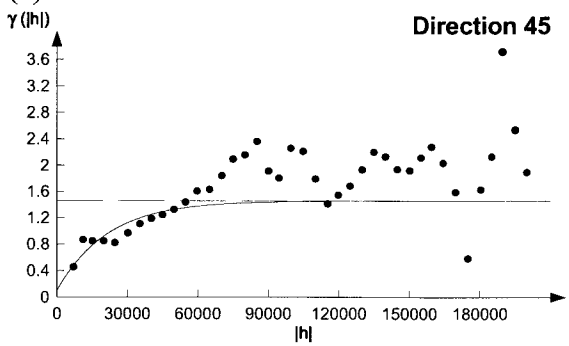
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 57000(h)$

(d)



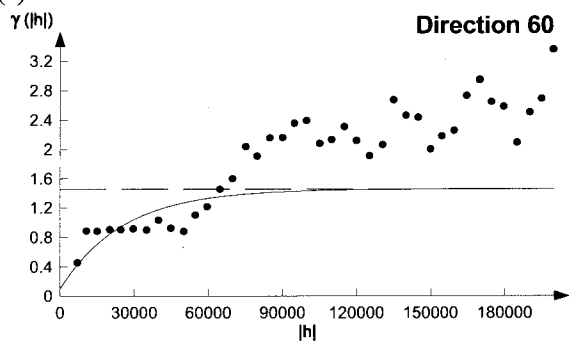
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 61000(h)$

(e)



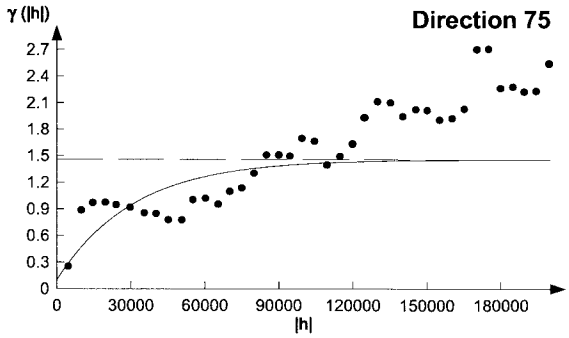
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 64000(h)$

(f)



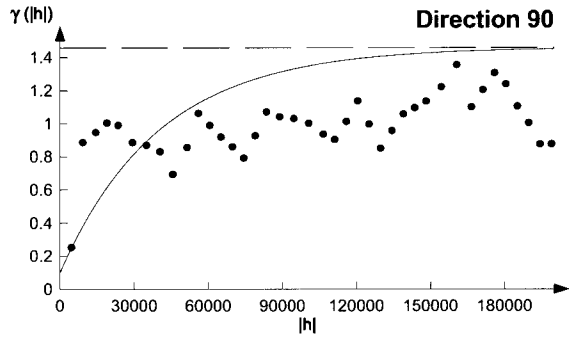
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 76000(h)$

(g)



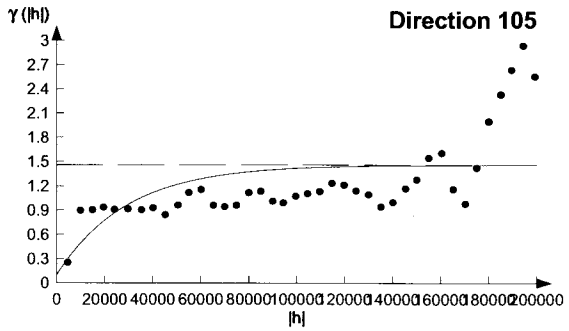
$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 93000(h)$

(h)

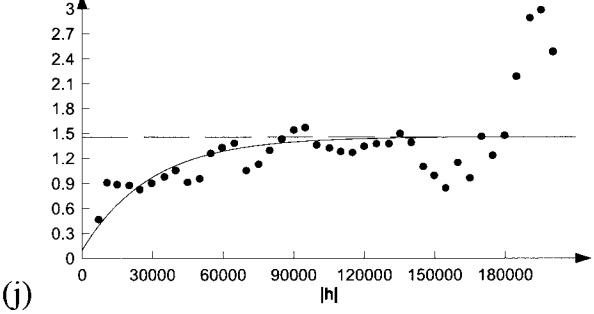


$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 120000(h)$

(i)



$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 102000(h)$



(j)

$\gamma(h) = 0.1 + 1.36 \times \text{Exp. } 87000(h)$

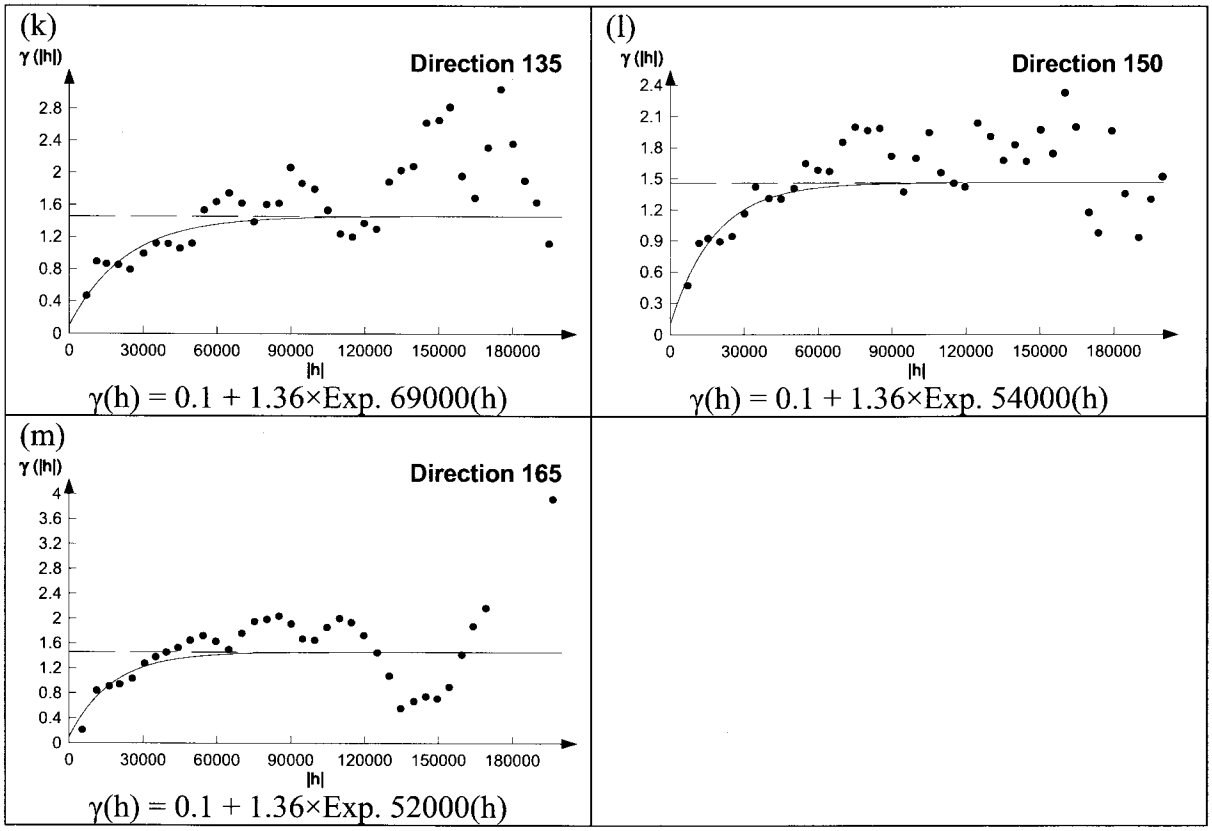


Fig. 4.22 (a) Omnidirectional variogram and (b) through (m) directional variograms for heat discharge at 5 km lag spacing

The correlation distance obtained for every 15-degree is plotted on a polar coordinate system in Fig. 4.23. The least square elliptical fit obtained for the directional correlation distance is also shown in the same figure. Figure 4.23 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 heat discharge. The nearly north-south major axis direction and semi-major and semi-minor axis lengths obtained for the correlation distance model are also given in the figure.

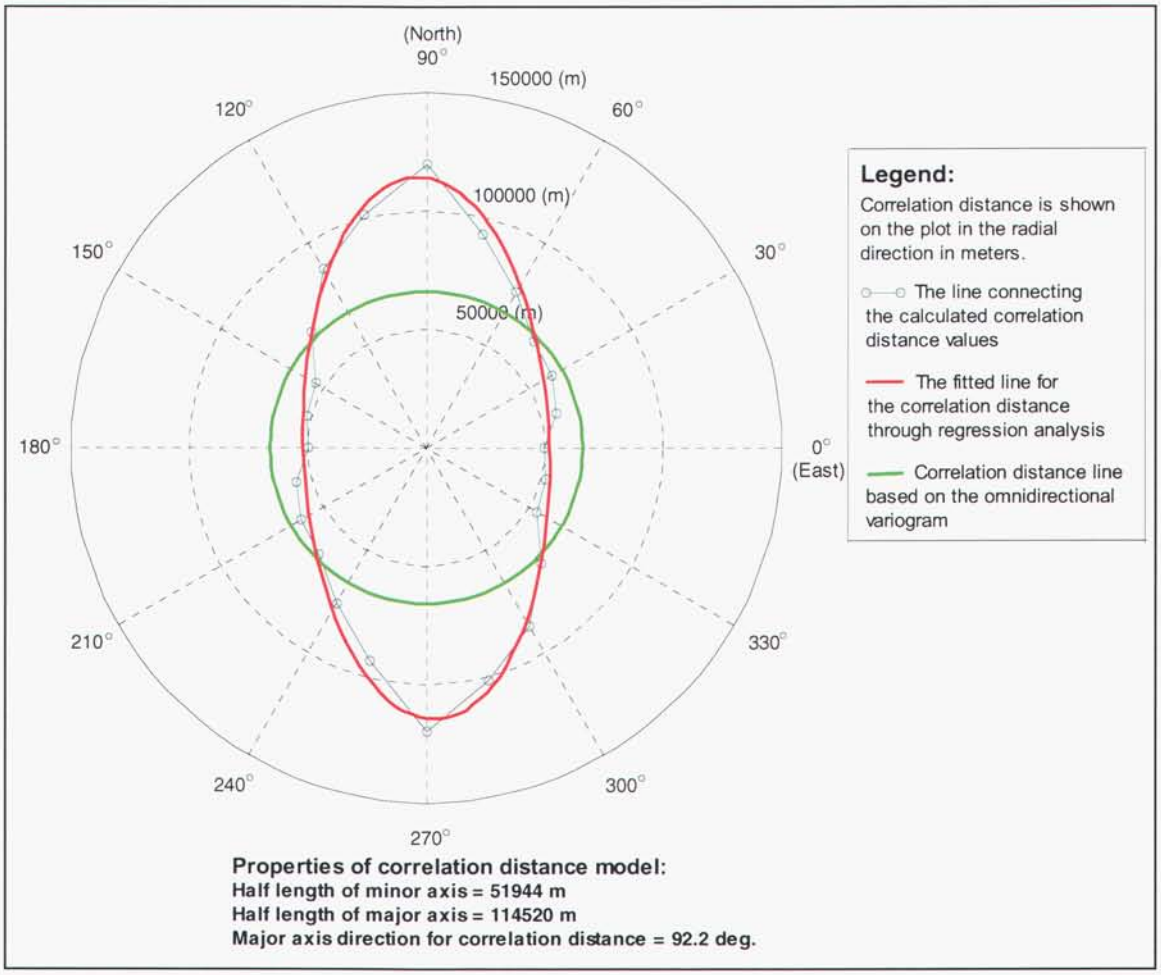


Fig. 4.23 Variation of correlation distance with direction for heat discharge

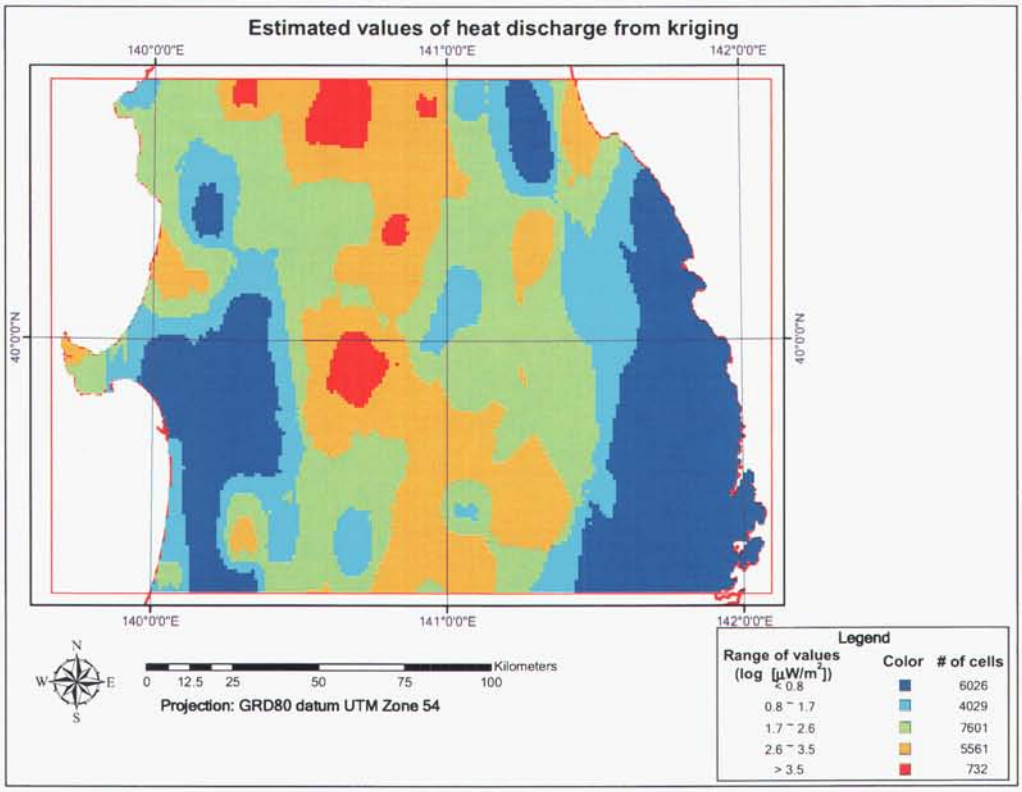


Fig. 4.24 Estimated values of heat discharge from kriging for the study area.

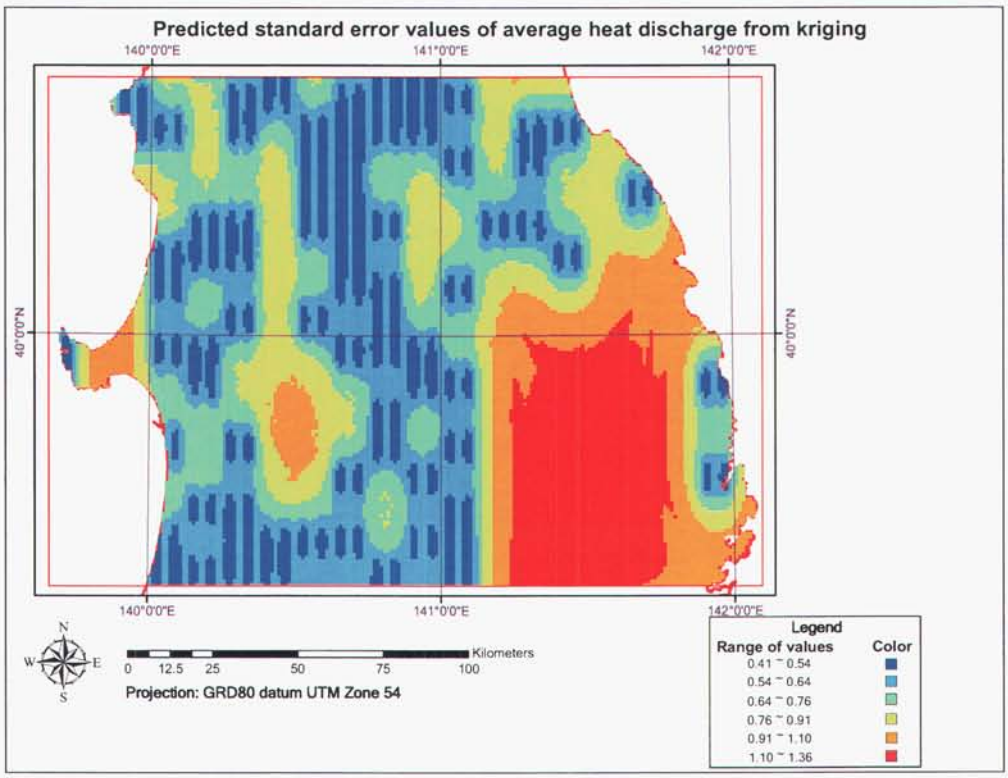


Fig. 4.25 Predicted standard error values of average heat discharge from kriging for the study area.

Properties of the aforementioned variogram model were used to perform kriging using ArcGIS Geostatistical Analyst (Johnston et al, 2001). Figure 4.24 shows the map obtained for the predicted values. The map obtained for the standard error (the square root of the variance) of the prediction is shown in Fig. 4.25. Note that the standard error (uncertainty) increases in areas with a lack of data.

4.5 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 3.6. For each 1 km cell that was fully within the volcanic rock area, a value of 1 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 Fig. 4.26.

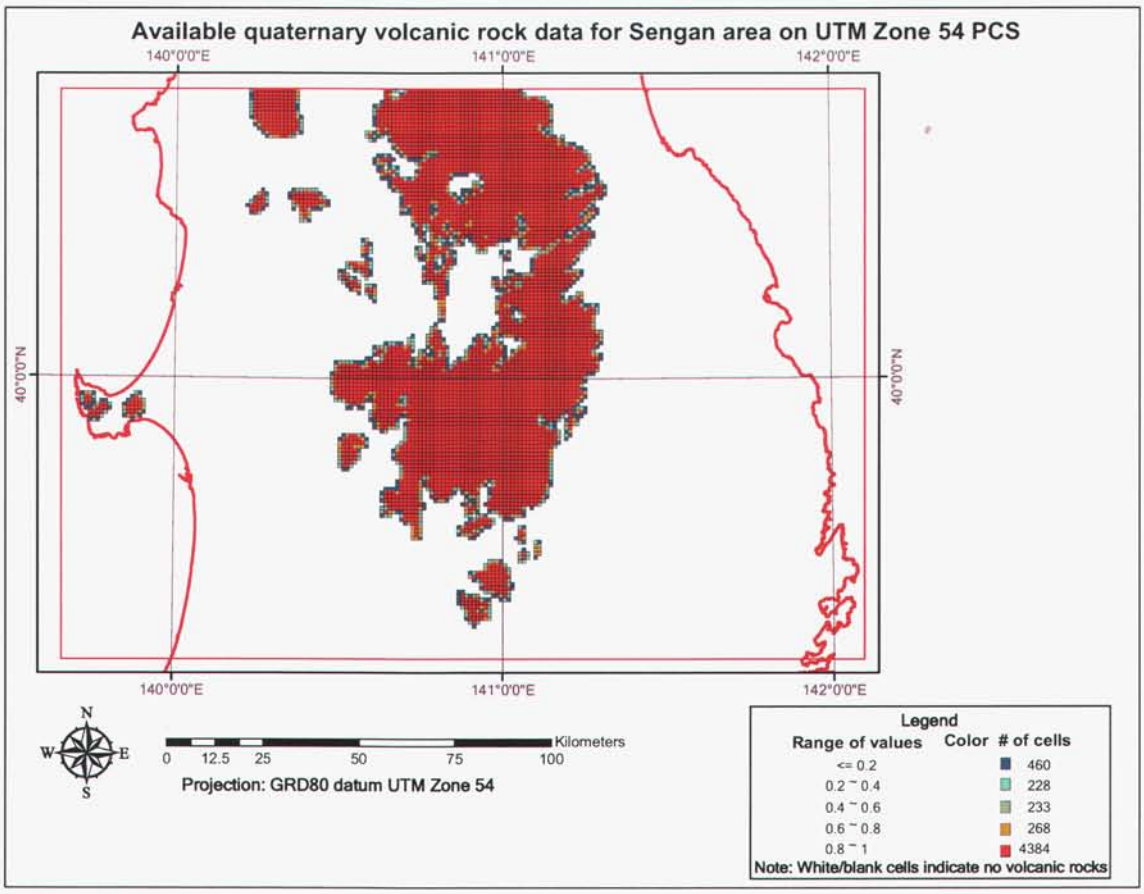


Fig. 4.26 Available quaternary volcanic rock data for Sengan region on 1 km square grid using UTM zone 54 PCS

4.6 Presence of Hydrothermal Alteration

To express the presence of hydrothermal alteration in 1 km square grid system, the 1 km grid mesh was superimposed on the combined map shown in Section 3.7. For each 1 km cell that was fully within the hydrothermal alteration area, a value of 1 was assigned to the cell. For each 1 km cell that was fully outside the hydrothermal alteration area, a value of 0 was assigned to the cell. If only a portion of 1 km cell fell within the hydrothermal alteration area, the cell was assigned a value between 0 and 1 according to the proportional coverage of the cell area by the hydrothermal alteration area. This final map obtained is shown in Fig. 4.27.

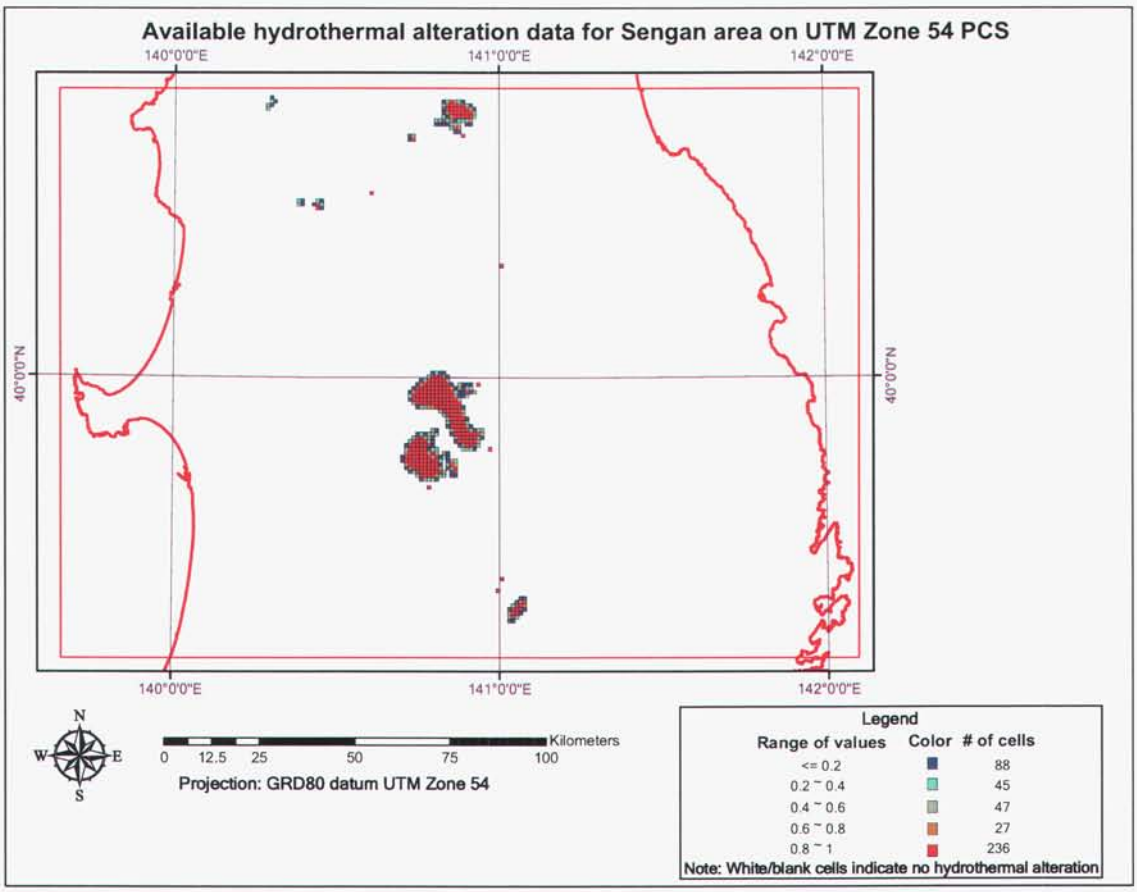


Fig. 4.27 Available hydrothermal alteration data for Sengan region on 1 km square grid using UTM zone 54 PCS

MULTIVARIATE CLASSIFICATION (TYPIFICATION) OF VOLCANISM FOR SENGAN REGION

5.1 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 methods, 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 x 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. One measure of resemblance used is the correlation coefficient, r_{ij}, in which the roles of objects and variables are interchanged. This can be done by transposing **X** so that rows become columns and vice versa, then calculating r_{ij}, in the conventional manner. Although called "correlation," this measure is not really a correlation coefficient in the conventional sense because it involves "means" and "variances" calculated across all the variables measured on any two objects, rather than the means and variances of two variables.

Another popular similarity measure between objects is a standardized m-space Euclidean distance, d_{ij}. This is computed by

$$d_{ij} = \sqrt{\frac{\Sigma_{k=1}^{m}(X_{ik} - X_{jk})^2}{m}} \qquad (5.1)$$

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

Computation of a similarity measurement using the correlation coefficient, Euclidean distance or some other technique 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. Different procedures are available in the literature to form these groups or clusters (Sneath and Sokal, 1973; Backer 1995; Gordon, 1999).

5.2 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 2**). In **Section 4**, at the center of each cell, values were calculated for each of the following geological variables: groundwater temperature, geothermal gradient, heat discharge, groundwater pH value, presence of volcanic rock and presence of hydrothermal alteration. These geological variables were identified as the variables significantly correlated to the volcanic activity. This means that to perform cluster analysis 23949 grid point locations (ie cases or objects) are available. For each grid point location, values for 6 geologic variables are available. This forms a complete variable vector for 23949 cases. 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 using the STATISTICA software package was performed. Figure 5.1 shows the dendrogram obtained for the final 50 groups. Figure 5.1 also shows that it is possible to reduce 23949 cases to 5 groups by selecting a value about 2.65 for the Euclidean distance. By selecting a value slightly lower than 3 for the Euclidean distance, the number of groups can be reduced to 3 (Figure 5.1). The mean values obtained for geologic variable vector for different grouping situations are shown in Figures 5.2 and 5.3. Descriptive statistics obtained for different clusters for different grouping situations are shown in Tables 5.1 and 5.2. Note that in each Table and Figure, the cluster numbers are arranged such that the volcanism level moves from a lowest possible volcanic disruption (Cluster 1) to a highest possible volcanic disruption (Cluster 5). Figures 5.4-5.7, and Figures 5.8-5.11 show the histograms obtained for different geologic variables for the 5-group and 3-group situations, respectively. Relative frequency distributions of groundwater temperature, geothermal gradient and heat discharge are expected to move to the right (higher values) with the increasing cluster number. Apart from the relative frequency distributions of clusters 3 and 4 of groundwater temperature, the relative frequency plots given in Figures 5.4, 5.6-5.8, 5.10-5.11 show the expected behavior. Relative frequency distributions of pH are expected to move to the left (lower values) with the increasing cluster number. Even though negligible differences are seen between some adjacent clusters, in overall, the relative frequency plots given in Figures 5.5 and 5.9 show the expected behavior.

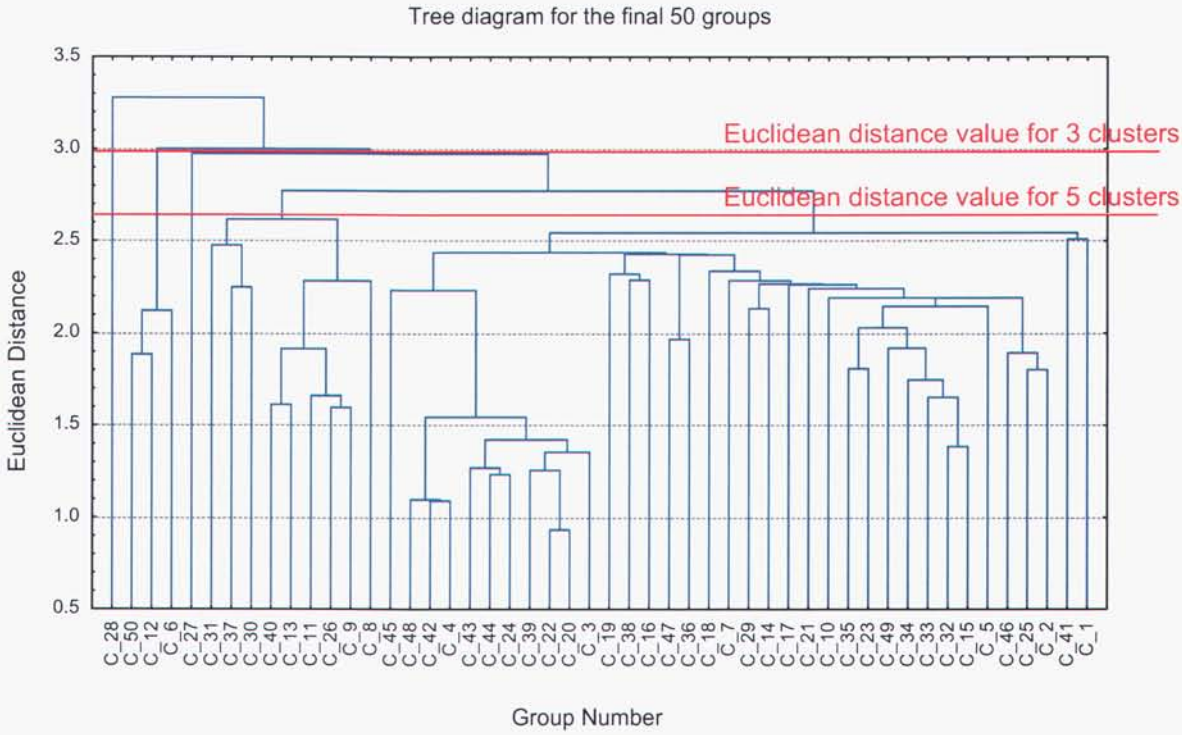


Fig. 5.1 Dendrogram obtained for the final 50 groups

Table 5.1 Cluster analysis results for 5-group clustering

Descriptive Statistics for Cluster 1, Cluster contains 4254 cases			
Variable	Mean	Standard Deviation.	Variance
Ground water Temperature	11.487	3.993	15.943
Geothermal Gradient	30.266	7.303	53.332
pH Value	7.481	0.521	0.271
Presence of Volcanic Rock	0.001	0.027	0.001
Presence of Hydrothermal Alteration	0.000	0.000	0.000
Heat Discharge	0.698	0.577	0.332
Descriptive Statistics for Cluster 2, Cluster contains 8952 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	23.634	7.193	51.744
Geothermal Gradient	54.689	8.925	79.657
pH Value	7.474	0.650	0.422
Presence of Volcanic Rock	0.079	0.260	0.067
Presence of Hydrothermal Alteration	0.000	0.013	0.000
Heat Discharge	1.441	0.817	0.668
Descriptive Statistics for Cluster 3, Cluster contains 6376 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	45.356	9.486	89.978
Geothermal Gradient	63.730	10.780	116.202
pH Value	7.687	0.662	0.439
Presence of Volcanic Rock	0.171	0.357	0.127
Presence of Hydrothermal Alteration	0.006	0.073	0.005
Heat Discharge	2.381	0.755	0.571
Descriptive Statistics for Cluster 4, Cluster contains 3192 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	36.167	12.153	147.702
Geothermal Gradient	108.608	14.827	219.830
pH Value	7.458	0.850	0.722
Presence of Volcanic Rock	0.524	0.480	0.231
Presence of Hydrothermal Alteration	0.011	0.104	0.011
Heat Discharge	2.696	0.614	0.377
Descriptive Statistics for Cluster 5, Cluster contains 1175 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	75.922	18.935	358.519
Geothermal Gradient	145.962	24.908	620.433
pH Value	5.613	1.676	2.807
Presence of Volcanic Rock	0.841	0.333	0.111
Presence of Hydrothermal Alteration	0.175	0.349	0.122
Heat Discharge	3.028	0.431	0.186

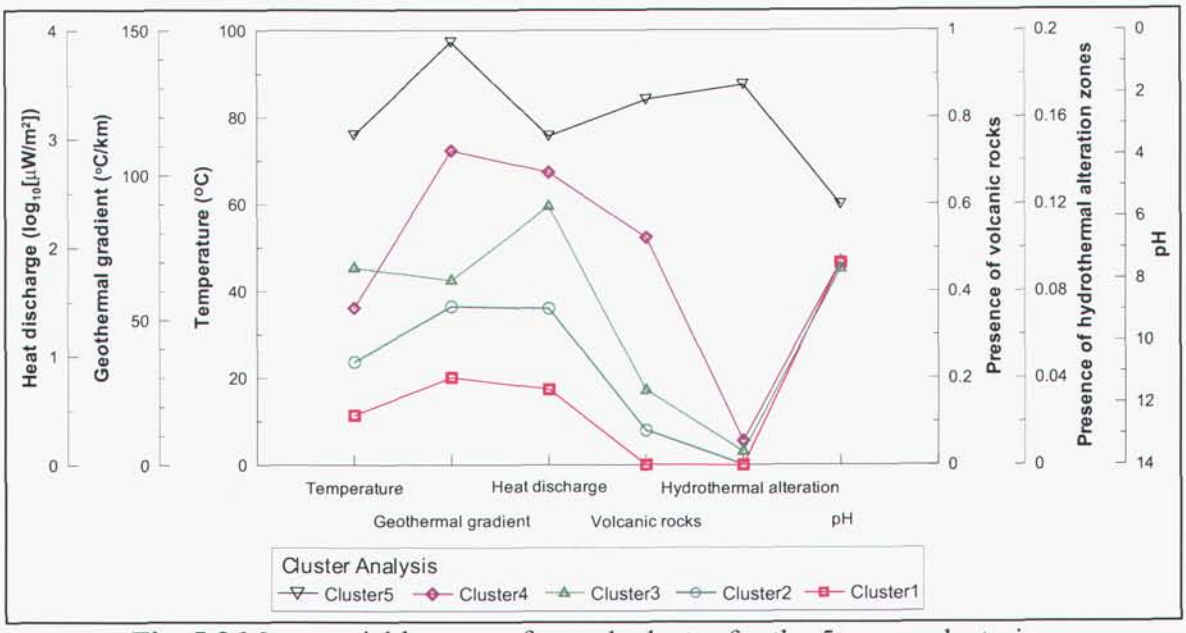


Fig. 5.2 Mean variable vector for each cluster for the 5-group clustering

Table 5.2 Cluster analysis results for 3-group clustering

Descriptive Statistics for Cluster 1, Cluster contains 8746 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	15.757	6.464	41.785
Geothermal Gradient	40.151	11.604	134.651
pH Value	7.442	0.588	0.346
Presence of Volcanic Rock	0.028	0.162	0.026
Presence of Hydrothermal Alteration	0.000	0.000	0.000
Heat Discharge	0.975	0.757	0.573
Descriptive Statistics for Cluster 2,Cluster contains 11675 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	37.636	12.034	144.806
Geothermal Gradient	64.178	12.201	148.860
pH Value	7.629	0.656	0.430
Presence of Volcanic Rock	0.155	0.344	0.118
Presence of Hydrothermal Alteration	0.005	0.063	0.004
Heat Discharge	2.115	0.843	0.710
Descriptive Statistics for Cluster 3,Cluster contains 3528 cases			
Variable	Mean	Standard Deviation	Variance
Ground water Temperature	50.188	23.642	558.956
Geothermal Gradient	125.401	22.479	505.316
pH Value	6.798	1.481	2.194
Presence of Volcanic Rock	0.682	0.443	0.196
Presence of Hydrothermal Alteration	0.066	0.231	0.054
Heat Discharge	2.832	0.556	0.309

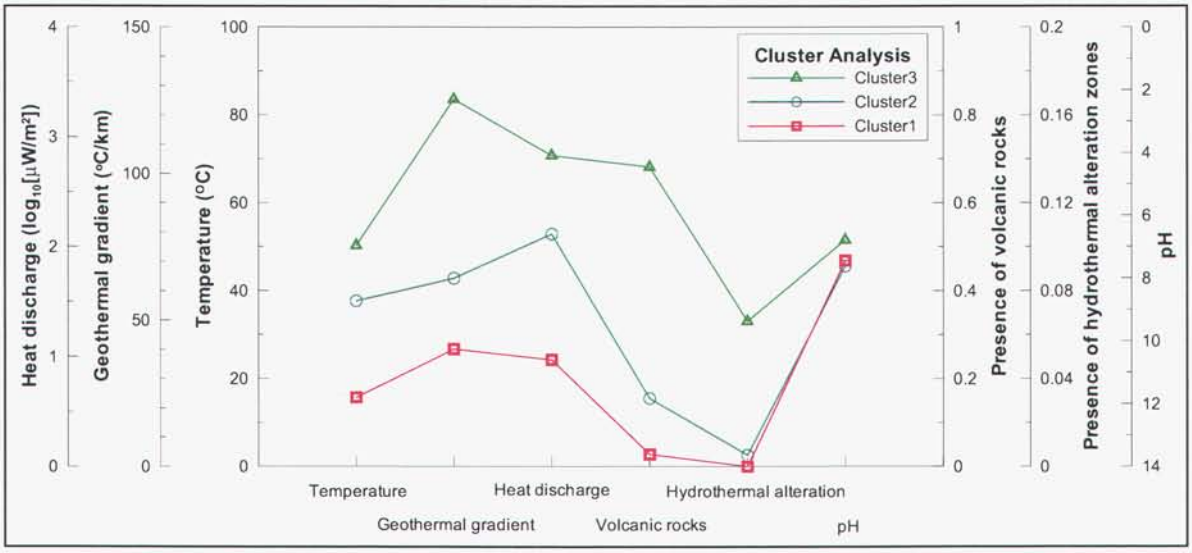


Fig. 5.3 Mean variable vector for each cluster for the 3-group clustering

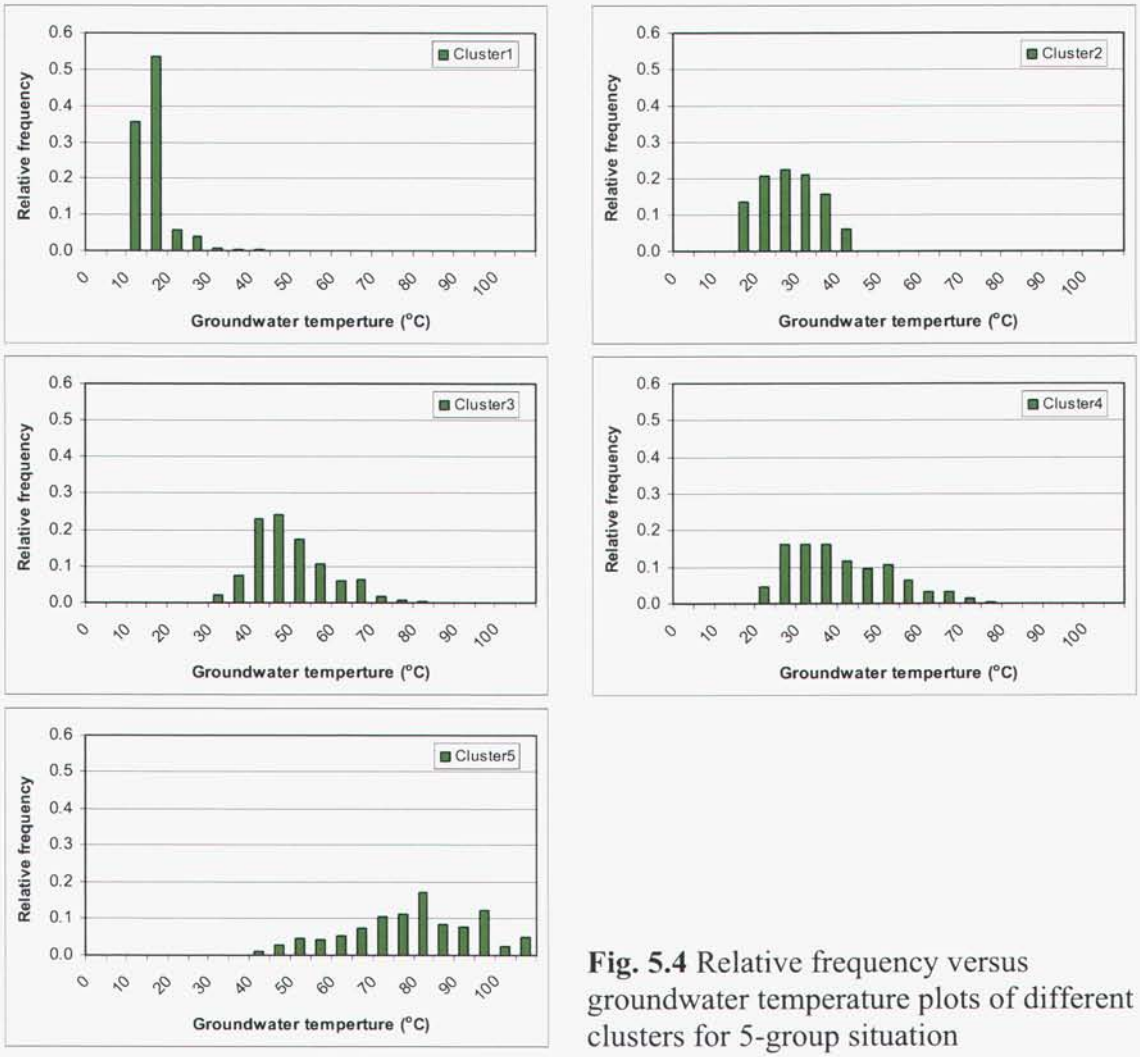


Fig. 5.4 Relative frequency versus groundwater temperature plots of different clusters for 5-group situation

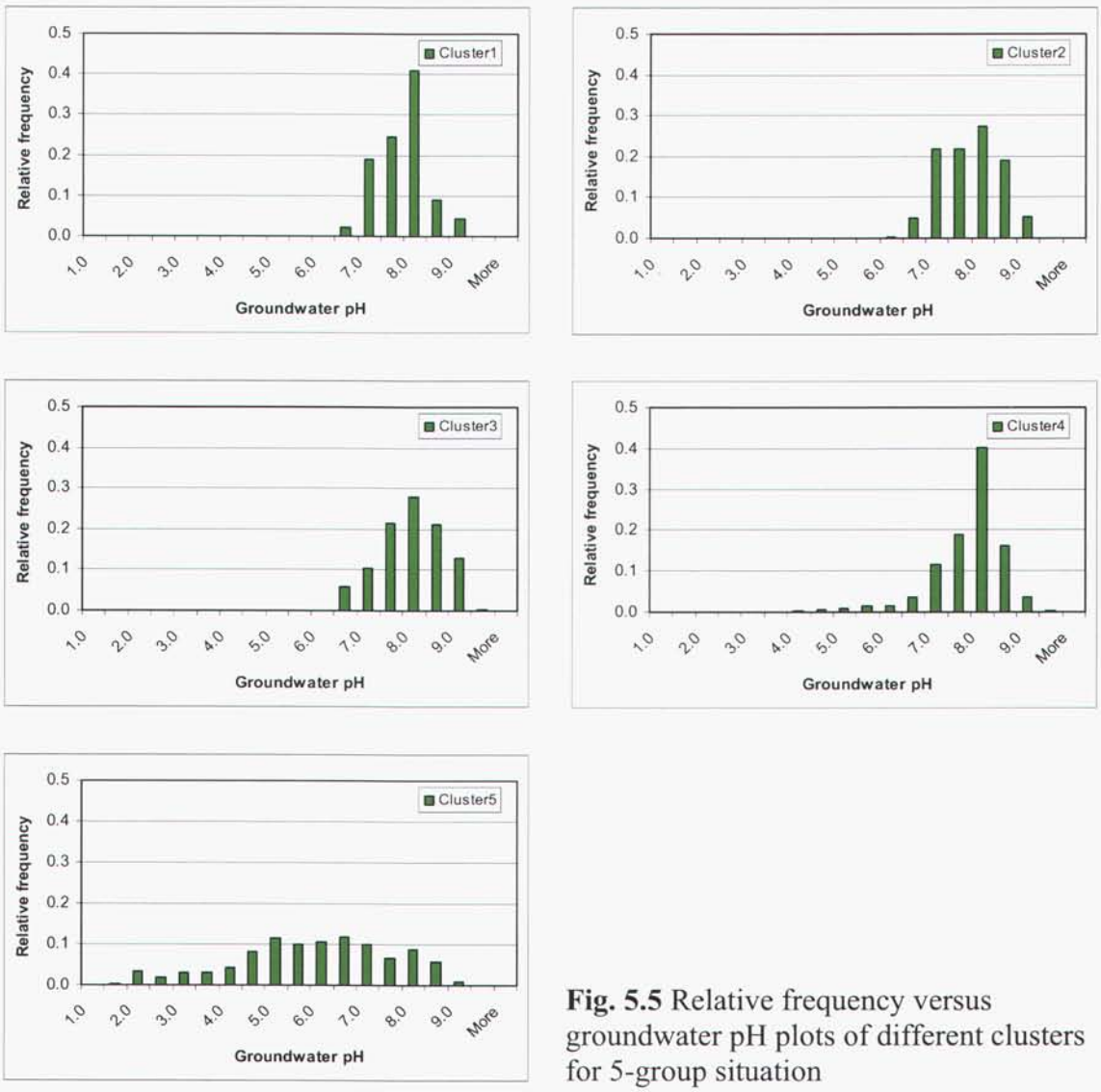


Fig. 5.5 Relative frequency versus groundwater pH plots of different clusters for 5-group situation

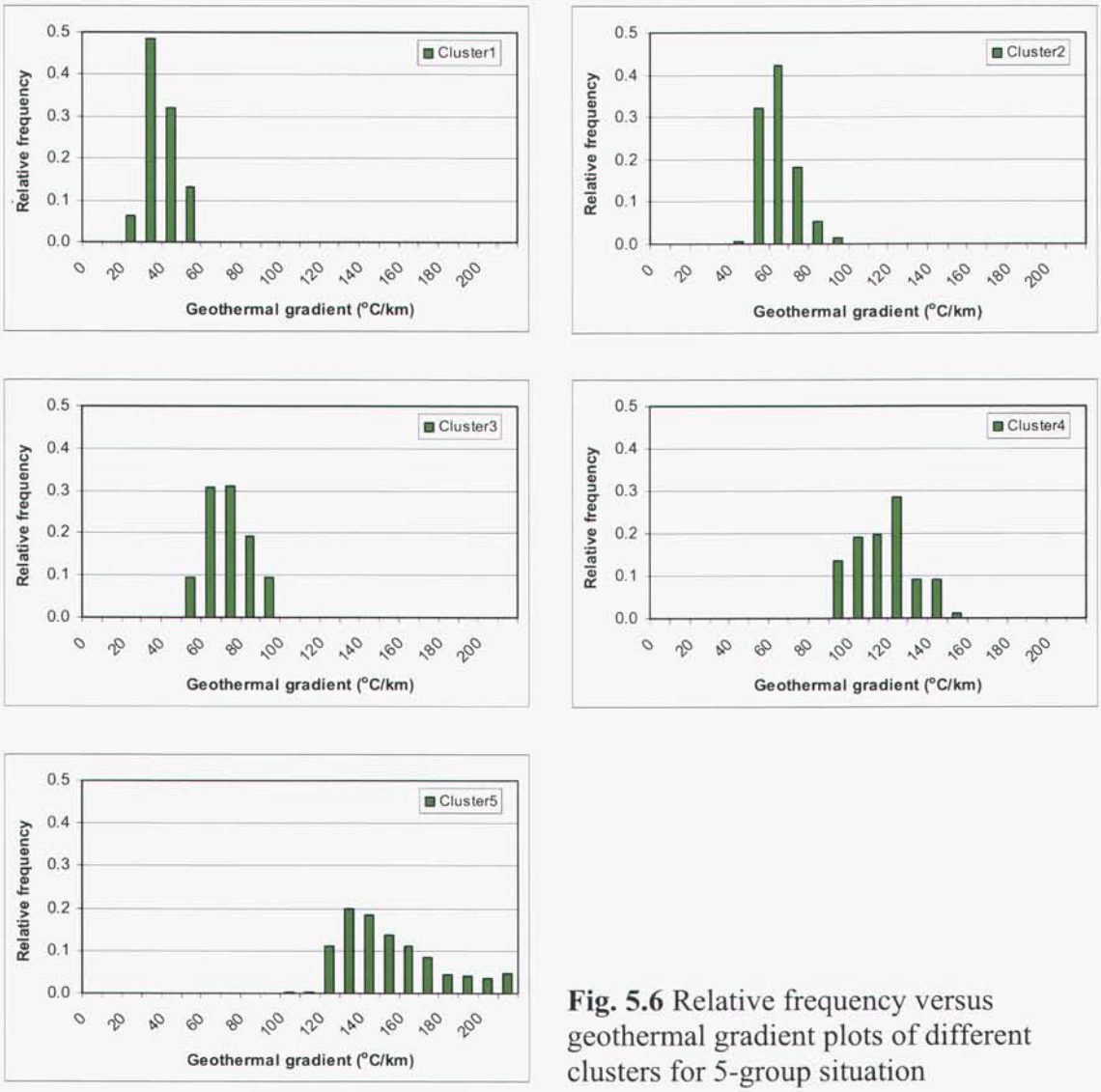


Fig. 5.6 Relative frequency versus geothermal gradient plots of different clusters for 5-group situation

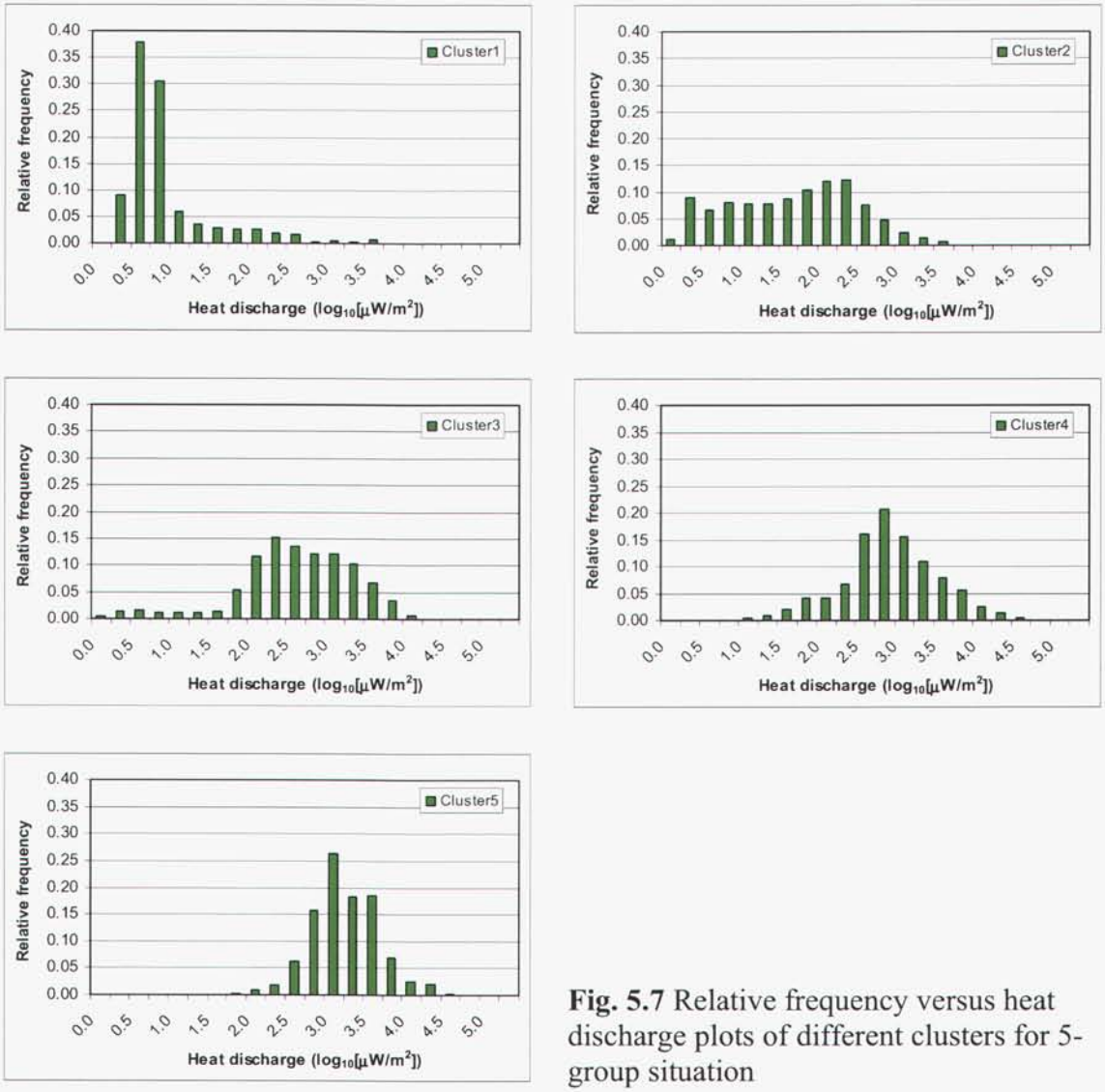


Fig. 5.7 Relative frequency versus heat discharge plots of different clusters for 5-group situation

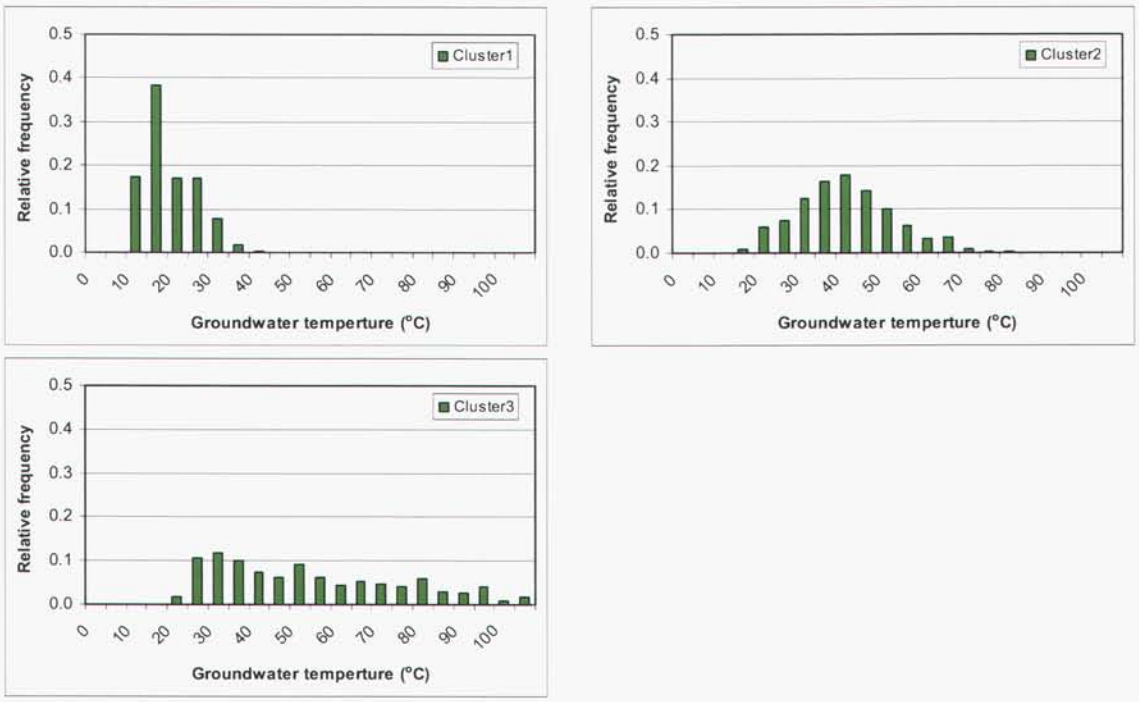


Fig. 5.8 Relative frequency versus groundwater temperature plots of different clusters for 3-group situation

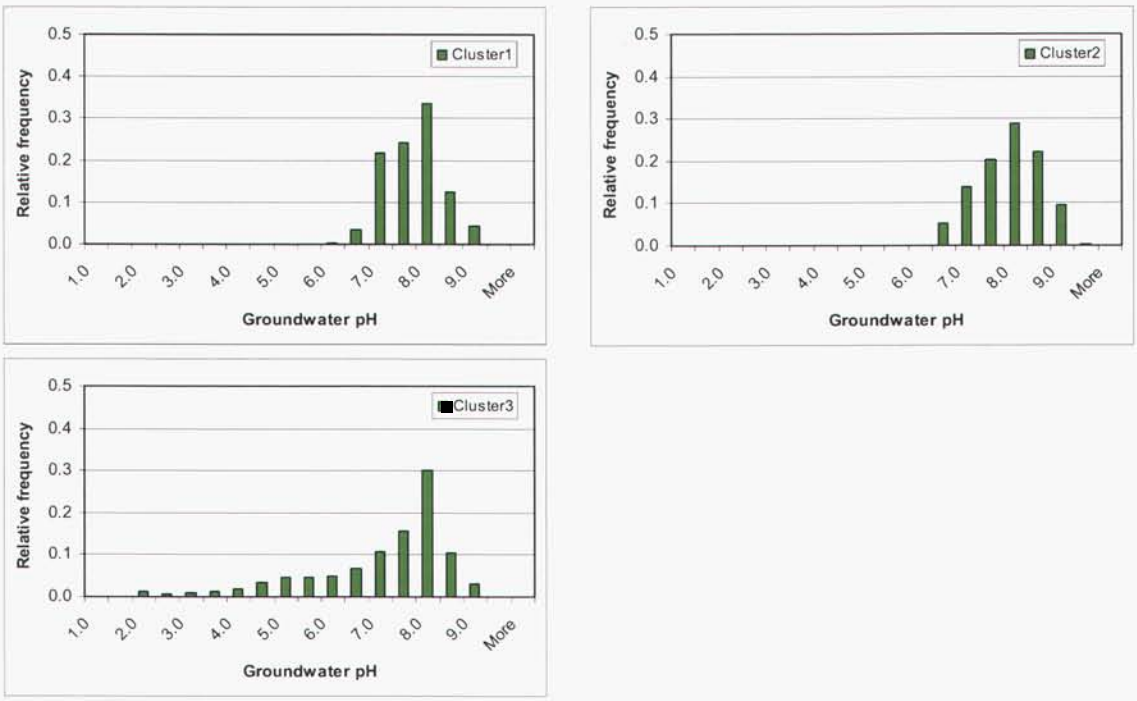


Fig. 5.9 Relative frequency versus groundwater pH plots of different clusters for 3-group situation

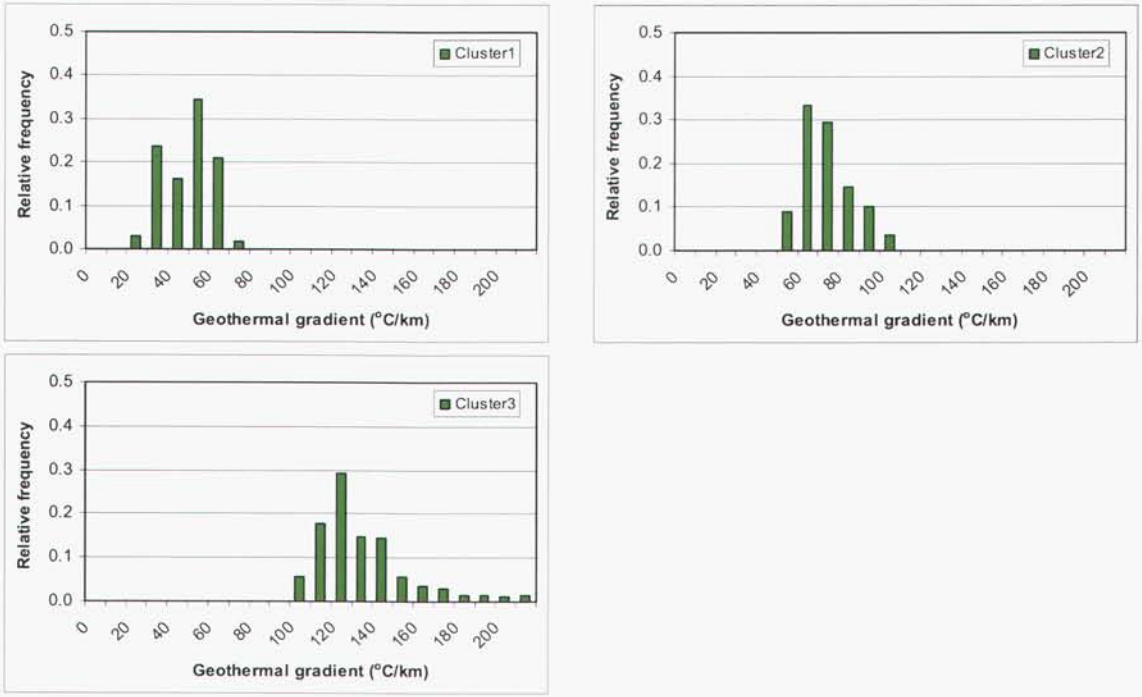


Fig. 5.10 Relative frequency versus geothermal gradient plots of different clusters for 3-group situation

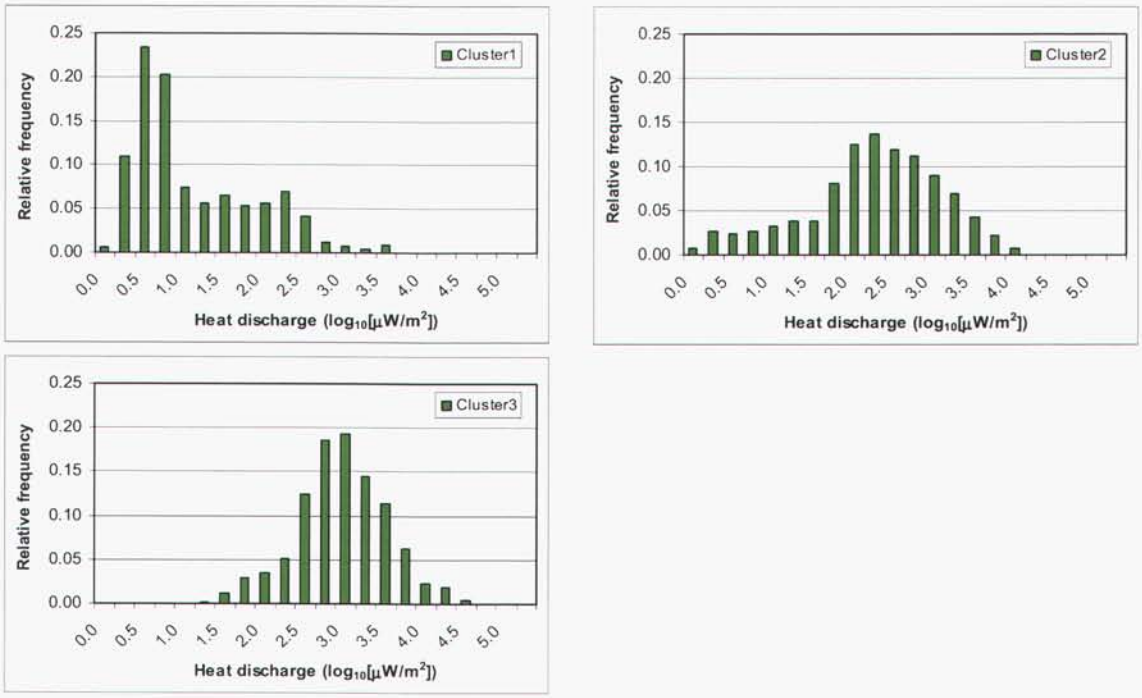


Fig. 5.11 Relative frequency versus heat discharge plots of different clusters for 3-group situation

5.3 Fundamentals of Discriminant Function Analysis (DFA)

The fundamentals of DFA are well covered in the book by Davis (2002). DA provides a rule to allocate samples or data points to certain predefined groups based on observed values of a set of variables. DFA calculates a function, which is a linear or non-linear combination of several variables, to differentiate or separate the membership of an object between the different groups. The different groups are known or defined before performing DFA. Note that the definition of the different groups, or classes, can be accomplished using CA as discussed in **Section 5.1**. Explanation of the general features of DFA is much easier for the case of two groups compared to several groups. Therefore, from here onwards only two groups are considered in explaining DFA.

DFA creates a discriminant function that transforms an original set of measurements on a group or an object into a single discriminant score (for a two group case). That score, represents the group's position along a line defined by the discriminant function. Therefore, the discriminant function can be thought of as a way of collapsing a multivariate problem into a univariate problem. DFA determines a transformation that gives the maximum ratio of the difference between two groups of multivariate means to the multivariate variance within the two groups. Considering the two groups to form clusters of points in multivariate space, DFA looks for the one orientation within that space along which the two clusters have the greatest separation while each cluster simultaneously has the least scatter. The dimensionality of the multivariate space determines whether a line, plane or hyperplane is used to discriminate the different groups.

Regression analysis can be used to determine the discriminant function. However, the dependent variable in this case is the differences between the multivariate means of the two groups, **D**. In matrix notation, the equation can be written as

$$\mathbf{D} = \mathbf{S}\boldsymbol{\lambda} \tag{5.2}$$

where **S** is an m x m matrix of pooled variances and covariances of the m variables and $\boldsymbol{\lambda}$ is the column vector of the unknown coefficients of the discriminant equation. The unknown $\boldsymbol{\lambda}$ can be determined as:

$$\boldsymbol{\lambda} = \mathbf{S}^{-1}\,\mathbf{D}. \tag{5.3}$$

For the two variable case, the discriminant function takes the form

$$R_i = \sum_{j=1}^{m} \lambda_j X_{ji} \tag{5.4}$$

where R_i is the discriminant score for a sample in the *ith* group and X_{ji} is the value of the *jth* variable of the sample belonging to the *ith* group. Substitution of the multivariate mean of group A into equation (5.4) for X_{ji} results in R_A. Similarly, R_B can be calculated by substituting the multivariate mean of group B. Substitution of the midpoint between the two group means into the discriminant function equation yields the discriminant index R_0.

In a two-dimensional example, the discriminant function can be plotted as a line on the scatter diagram of the two original variables. The slope of the line is given by λ_2/λ_1. The discriminant index is the point along the discriminant function line that is exactly halfway between the centers of the two groups A and B. In fact, the location of every sample (all X_{ji}'s) along the discriminant function can be found with equation (5.4). These locations can be compared with the location of R_0 to determine whether each sample is classified into the appropriate group. The samples that do not get classified correctly are called misclassified samples. The misclassification ratio, or percent of observations that the discriminant function places into the wrong group, is sometimes taken as an indication of the function's discriminatory power.

Note that the discriminant function given by equation (5.4) is one of many possible solutions available for a discriminant function that maximizes the difference between the two groups. However, all these possible solutions are proportional to the discriminant function given by equation (5.4). This means $\boldsymbol{\lambda}$ given by equation (5.3) as well as all sets $c\boldsymbol{\lambda}$ (where c is any arbitrary constant) will produce applicable solutions to the discriminant function. Therefore, different computer programs may produce different sets of coefficients for $\boldsymbol{\lambda}$ as well as the discriminant function. However, these different sets of results should be proportional to each other. For further information on this aspect, the reader is referred to Davis (2002).

If the means of the two groups are very close together, it will be difficult to tell them apart, especially if both groups have large variances and there is considerable overlap in the areas that they occupy within multivariate space. In contrast, if the two means are well separated and the scatter around the means is small, discrimination will be relatively easy. Hypothesis tests are available to test the significance of the separation between the two groups. The reader is referred to Davis (2002) for details on the hypothesis tests.

5.3.1. An example to illustrate the fundamentals of DFA

Table A.1 provides the data set used to illustrate the fundamentals of DFA. It provides grain-size statistics on samples coming from two groups of sands: group A--modern beach sand from the Gulf Coast of Texas; group B--sands from offshore in the Gulf of Mexico. Both data sets consist of two variables, the median grain size and the grain-size sorting coefficient. Figure 5.12 shows the data for the two groups, **Beach sand** and **Offshore sand** using the two variables as the axes of the plot. With respect to each variable, the scatter of each of the two sand groups is quite large. It is clear from the plot that an adequate separation between the two sands cannot be achieved by considering only one of the two variables. However, the inclined line, which is a linear function of the two variables, shown on the plot provides a satisfactory separation between the two sands. This line has a slope of $(-\lambda_1/\lambda_2)$ and passes through the mean of all the sampling points. Figure 5.13 shows the location of each data point projected on to the discriminant function line along with R_0, R_A and R_B values. This figure was obtained by applying the data given in Table A.1 to the discriminant analysis computer program in the STATISTICA software package. Note that a few samples of group A are located on the group B side of R_0 and a few samples of group

B are located on the group A side. These samples have been misclassified by the discriminant function.

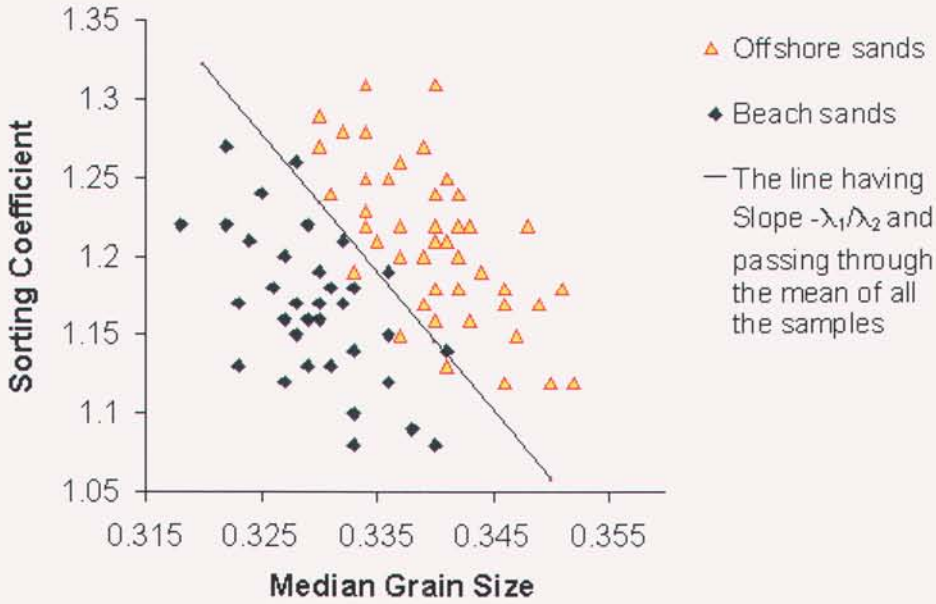


Fig. 5.12 Scatter plot between sorting coefficient and median grain size for the samples coming from the two groups of sands.

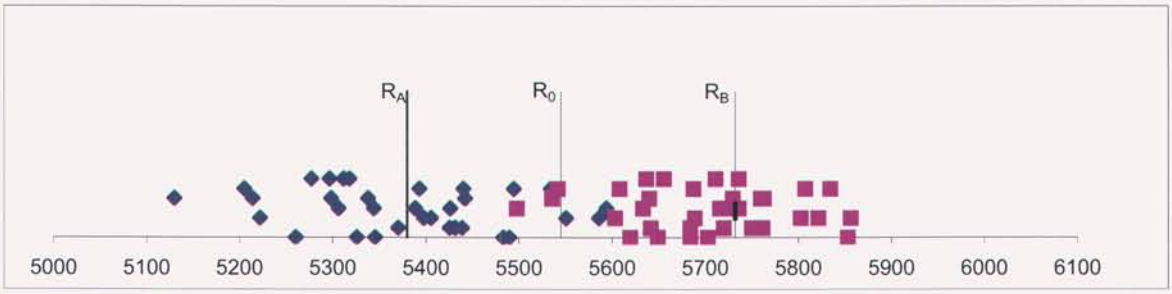


Fig. 5.13 Raw discriminate scores (plotted along the discriminant line) of sample belonging to the two groups of sands

5.4 Verification of the Multivariate Classification of Volcanism for Sengan Region through DFA

In **Section 5.2,** multivariate classification of volcanism for Sengan area was performed using cluster analysis under the following two cases: 1. five-group typification and 2. three-group typification. In this section, multivariate classification of volcanism for the same area will be performed through DFA (see Reyment and Savazzi, 1999) using the same data in an attempt to verify the results obtained through cluster analysis. In **Section**

5.3, the DFA procedure was explained for a two-group situation. The DFA procedure for a multiple group (more than 2) situation is more complex. It is possible to automatically determine some optimal combination of variables so that the first discriminant function (Root 1) provides the greatest overall discrimination between groups, the second function (Root 2), the second most, and so on. Moreover, the functions will be independent or orthogonal, that is their contributions to the discrimination between groups do not overlap. Computationally, such an analysis is called canonical correlation analysis (Khattree and Naik, 2000; StatSoft, 1997) and it determines the successive functions (Roots) and canonical principal values (eigen values). The maximum number of functions will be equal to the number of groups minus one, or the number of variables in the analysis, whichever is smaller. Each function can be given by a linear equation of the type:

$$\text{Function (Root)} = a + b_1 * x_1 + b_2 * x_2 + \ldots + b_m * x_m \qquad (5.5)$$

where a is a constant and b_1 through b_m are regression coefficients.

The two functions (Roots) obtained for the 3-group situation are given in Table 5.3. This table shows that 92.2% of the overall discrimination is provided by Root 1. Root 2 provides the remaining 7.8% of overall discrimination. These two discriminant functions were applied to all 23949 variable vectors to obtain the results shown in Fig. 5.14a through c for the 3-group situation. These figures show very little overlapping between the three groups. Overlapping indicates areas of misclassifications. Only 0.98 percent of points were found to be misclassified through discriminant analysis when compared to the original cluster analysis grouping. All the misclassifications were found to be between adjacent groups in going from lowest volcanic susceptibility group to the highest volcanic susceptibility group. The five functions (Roots) obtained for the 5-group situation are given in Table 5.4. This table shows that 96.1% of the overall discrimination is provided by Roots 1 and 2. Roots 3 and 4 provide the remaining 3.9% of overall discrimination. Figure 5.15 shows the plot between Root 1 and Root 2 for all the 23949 variable vectors for 5-group situation. This figure also shows a very little overlapping or misclassifications between the five groups. Only 1% percent of points were found to be misclassified through discriminant analysis compared to the cluster analysis grouping. All the misclassifications were again found to be between adjacent groups. Therefore, these discriminant function analyses results verify the accuracy of the results obtained through cluster analysis for the different groupings.

Table 5.3 Raw coefficients for variables from canonical analysis to form Roots for 3-group clustering

Variables	Root 1	Root 2
Groundwater Temperature	0.018653	0.053908
Geothermal Gradient	0.064564	-0.02369
pH Value	0.140888	0.462183
Presence of Volcanic Rock	0.400853	-0.78951
Presence of Hydrothermal Alteration	-1.22975	-1.4414

Heat Discharge	0.080293	0.508263
Constant	-5.99967	-4.36238
Eigen Value	4.301879	0.365482
Cumulative Proportion	0.921694	1

Table 5.4 Raw coefficients for variables from canonical analysis to form Roots for 5-group clustering

Variables	Root 1	Root 2	Root 3	Root 4
Groundwater Temperature	0.063682	-0.09105	-0.02006	-0.00014
Geothermal Gradient	0.075316	0.041725	0.00032	-0.02312
pH Value	-0.11804	0.056559	0.89146	0.303091
Presence of Volcanic Rock	0.008005	0.510363	-0.0298	3.211632
Presence of Hydrothermal Alteration	-0.2989	-0.36509	-3.65613	2.093632
Heat Discharge	-0.0104	0.02308	0.923393	0.013315
Constant	-5.95882	-0.37332	-7.63703	-1.40744
Eigen Value	9.187321	0.823358	0.390034	0.013379
Cumulative Proportion	0.882201	0.961263	0.998715	1

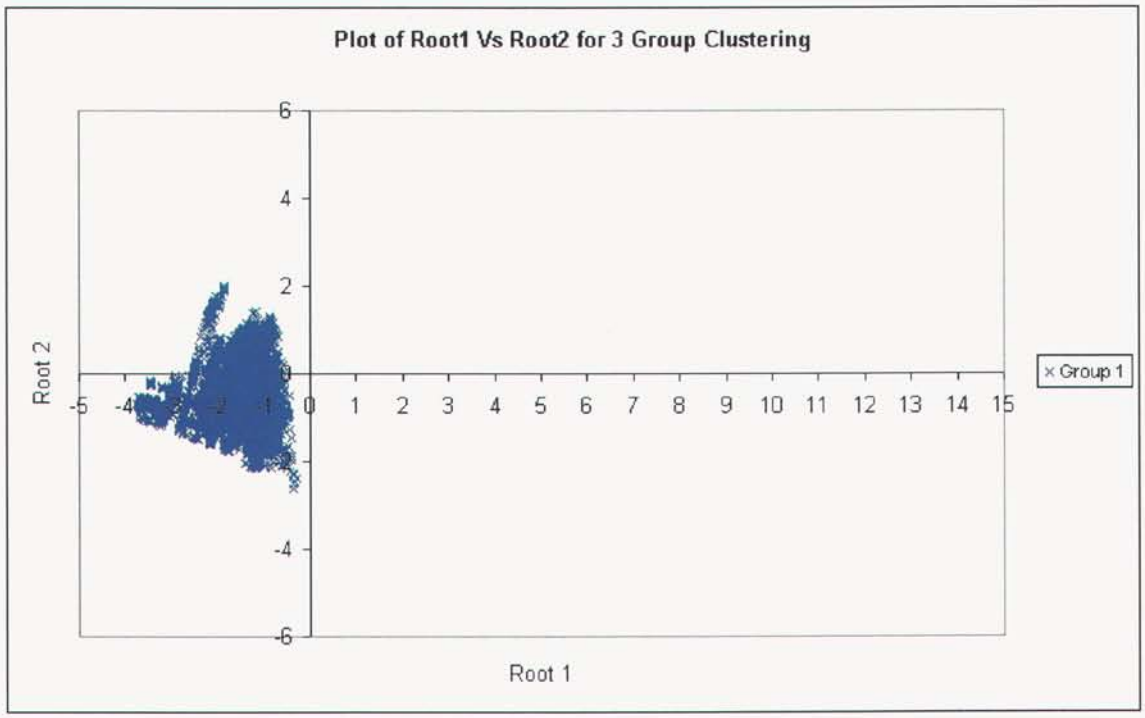


Fig. 5.14 (a) Cluster 1 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis

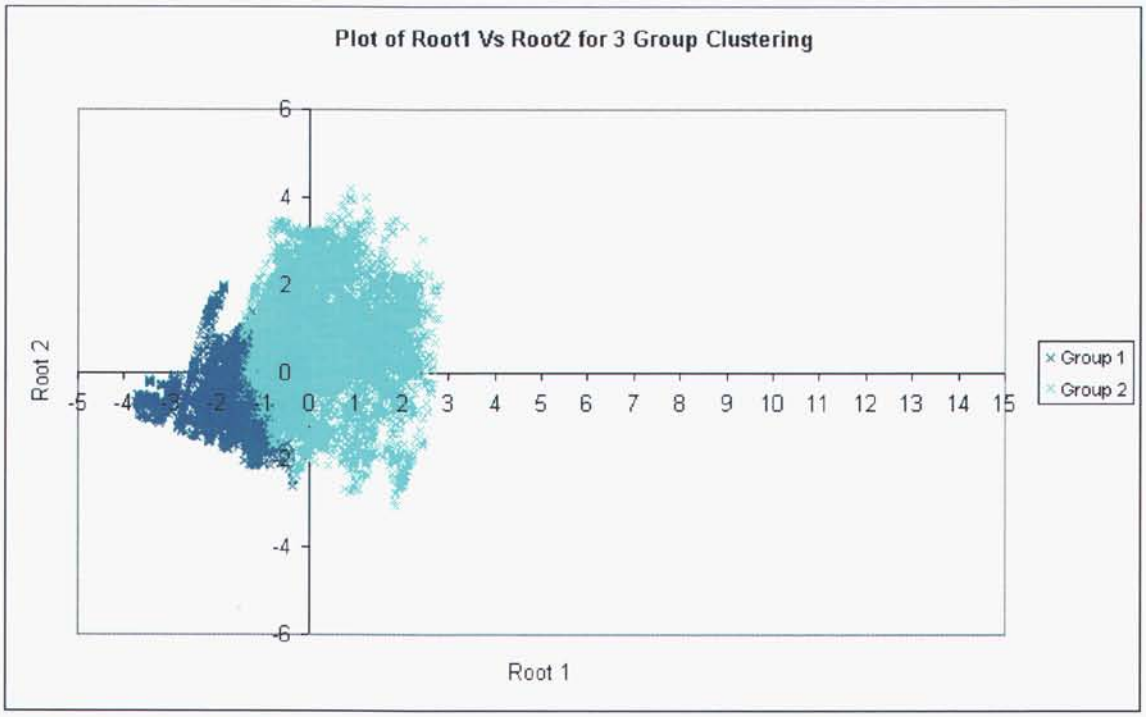


Fig. 5.14 (b) Clusters 1 and 2 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis

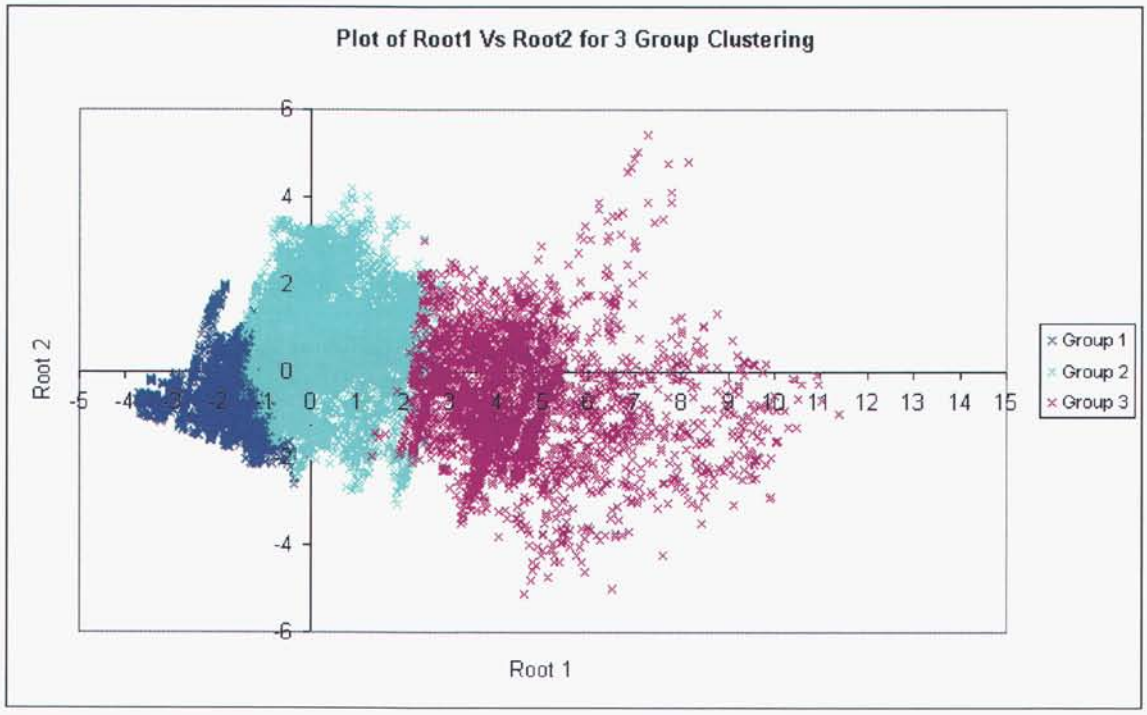


Fig. 5.14 (c) Clusters 1, 2 and 3 on the plot between Root 2 and Root 1 from the 3-cluster canonical analysis

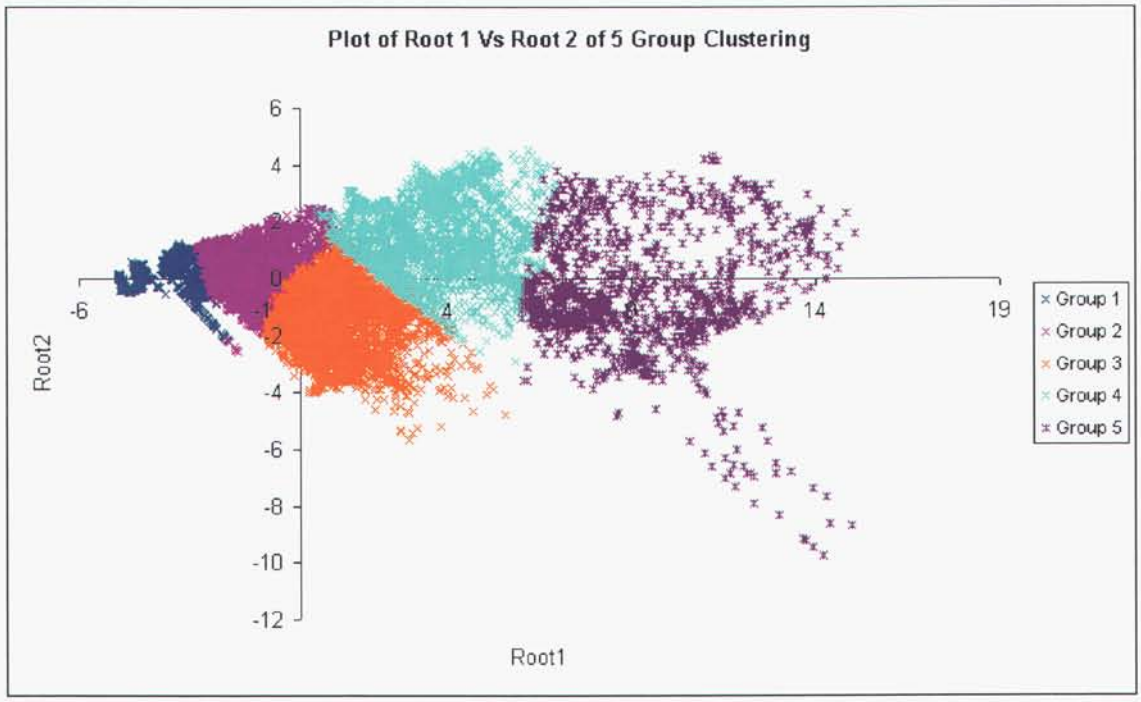


Fig. 5.15 Clusters 1 through 5 on the plot between Root 2 and Root 1 from the 5-cluster canonical analysis

6. REGIONALIZED MAPPING OF VOLCANISM FOR SENGAN REGION

6.1 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 **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 grid point 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 ith 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 **x** is known or given for the sample, then according to Bayes' theorem, the posterior probability of the sample coming from ith group is given by

$$p(i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{\sum_{j=1}^{g} q_j f_j(\mathbf{x})} \tag{6.1}$$

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 $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_\mathbf{i}$, $f_i(\mathbf{x})$ can be given as

$$f_i(\mathbf{x}) = (2\pi)^{-p/2}\left|\boldsymbol{\Sigma}_\mathbf{i}\right|^{-1/2}\exp\left(-0.5d_i^2(\mathbf{x})\right) \tag{6.2}$$

where

$$d_i^2(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_\mathbf{i}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \tag{6.3}$$

is the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}_i$. Substituting equation (6.2) into (6.1) and canceling the constant factor $(2\pi)^{-p/2}$ yields

$$p(i|\mathbf{x}) = \frac{q_i\left|\boldsymbol{\Sigma}_\mathbf{i}\right|^{-1/2}\exp\left(-0.5d_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g} q_j\left|\boldsymbol{\Sigma}_\mathbf{j}\right|^{-1/2}\exp\left(-0.5d_j^2(\mathbf{x})\right)} \tag{6.4}$$

$$= \frac{\exp\left(-0.5D_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g}\exp\left(-0.5D_j^2(\mathbf{x})\right)} \tag{6.5}$$

where

$$D_i^2(\mathbf{x}) = d_i^2(\mathbf{x}) + \ln\left|\boldsymbol{\Sigma}_\mathbf{i}\right| - 2\ln q_i \tag{6.6}$$

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 probability 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 may be used to estimate q_i values. Group means and covariance matrices can be estimated from the geologic variable data available for each group defined at the end of the typification step. Equations 6.7 and 6.8 can be used to estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_\mathbf{i}$, respectively.

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{x}_k \tag{6.7}$$

and

$$\hat{\mathbf{\Sigma}}_{\mathbf{i}} = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)' \qquad (6.8)$$

In equations 6.7 and 6.8, n_i is the number of samples from group i. When the group covariance matrices are not assumed to be equal (i.e. different $\mathbf{\Sigma}_i$ values), the analysis performed through equations (6.5) and (6.6) 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, $\mathbf{\Sigma}_i = \mathbf{\Sigma}$, then equations (6.3) and (6.4) reduce to equations (6.9) and (6.10) given below, respectively.

$$d_i^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \qquad (6.9)$$

$$p(i|\mathbf{x}) = \frac{\exp\left(-0.5d_i^2(\mathbf{x})\right)}{\sum_{j=1}^{g}\exp\left(-0.5d_j^2(\mathbf{x})\right)} \qquad (6.10)$$

The implication of equations (6.9 and 6.10) 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 $\mathbf{\Sigma}$ is estimated by the pooled within-groups covariance matrix given as

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-g}\sum_{i=1}^{g}(n_i - 1)\hat{\mathbf{\Sigma}}_{\mathbf{i}} \qquad (6.11)$$

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.

6.2 Results of Regionalized Mapping for the Sengan Region

In Chapter 5, two cases of typification were performed: 1. three-group typification and 2. five-group typification. For each of these cases, both the quadratic and linear discriminant analyses were performed. Values of q_i used for each case are given in Tables 6.1a and b and 6.2a and b. Negligible differences were found between the quadratic and linear discriminant analysis results. Probabilistic regionalization mapping results obtained for the two cases are shown in Figs. 6.1a through e and Fig. 6.2a through c. Recorded volcano

data are plotted in the same figures for comparison. 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. Tables 6.3 and 6.4 show the distribution of recorded volcanic data among the different groups of volcanic susceptibility level obtained under 5 group and 3 group situations. These tables provide the volcanic data information under two different 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) for 5-group situation. Only two cells out of a total of 73 cells are located in the lowest volcanic susceptibility region (group 1) for 3-group situation. Note that both these cells are located in a low probability region. Sixty seven cells out of a total of 73 cells are located in high and moderate volcanic susceptibility regions (groups 3, 4 and 5) for 5-group situation. Seventy one out of a total of 73 cells are located in high and moderate volcanic susceptibility regions (groups 2 and 3) for 3-group situation. Also, for both the 5-group and 3-group classifications, the number of volcanoes increases as the probability of volcanism increases with each class, 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.

Table 6.1a Values of q_i used for quadratic discriminant analysis for 3-group situation

Group Number	Number of Data	Prior Probability
1	8746	0.3652
2	11675	0.4875
3	3528	0.1473
Total	23949	1.000

Table 6.1b Values of q_i used for linear discriminant analysis for 3-group situation

Group Number	Prior Probability
1	0.333
2	0.333
3	0.333
Total	1.000

Table 6.2a Values of q_i used for quadratic discriminant analysis for 5-group situation

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

Table 6.2b Values of q_i used for linear discriminant analysis for 5-group situation

Group Number	Prior Probability
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2
Total	1.000

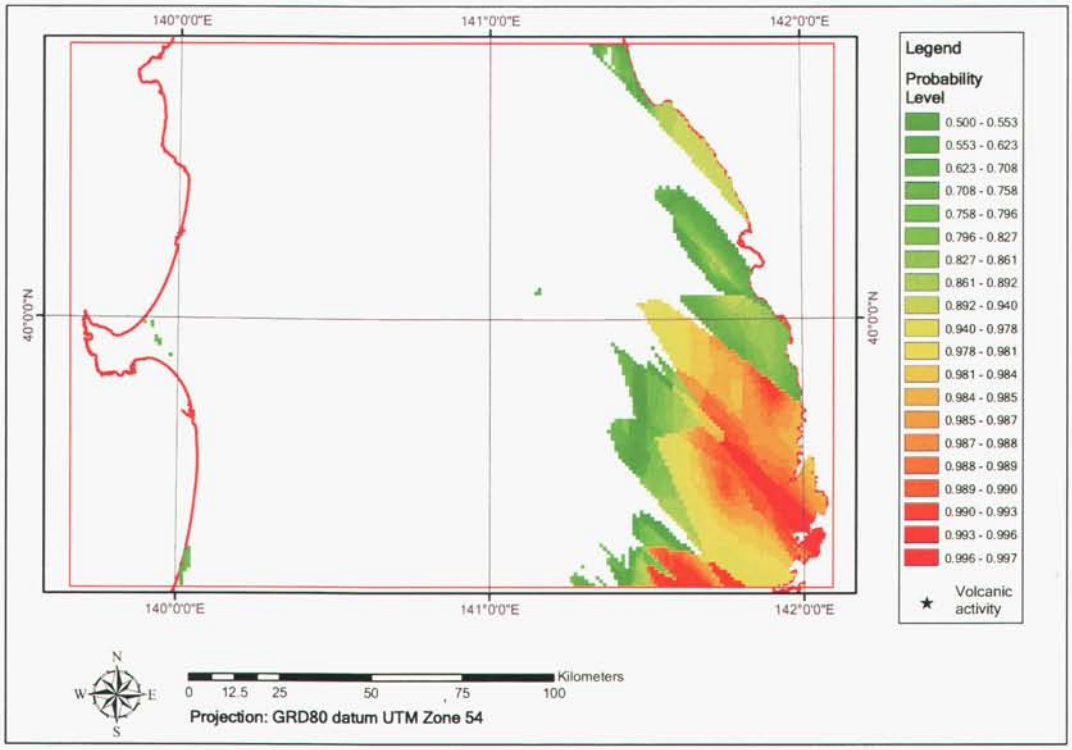


Fig. 6.1 (a) Regionalized probability distribution for group 1 for 5-cluster analysis

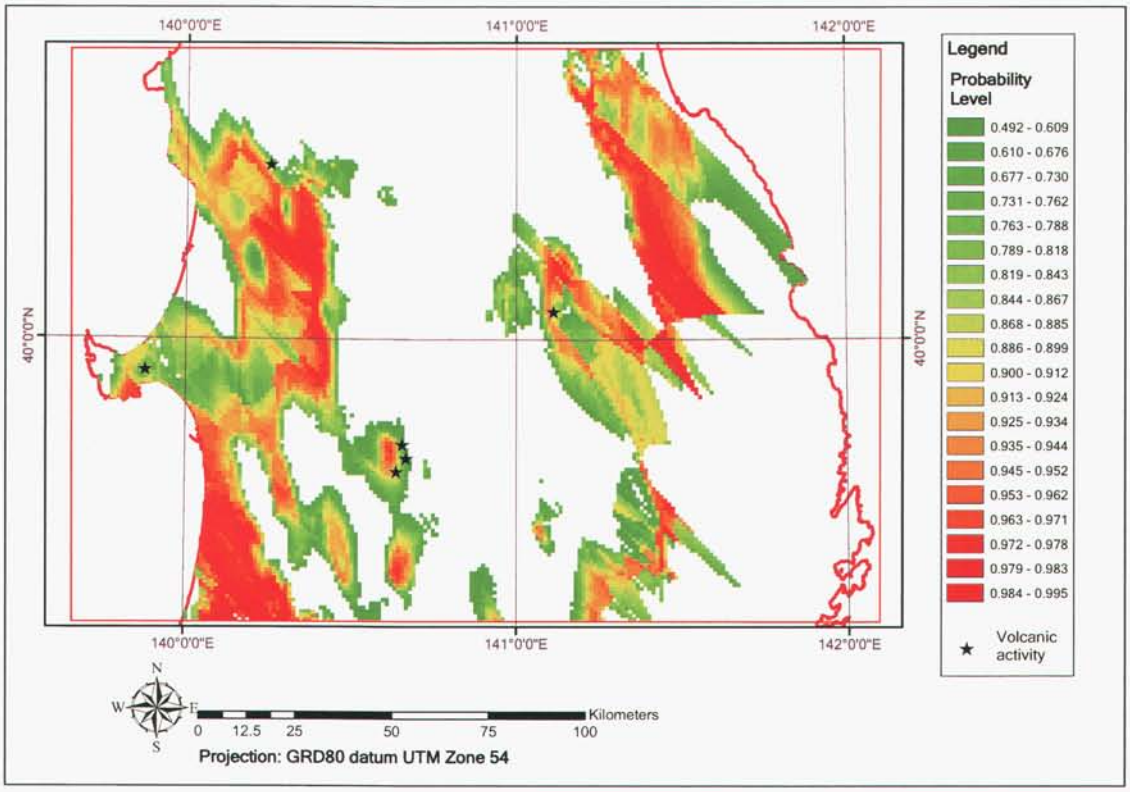


Fig. 6.1 (b) Regionalized probability distribution for group 2 for 5-cluster analysis

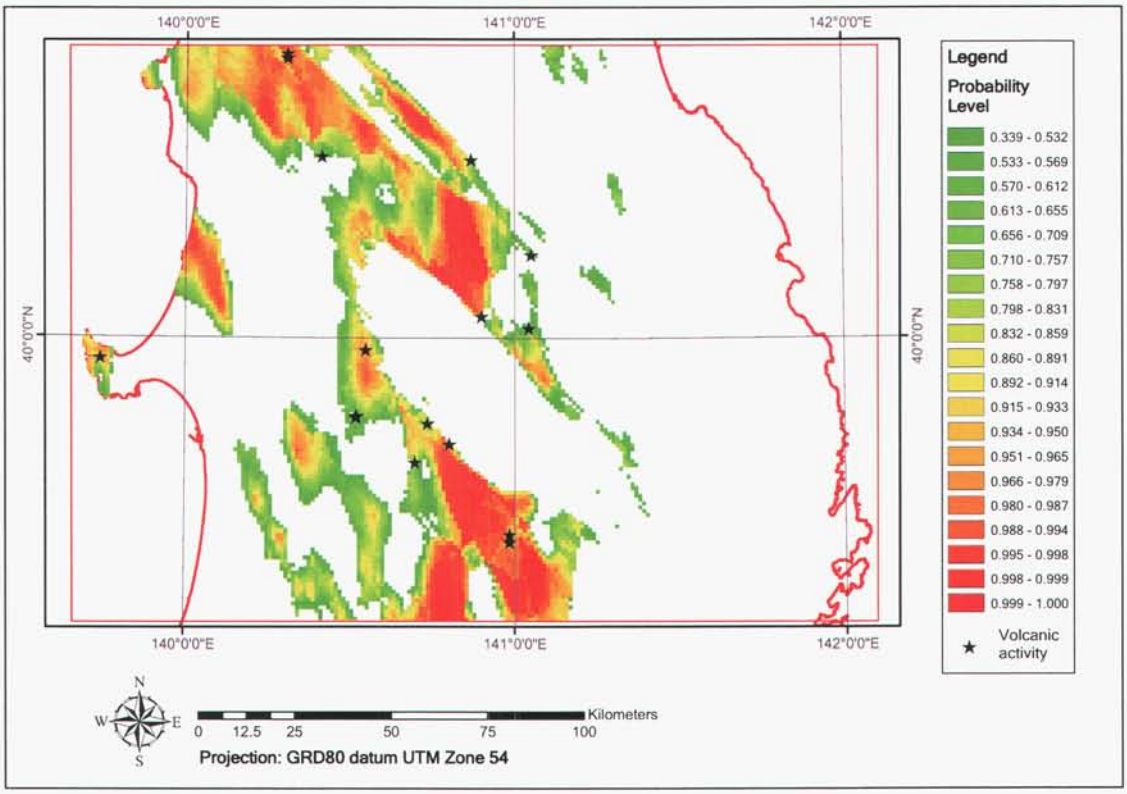


Fig. 6.1 (c) Regionalized probability distribution for group 3 for 5-cluster analysis

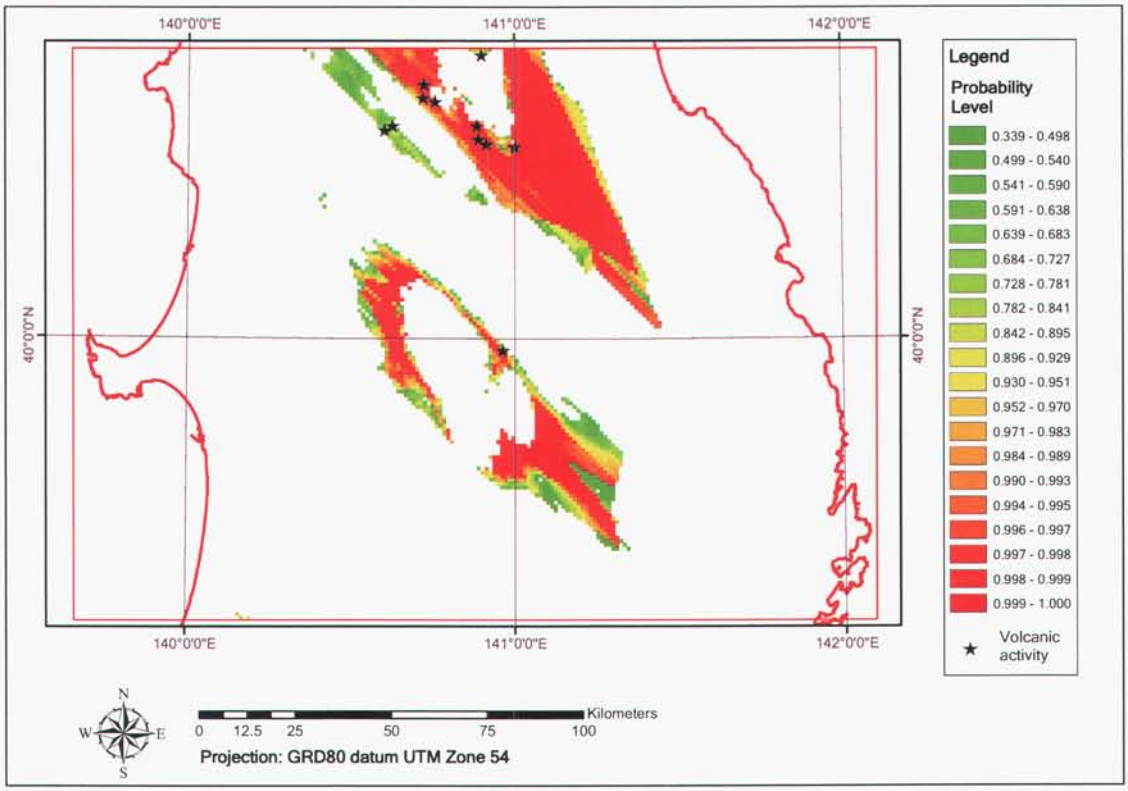


Fig. 6.1 (d) Regionalized probability distribution for group 4 for 5-cluster analysis

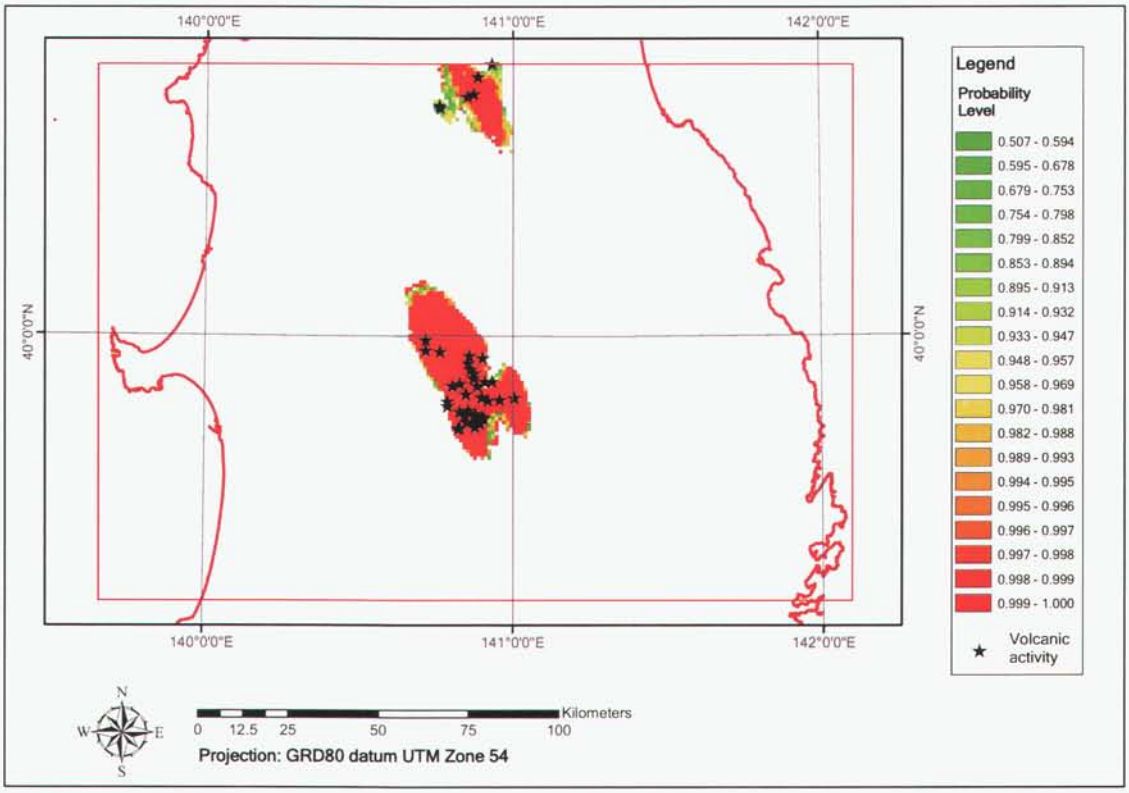


Fig. 6.1 (e) Regionalized probability distribution for group 5 for 5-cluster analysis

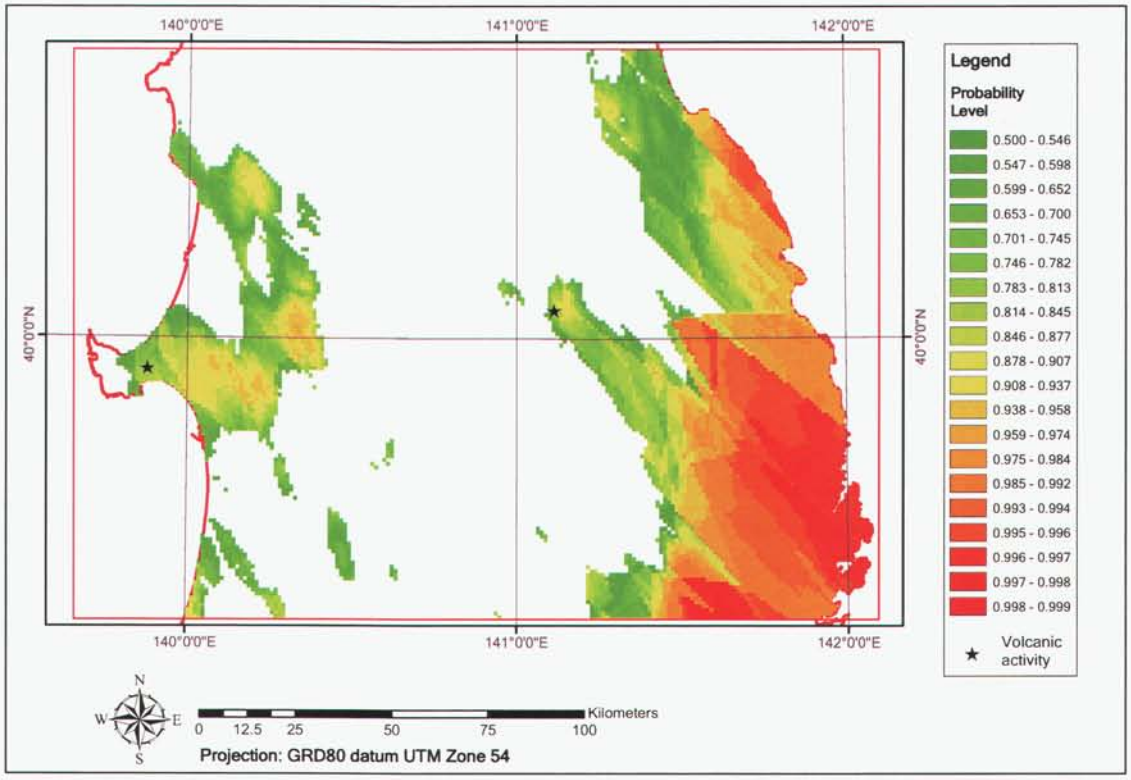


Fig. 6.2 (a) Regionalized probability distribution for group 1 for 3-cluster analysis

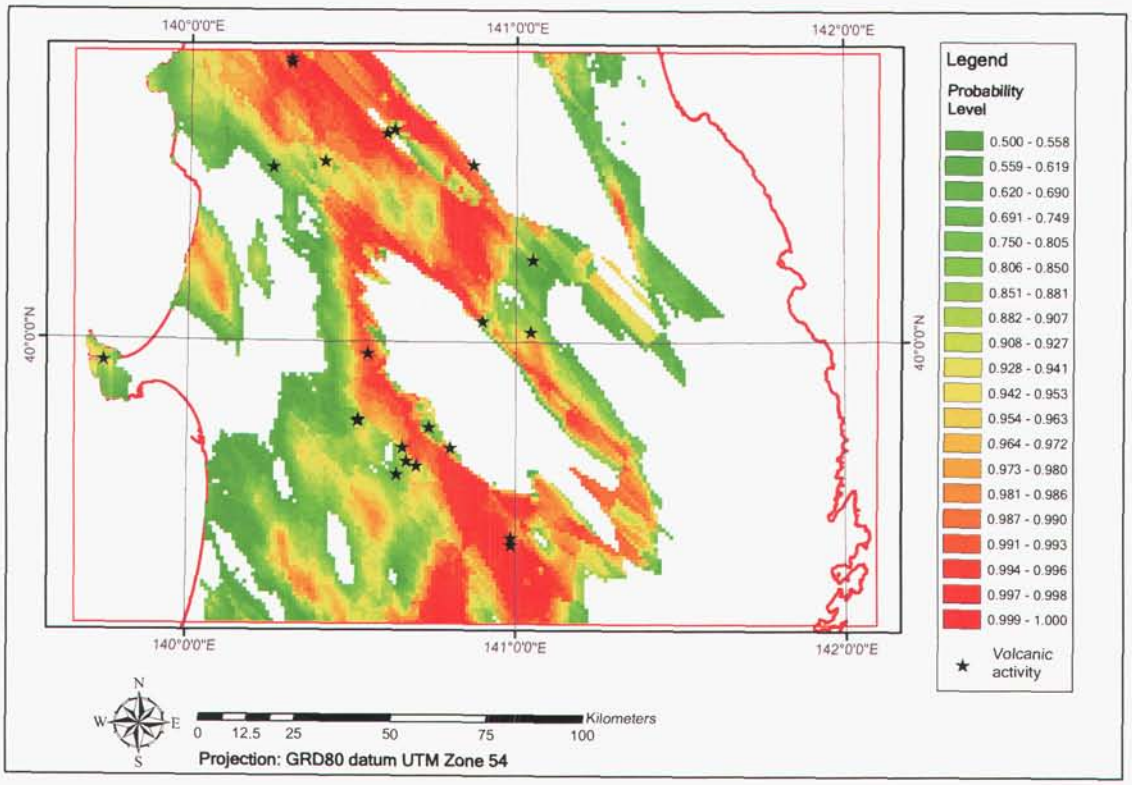


Fig. 6.2 (b) Regionalized probability distribution for group 2 for 3-cluster analysis

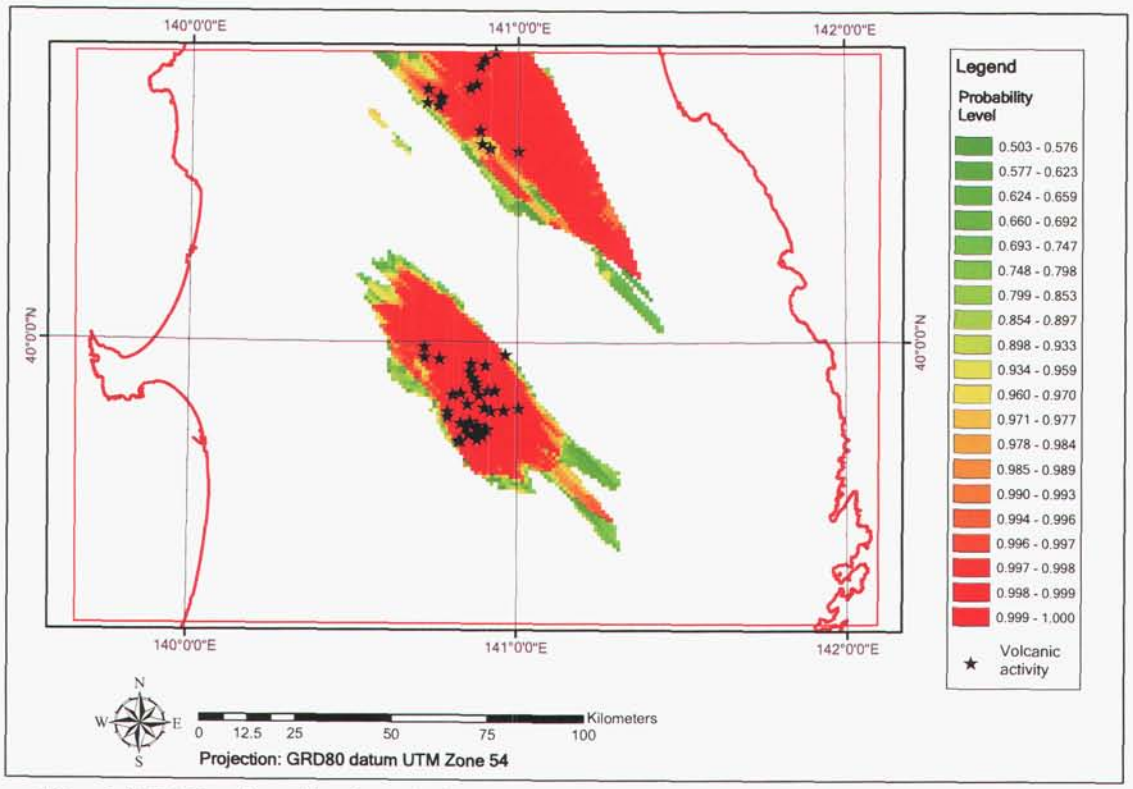


Fig. 6.2 (c) Regionalized probability distribution for group 3 for 3-cluster analysis

Table 6.3 Distribution of volcanic data locations among the five different groups obtained under 5-cluster analysis

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

Table 6.4 Distribution of volcanic data locations among the three different groups obtained under 3-cluster analysis

Volcanic data	Group 1	Group 2	Group 3
# of points	4	34	75
# of cells	2	22	49

6.3 Uncertainty Evaluations of Regionalized Mapping

Two measures are suggested in the literature to assess the uncertainty of the allocation process of the sample based on the calculated posterior probabilities. The first measure is given by (Bohling et al., 1990)

$$1 - \max_i p(i|\mathbf{x}) \tag{6.12}$$

The second measure is based on entropy of classification (Jaynes, 1957; Kitanidis, 1994) and is given by

$$H = \left(-\sum_{k=1}^{g} p_k \ln p_k \right) \Big/ \ln g \tag{6.13}$$

In equation (6.13), $H \rightarrow 0$ when $p_k \rightarrow 1$ for any k and H reaches its maximum value of 1 when all the posterior probabilities are equal. Therefore, H ranges between 0 and 1 with larger values indicating greater uncertainty. Unlike maximum posterior probability, the entropy measure accounts for the entire set of posterior probabilities. Mapping the entropy as calculated in equation (6.13) can be used to improve the spatial definition of group boundaries (Bohling, 1997; Olea, 1999).

The obtained results for the two measures of uncertainty are shown in Figs. 6.3 and 6.4. Each recorded volcanic activity data along with the volcanic activity group it belongs to under the 3-group situation are also shown in the same figures. The histogram given in Figs. 6.5 show the uncertainty level of the group estimations at different locations where volcanic data are available for each group obtained under the 3-group situation. Note that only 2 out of a total of 73 volcanic data cells are located in the lowest volcanism group (group 1). Figures 6.3, 6.4 and 6.5 show that the uncertainty of the mapping estimations is

relatively high at the locations of these two data cells. Twenty two out of a total of 73 volcanic data cells are located in the moderate volcanism group (group 2). Figures 6.3, 6.4 and 6.5 show that the uncertainty of the mapping estimations is relatively moderate on the average for these 22 data cell locations. Forty nine out of a total of 73 volcanic data cells are located in the high volcanism group (group 3). Figures 6.3, 6.4 and 6.5 show that the uncertainty of the mapping estimations is relatively low on the average for these 49 data cell locations. Similar figures of uncertainty measures are repeated for the 5-group situation (Figs. 6.6, 6.7 and 6.8). Figures 6.6, 6.7 and 6.8 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 three 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 6.6, 6.7 and 6.8 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 regionalization mapping to reduce the uncertainty level of the estimations.

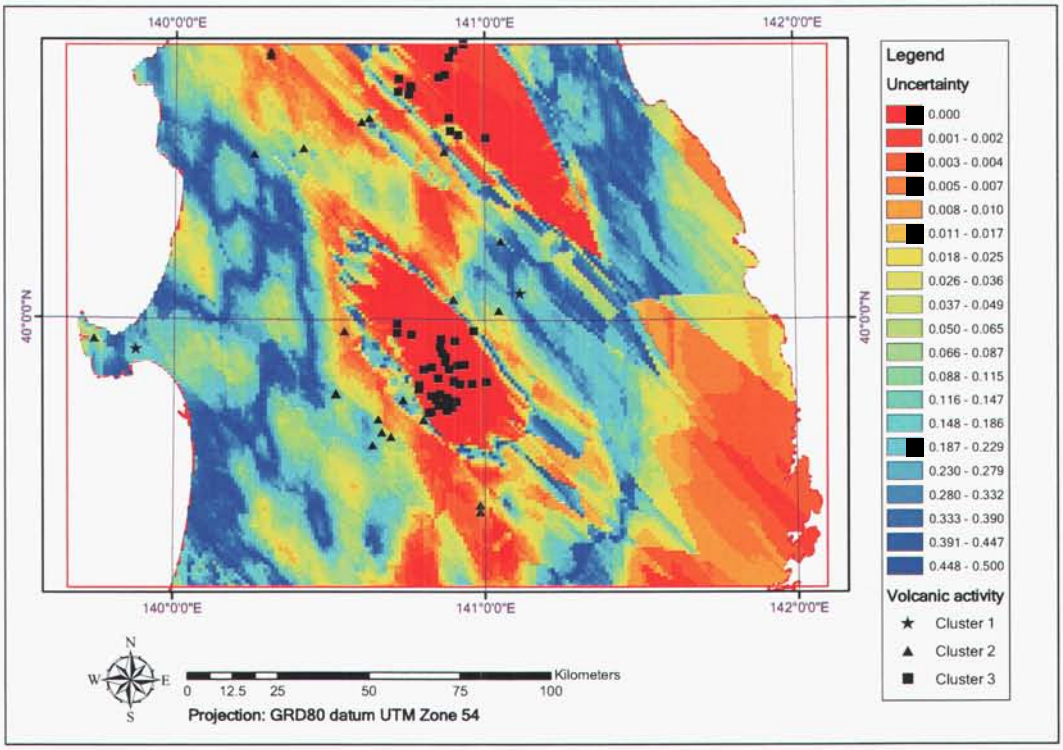


Fig. 6.3 Regionalized distribution of the first measure of uncertainty for 3-group situation along with the available volcanic activity data located in different clustering groups

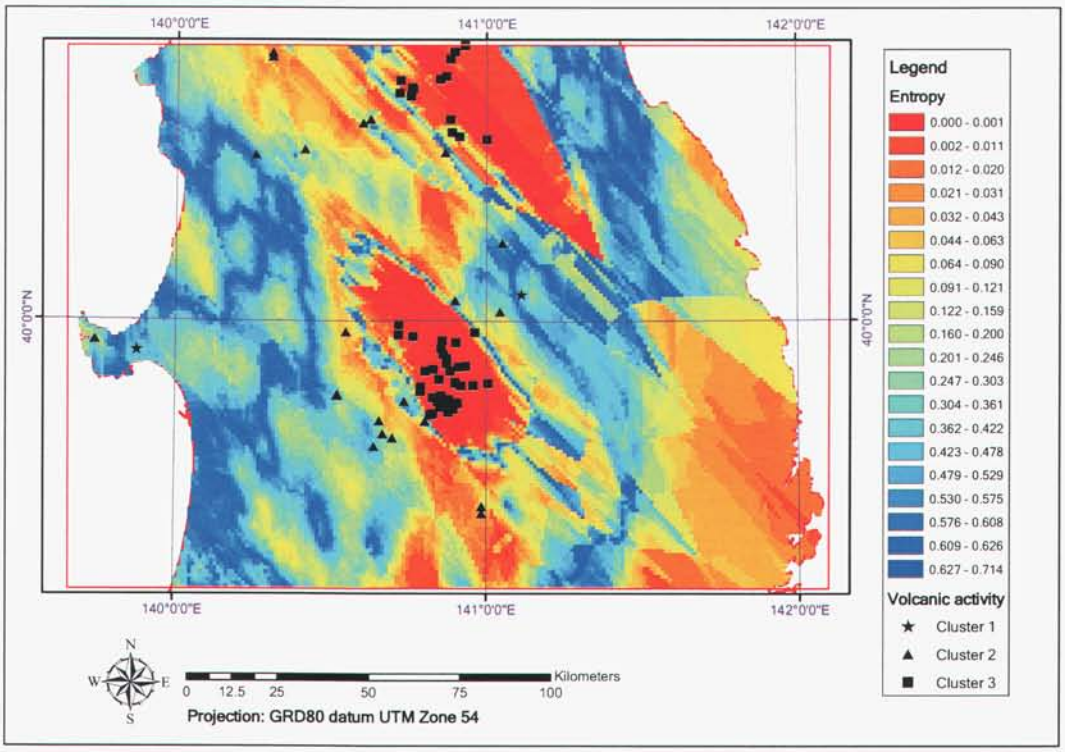
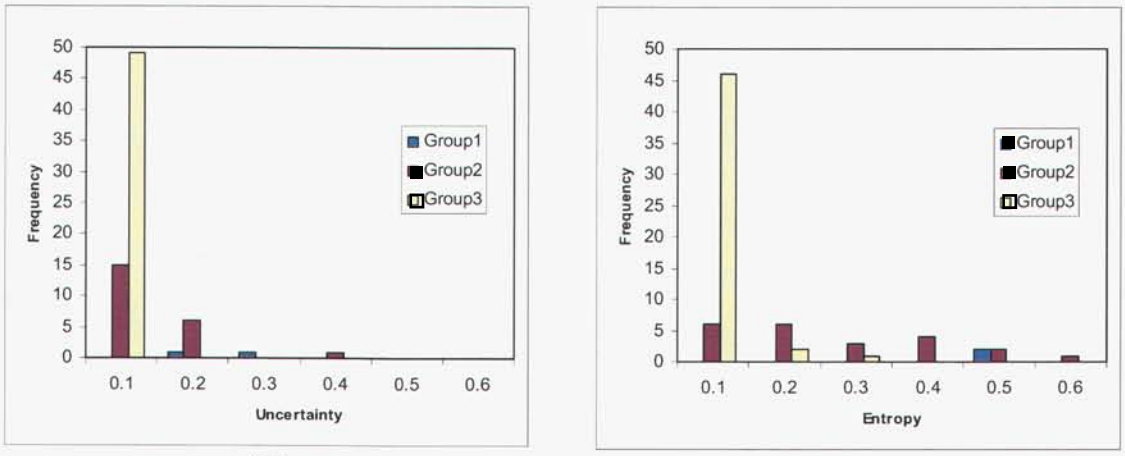


Fig. 6.4 Regionalized distribution of entropy (the second measure of uncertainty) for 3-group situation along with the available volcanic activity data located in different clustering groups



(a) (b)

Fig. 6.5 Histograms of the frequency of volcanic data located in different regionalized groups and the associated uncertainty level (a) the first measure of uncertainty, (b) entropy for each group for 3-group situation

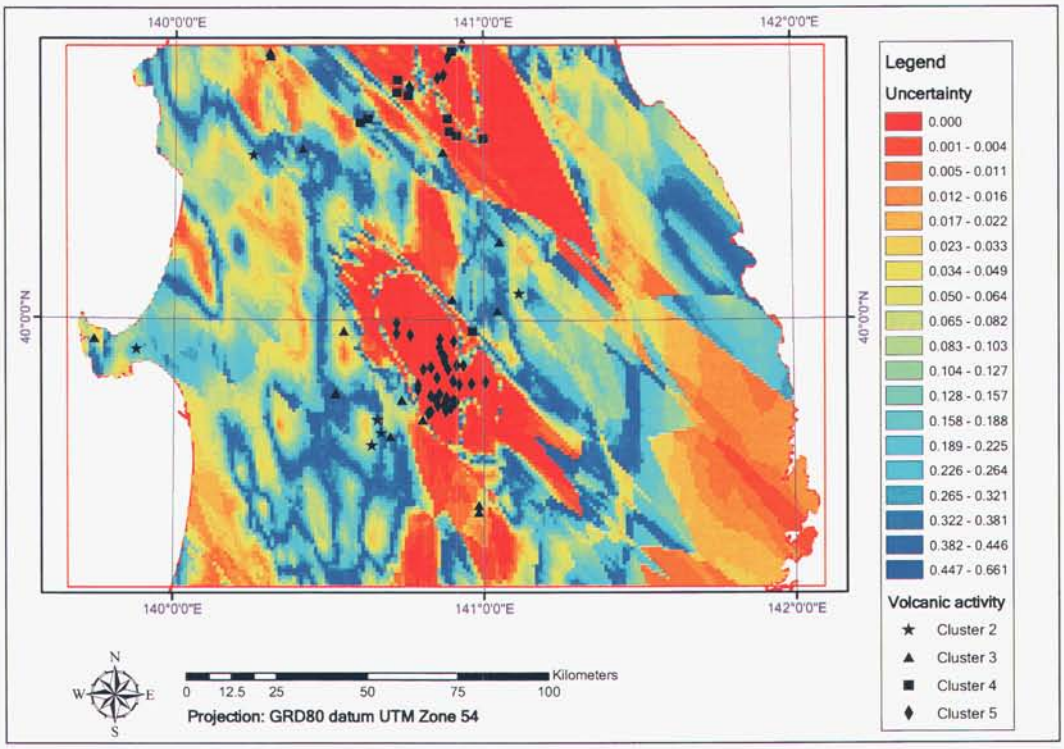


Fig. 6.6 Regionalized distribution of the first measure of uncertainty for 5-group situation along with the available volcanic activity data located in different clustering groups

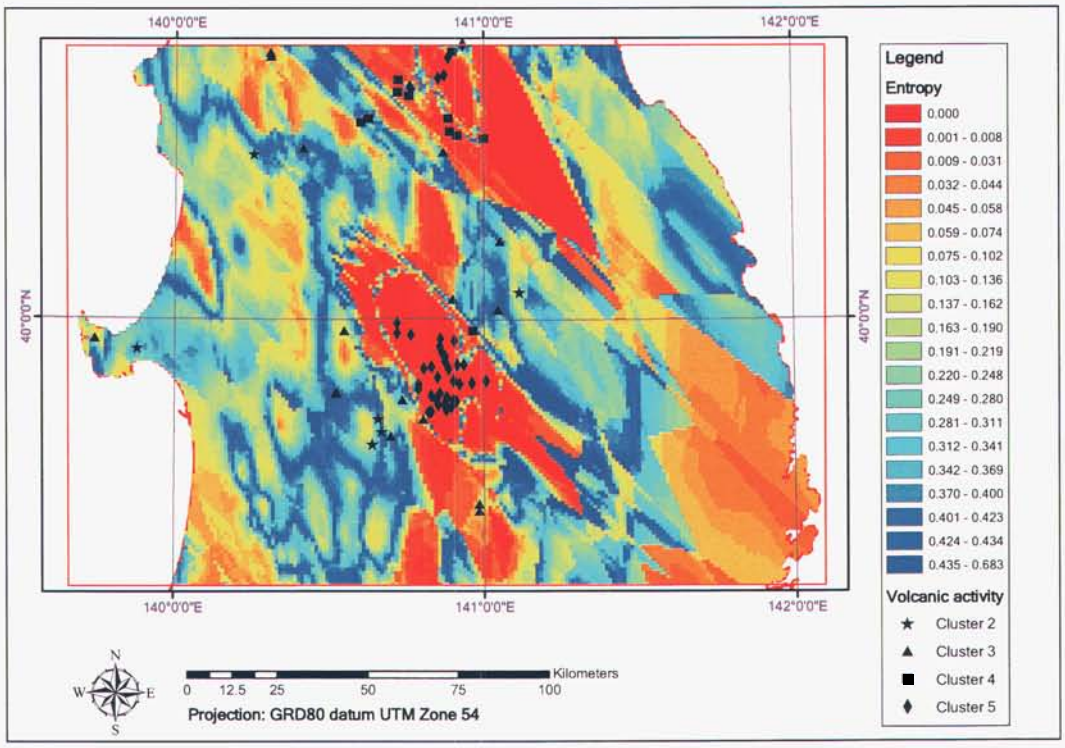
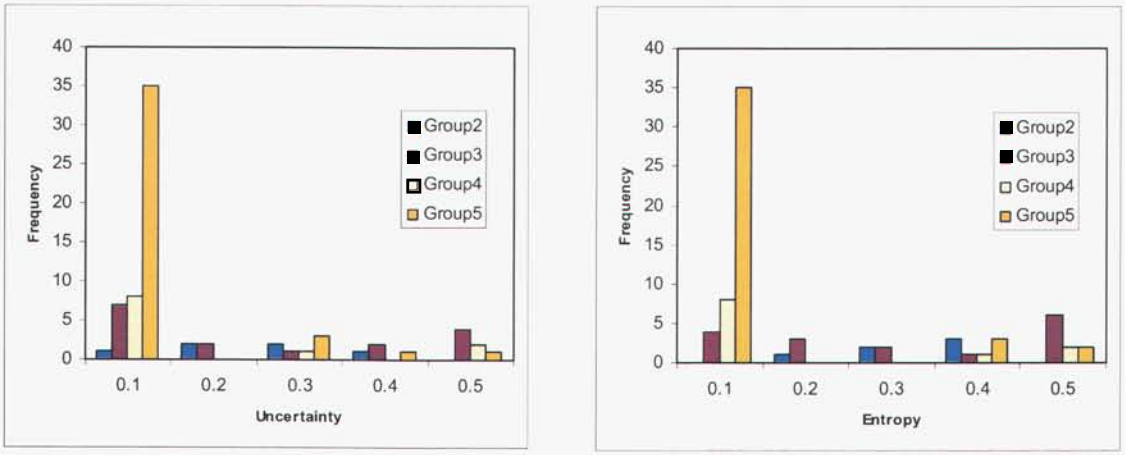


Fig. 6.7 Regionalized distribution of entropy (the second measure of uncertainty) for 5-group situation along with the available volcanic activity data located in different clustering groups



(a) (b)

Fig. 6.8 Histograms of volcanic data located in different regionalized groups and the associated uncertainty level (a) the first measure of uncertainty, (b) entropy for each group for 5-group situation

Scatter plots between the entropy of classification and standard error of the predicted value from kriging were created for each of the first four important geologic variables for the 1 km cell centers belonging to each of the groups (clusters) 1 through 5. Similar plots were obtained for the 3-group situation. None of these plots showed a strong correlation between the entropy and the standard error of each predicted variable. Note that the entropy provides the uncertainty of the 1 km cell belonging to the considered cluster by combining the estimated values obtained from either kriging or some other technique for all the six geologic variables. On the other hand, the standard error of the prediction is calculated separately for each variable. It is important to note that the spatial distribution of the available data was not the same for each of the chosen 6 important geologic variables. That means the spatial distribution of the standard error of the prediction is different for each of the aforementioned six important geologic variables. The entropy can be considered as a measure of classification uncertainty that combines uncertainty across all six important geologic variables. In addition, for the presence of volcanic rock and presence of hydrothermal alteration variables, zero values were assigned for the standard error of the kriging in order to complete the geologic variable vectors for all the 1 km cell locations. Due to these reasons it was not expected that strong correlations between the entropy and the standard error of each predicted variable would exist.

7. SUMMARY AND CONCLUSIONS

This report provides the procedures used and results obtained for the U of A study on hierarchical probabilistic regionalization of volcanism for Sengan region in Japan. Multivariate statistical techniques and geostatistical interpolation techniques were applied on the geologic data provided by NUMO to achieve the required goals of the project. The summary of the steps used along with the results obtained and the conclusions arrived at are given below:

1) UTM zone 54 projected coordinate system and a 1 km square regular grid system were selected to show the available geologic data for Sengan region using ArcGIS 8.x maps.
2) Maps for each geological variable were visually compared with the map of recorded volcanic activity data (Fig. 3.1) to determine the geologic variables that are correlated to volcanic activity. These variables: geothermal gradient, groundwater temperature, heat discharge, groundwater pH value, presence of volcanic activity and presence of hydrothermal alteration were labeled as the important variables for volcanism. Definition of these variables from the available data is consistent with the initial variables identified in the volcanology workshop held in September of 2003 (see Arnold, et al., 2003).
3) For each of the aforementioned 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. Maps obtained for the important geologic variables are given in Figs. 4.6, 4.12, 4.18, 4.24, 4.26 and 4.27.
4) Cluster analysis was performed on the 23949 complete variable vectors to classify them to (a) 5 groups and (b) 3 groups of volcanism level spanning from lowest possible volcanism to highest possible volcanism with increasing group number. The obtained results are shown in Figs. 5.1 through 5.11 and Tables 5.1 and 5.2.
5) Discriminant function analysis performed on the groups obtained in step 4 verified the results obtained through cluster analysis. For the 5-group situation, only 1 % of the variable vectors were found to be misclassified when compared with the cluster analysis results. For the 3-group situation, only 0.98 % of the variable vectors were found to be misclassified when compared with the cluster analysis results. The discriminant plots obtained for the 3-group and 5-group situations are shown in Figs. 5.14 and 5.15.
6) Volcanism group results obtained through cluster analysis were used with Bayes' theorem to construct maps showing the probability of group membership for each of the volcanism groups obtained in step 4 for both the 5-group and 3-group classifications. These maps are shown in Figs. Figs. 6.1a through e and Figs. 6.2a through c. These maps show good comparisons with recorded volcanic data of the Sengan region. Note that no volcanic data exist in the group 1 region of the 5-group situation. The highest probability area (i.e the lowest uncertainty) within group 1 has the chance of being the no volcanism region.

7) Maps were constructed to express uncertainty of the regionalized mapping estimations. These maps express uncertainty through the classification entropy as calculated with Eqn. 6.13 (Figs. 6.4 and 6.7) and the complement of the maximum posterior probability of group membership calculated with Eqn. 6.12 (Figs. 6.3 and 6.6). The recorded volcanism data are also plotted on the same maps to see the uncertainty level of the estimations at the locations where volcanism exists. The two uncertainty measures lead to the same conclusions. The volcanic data cell locations that are in the high volcanism regions (group 4 and 5 of 5-group situation and group 3 of 3-group situation) show on the average relatively low uncertainty mapping estimations (Figs. 6.5 and 6.8). On the other hand, the volcanic data cell locations that are in the low volcanism regions (group 2 of 5-group situation and group 1 of 3-group situation) show on the average relatively high uncertainty mapping estimations (Figs. 6.5 and 6.8). The volcanic data cell locations that are in the medium volcanism regions (group 3 of 5-group situation and group 2 of 3-group situation) show on the average relatively moderate uncertainty mapping estimations (Figs. 6.5 and 6.8). 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 regionalization mapping to reduce the uncertainty level of the estimations.

8. ACKNOWLEDGEMENTS

Funding for this work was provided by the Nuclear Waste Management Organization of Japan (NUMO). 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

9. REFERENCES

Anderberg, M.R., 1973, *Cluster Analysis for Applications*: Academic Press, New York, 359p.

Anderson, T.W., 1984, *An Introduction to Multivariate Statistical Analyses, 2nd ed.*: John Wiley & Sons, Inc., New York, 704 p.

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, U.K., 367p.

Bishop, C.M., 1995, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 482p.

Bohling G. C., 1997, GSLIB-Style Programs for Discriminant Analysis and Regionalized Classification, *Computer & Geosciences*, 23,no 7, pp 739-761.

Bohling, G. C., J. Harff, J.C. Davis, 1990, Regionalized Classification: Ideas and Applications, Proceedings, Fifth Canadian /American Conference on Hydrology: ed. S. Bachu, pp 229-242, NWWA, Dublin, Ohio.

Davis, J.C., 2002, *Statistics and Data Analysis in Geology*, 3rd ed.: John Wiley & Sons, Inc., 638p.

Deutsch, C.V., A.G. Journel, 1998, GSLIB: *Geostatistical Software Library and User's Guide, Second Edition*, Oxford University Press, New York, 340p.

Everitt, B.S., 1993, *Cluster Analysis, 3rd ed.*: Edward Arnold, Cambridge (U.K), 170p.

Fernanadez, J.A.M., C.B.Vidal, V.P.Glahn, 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, A.D., 1999, *Classification, 2nd ed.*: Chapman & Hall/CRC, Boca Raton, Fla., 256p.

Harff, J., J.C.Davis, 1990, Regionalized Mapping in Geology by Multivariate Classification, *Mathematical Geology*, V. 22, No 5, pp 573-588.

Harff, J.E., J.C.Davis, W.Eiserbeck, 1993, Prediction of Hydrocarbons in Sedimentary Basins, *Mathematical Geology*, V.25, No 7, pp 925-936.

Harff, J., J.C.Davis, R.A.Olea.1991, Quantiative Assessment of Mineral Recourses with an Application to Petroleum Geology: *Nonrenewable Resources*, V.1, No 1, pp 74-84.

Harff, J., J.C.Davis, L.Watney, G.C. Bohling, J.C.Wong, 1989, Regionalization of Western Kansas Based on Multivariate Classification of Statigraphic Data from Oil Wells: Open-File Report.89-21, Kansas Geological Survey, Univ of Kansas, Lawrence, Kansas, 26p.

Harff, J., W.Eiserbeck, K.Hoth, J.Springer, 1990, Computer-Assisted Basin Analysis and Regionalization Aid the Search for Oil and Gas, *Geobyte*, pp 11-14.

Isaaks, E.H., R.M.Srivastava.1989, *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 550p.

Jaynes, E.T. 1957, Information Theory and Statistical Mechanics, *Phys. Review, V.* 106, pp. 620-630.

Johnston, K., J. M. Ver Hoef, K. Krivoruchko and N. Lucas N, 2001, Using ArcGIS Geostatistical Analyst. ESRI.

Journel, A.G., C.J.Huijbregts, 1978, *Mining Geostatistics*: Academic Press, New York, 600p.

Khattree, R., D. N. Naik, 2000, *Multivariate data reduction and discrimination with SAS Software*, John Wiley, New York.

Kitanidis, P.K., 1994, The Concept of Dilution Index, *Water Resources Research*, V. 30, No. 7, pp. 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, 211p.

McCoy J. and J. Kevin J., 2001, Using ArcGIS Spatial Analyst. ESRI.

McLachlan, G.J., 1992, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Inc., New York, 526p.

Moline, G.R., J.M.Bahr, 1995, Estimating Spatial Distributions of Heterogeneous Subsurface Characteristics by Regionalized Classification of Electrofacies, *Mathematical Geology*, V.27, No 1, pp 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.

Reyment, R.A. and E. Savazzi, 1999, *Aspects of Multivariate Statistical Analysis in Geology*, Elsevier , Amsterdam, 285 pp.

SAS, 1989, SAS/STAT User's Guide, Vol.1, (version 6), 4th ed. SAS Institute Inc., Cary, North Carolina.

SAS, 2002, SAS/STAT User's Guide, Version 9. SAS Institute Inc., Cary, North Carolina.

Schalkoff, R.J., 1997, *Artificial Neural Networks*: McGraw-Hill, Book co. New york, 422p.

Sneath, P.H.A., R.R. Sokal, 1973, *Numerical Taxonomy*, W.H. Freeman & Co., San Francisco, 573p.

StatSoft, 1997, STATISTICA manual, '97 edition.

Zaknich, A., 2003, *Neural Network for Intelligent Signal Processing*: World Scientific, Singapore, 508p.

Appendix A

Table A.1 Input data used for discriminant analysis (file SANDS.TXT from the URL www.wiley.com/college/davis)

No	Sample	Mean grain size	Sorting Coefficient
1	A	.333	1.080
2	A	.340	1.080
3	A	.338	1.090
4	A	.333	1.100
5	A	.323	1.130
6	A	.327	1.120
7	A	.329	1.130
8	A	.331	1.130
9	A	.336	1.120
10	A	.333	1.140
11	A	.341	1.140
12	A	.328	1.150
13	A	.336	1.150
14	A	.327	1.160
15	A	.329	1.160
16	A	.330	1.160
17	A	.323	1.170
18	A	.328	1.170
19	A	.332	1.170
20	A	.331	1.180
21	A	.326	1.180
22	A	.333	1.180
23	A	.330	1.190
24	A	.336	1.190
25	A	.327	1.200
26	A	.324	1.210
27	A	.332	1.210
28	A	.322	1.220
29	A	.329	1.220
30	A	.325	1.240
31	A	.328	1.260
32	A	.322	1.270
33	A	.318	1.220
34	A	.330	1.170
35	B	.339	1.120
36	B	.346	1.120
37	B	.350	1.120
38	B	.352	1.130
39	B	.341	1.150
40	B	.347	1.150
No	Sample	Mean grain Size	Sorting Coefficient
41	B	.337	1.160
42	B	.343	1.160
43	B	.340	1.170
44	B	.346	1.170
45	B	.349	1.170
46	B	.339	1.180
47	B	.342	1.180
48	B	.346	1.180
49	B	.351	1.180
50	B	.340	1.190
51	B	.344	1.190
52	B	.333	1.200
53	B	.337	1.200
54	B	.339	1.200
55	B	.342	1.200
56	B	.339	1.210
57	B	.340	1.210
58	B	.341	1.210
59	B	.335	1.220
60	B	.337	1.220
61	B	.340	1.220
62	B	.343	1.220
63	B	.334	1.220
64	B	.348	1.220
65	B	.337	1.220
66	B	.342	1.230
67	B	.334	1.240
68	B	.340	1.240
69	B	.342	1.240
70	B	.331	1.250
71	B	.336	1.250
72	B	.341	1.250
73	B	.334	1.260
74	B	.337	1.270
75	B	.339	1.270
76	B	.330	1.280
77	B	.334	1.280
78	B	.332	1.290
79	B	.330	1.310
80	B	.334	1.310
81	B	.340	1.210

Distribution:
Internal

Dept. 6115 MS 0735
Ray Finley
Tom Lowry
Sean McKenna (4)
Barry Roberts

Dept. 6851 MS 0778
Bill Arnold

Dept. 6874 MS-0779
Hong-Nian Jow

Dept. 6805 MS 0731
NWM Library (5)

Dept. 8945-1 MS 9018
Central Technical Files

Dept. 9616 MS 0899
Technical Library (2)

External
Kazumi Kitayama
NUMO Headquarters
MitaNN Bldg., 1-23 Shiba 4-Chome,
Minato-ku, Tokyo, Japan, 108-0014.

Hiroyuki Tsuchi (3)
NUMO Headquarters
MitaNN Bldg., 1-23 Shiba 4-Chome,
Minato-ku, Tokyo, Japan, 108-0014.

Junichi Goto (3)
NUMO Headquarters
MitaNN Bldg., 1-23 Shiba 4-Chome,
Minato-ku, Tokyo, Japan, 108-0014.

Dr. Katsuhiko Ishiguro
NUMO Headquarters
MitaNN Bldg., 1-23 Shiba 4-Chome,
Minato-ku, Tokyo, Japan, 108-0014.

Keiichiro Wakasugi
NUMO Headquarters
MitaNN Bldg., 1-23 Shiba 4-Chome,
Minato-ku, Tokyo, Japan, 108-0014.

Dr. Pinnaduwa H.S.W. Kulatilake (5)
The University of Arizona
Department of Mining and Geological Engineering
1235 E. North Campus Drive, Mines 229
PO Box 210012
Tucson, AZ 85721

Dr. Tim Coburn
Department of Management Science
Abilene Christian University
ACU Box 29315
Abilene, TX 79699

Dr. Erik Webb
201 3rd Street NW, Suite 710
Albuquerque, NM 87104