Training course on

Documentation and Information Management of Plant Genetic Resources

22 November – 3 December 1998

ICARDA, Aleppo, Syria

Version 1.0

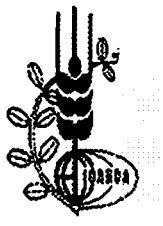
November 1998

Analysis of Plant Genetic Resources Data

Abdallah Bari [1] and Murari Singh [2]

[1] International Plant Genetic Resources Institute (IPGRI), Central and West Asia, and North Africa (CWANA), PO Box 5466, Aleppo, Syria

[2] International Centre for Agricultural Research In the Dry Areas (ICARDA), PO Box 5466, Aleppo, Syria



Version 1.0

November 1998

Analysis of Plant Genetic Resources Data

Abdallah Bari [1] and Murari Singh [2]

[1] International Plant Genetic Resources Institute (IPGRI), Central and West Asia, and North Africa (CWANA), PO Box 5466, Aleppo, Syria

[2] International Centre for Agricultural Research In the Dry Areas (ICARDA), PO Box 5466, Aleppo, Syria

1. INTRODUCTION

Rational conservation of bio-diversity is a fundamental element for preserving the existence of humanity and other species exhibiting life on the planet. In the Food and Agricultural Organisation (FAO) Global Plan of Action, it is mentioned that surveying and inventorying are prerequisite of a rational conservation approach. Surveys have been carried out by researchers around the world for collecting the germplasm of various plant species with an objective to maintain them in-situ and/or ex-situ in order to averse or minimise chances of their loss. Gene banks collect large volume of the plant material. Acquiring knowledge of the genetic material is key to the exploitation of the germplasm resources from various points of view and objectives.

Information generated through analyses of these data can facilitate in the planing and decision making with regard to conservation and utilisation. Diversity has been used as criteria number in locating sites for conservation as well as for carrying out monitoring activities. It helps to see how much harbours a site or an accession in terms of variability. It is also an indicator of the wellbeing of ecological systems. To be able to measure and locate this diversity, statistical/mathematical tools are of immense value.

This report presents various methods of data collection, preparation/structure of the databases for storing, updating and retrieving the data, and data analysis methods. Here, we discuss a case study from Cyprus with area of collection covering the range of 32.00° N to 34.52° N (longitude) and 34.5° E to 35.5° E (latitude). We have kept the data analysis methods restricted

to the aspects of evaluating diversity of the species in a region, rank and abundance of the species, abundance of a species in association with the site of collection of the data, with geographical region and with the rainfall pattern. The estimation of expected number of species in a given region is given. The abundance of species showing spatial variation and therefore have been modelled using spatial models.

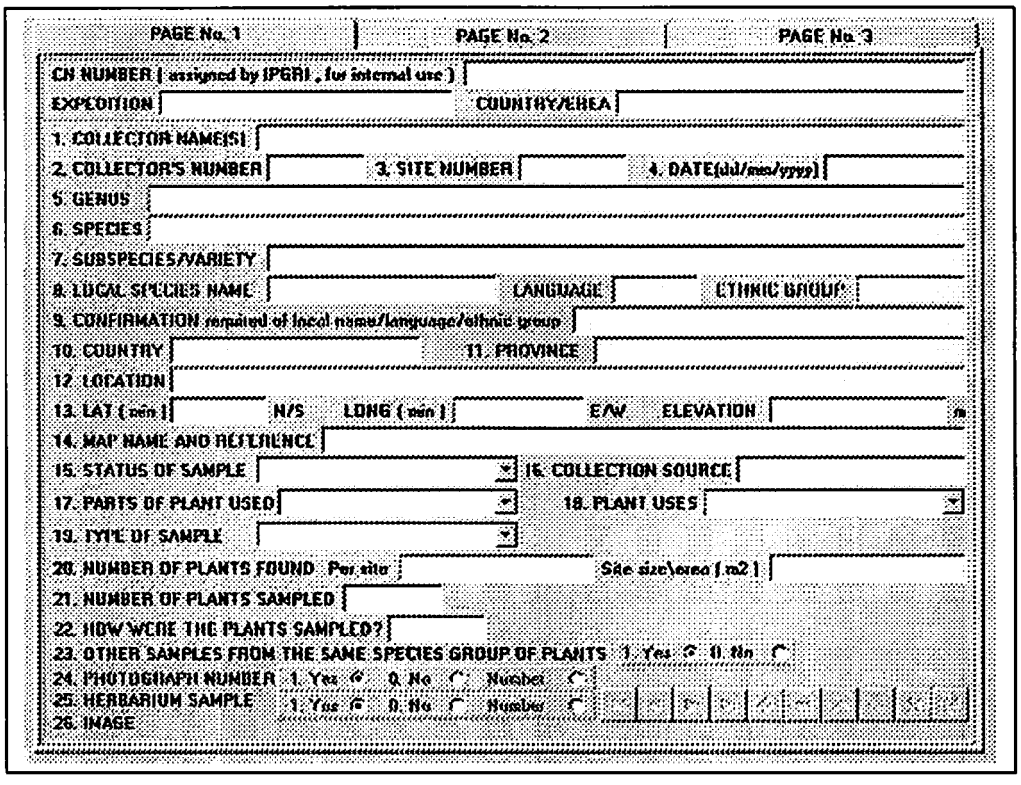
2. DATA COLLECTION

The most commonly practised method of collecting data on plant genetic resources is through survey where one basically encounters the data. It can provide a reasonably clear picture of the ecosystem under investigation. In this scheme, a surveyor travels during the appropriate seasons for searching plant species, through the accessible routes , stops and records the data , and proceeds further to gather more.

Survey is a way to get a picture on an ecosystem. This can be done by measuring parameters of its constituents such as soil, vegetation, species, micro-climate and socio-economic activities. This can be either by:

-Ground survey:	Using a field book and a pencil
	Precise detailed information on the habitat/soil type/bed rock/species' name
- Air	Using cameras
	Reconnaissance : distribution and nature of plant and animal communities
- Satellite	Remotely sensed imagery
	Image processing

A survey questionnaire form (exhibited in the following) is often used to record the primary (raw) data on identification descriptors, sample descriptors, site descriptors.



3 DATA MANAGEMENT

3.1 Database

Structurally a database is an organised collection of related information, or data, for efficient storage, update and retrieval of the data. A database management system (DBMS) refers to a software that is used to import, manage and analyse a database (*.DBF, *.DB, *.MDB, *.XLS). In a related database information is organised in tables. A table consists of rows and columns. A row contains information about an individual thing/place. A column contains specific item of information about records (rows). Once the raw data (survey data) has been collected it is then collated using commercial DBMS (Database Management System) software such as dBASE IV, dbase V, PARDADOX or ACCESS or using a tailored DBMS application developed specifically for plant genetic resources such as GMS. These raw data (*.DBF, *.DB, *.MDB, *.XLS) are usually compiled as a database of one or more tables linked by one or more common fields or identifiers.

3.2 DATA MANAGEMENT

To be able to generate information from the raw data (*.DBF, *.DB, *.MDB, .XLS..), the tables need to be prepared in the appropriate format (s) (*. TXT, *. CSV, *.DAT) that is required for statistical (such as GENSTAT, SAS, SYSTAT, STAGRAPHICS) and geographical information system (GIS) software.

Programmers use the word or appellation "table" in relational databases management system, documentation people use the word databases (one table or more tables) and statisticians use the word "matrix" (X). A matrix $X_{(i)}$ consists of (n x p) scores/measurements.

$$Xnp = \begin{bmatrix} x_{11} & & x_{1p} \\ & x_{ij} & \\ x_{n1} & & x_{np} \end{bmatrix}$$

Where x_{ij} is the measurement or score of the ith row (entity/individual) for the jth column (variable/attribute).

Data may consist of numerical, categorical and/or geographical information, depending on the score or measurement being: quantitative, such as counts (days to flowering, maturity); categorical, such presence/absence; geographical such as coordinates (LON/LAT).

Generation of various data matrices

Various data matrices may need to be generated from the raw data, for example we want to have from sample by descriptors table a species by site table (species-site table), where xij is the ith species' abundance at the jth site or agror-eclogical factors by site table (ecology-site table) where xij is the ith site and jth variable. This could be done by uisng statistical methods such as frequency tabulation.

Standardisation

The intensity of sampling may vary from site to site thus comparing the means between samples or sites may be misleading, thus the need to center the data by extracting the means of either row or column basis of data values. The data could also be standardised by either the range or the variance.

Data transformation/ "re-expression"

The raw data may need to be transformed or re-expressed as suggested by Tukey (199?) prior to its analysis. The common transformation is the (log(xij + c)). It is usually used when we have counts (quantitative data) to reduce the upper limit of a data set so that it is close to the lower

limit, in other words to reduce the importance of large values in comparison to small values. There is a problem however, if we have zero values, thus the use of a constant c.

Presence/absence matrices
These are usually developed from the raw data to find out similarity or dissimilarity in terms of species presence or absence sites. This could be a binary data.

Table data to map data (map data to a table data)
A table with geograghic data can be transformed to a map data. For this table/matrix in some cases should be in a format where the first columns are in the forms of (ID, X_i, Y_i) or and {ID (X_i,Y_i)}. GIS links adjacent point (X_i,Y_i) by drawing a straight line. When a layer is loaded automatic adjustments of scale and coordinate system takes place. This creates an ensemble of objects /layers on top of each other. The layer at the bottom is the base layer (map). This could be a vector (digitized) or raster (scanned) file.

Data from Cyprus

A database on wild wheat (*Aegilops* spp.) found in the island of Cyprus has been developed and used to create dot distribution maps (Della and Bari, 1993). The database consists of both herbarium data and germplsam data stored in the ARI (Agricultural Research Institute), Cyprus gene-bank. The herbarium data are from both the flora and the herbarium voucher specimens collected over many years by botanists in the island while the germplasm data are mainly on the germplasm expedition held during 1989. The study focuses on the germplasm data as it is the type of data that is available in the gene-banks. In this report, we used 1989 data from Cyprus. Various fields in data table is as follows:

Identification number
Taxonomic name
Location
Longitude
Latitude
Altitude
Rainfall

The map of Cyprus is given in the following.

[include Cyprus Map here]

4. DATA ANALYSES

4.1 Type of data and analysis approaches

Plant genetic resources (PGR) data are qualitative such as presence or absence of a species, and quantitative such as the number of species present at a site, bio-mass and root-mass of the plant sampled. Environmental variables such as annual rainfall or its distribution at the site of collection, and geo-reference such as longitude, latitude and altitude are often used to model the data. Depending on the objective of the study and the available data resources, appropriate statistical analysis methods are called for. These may include: developing a relationship between dependent variable (such as root-mass, abundance of a species) and independent variables (such as rainfall, temperature, geographical position) using a general or generalised linear model describing the abundance of the species by fitting an appropriate distribution; obtaining an estimate of number of species that might be expected in a region; evaluating and comparing diversity among the species by suitable diversity indices and test of homogeneity across regions; establishing an association between the abundance of a species with the site of collection using some data reduction techniques such as correspondence analysis; describing the spatial distribution of the species abundance using a spatial method and prediction at locations which were not accessed or were non-accessible. We shall present applications of some of these methods on Cyprus data.

4.2 Behaviour of species presence with groups of sites of collections

One might be interested in examining how presence/absence of a species varies with the collection sites. One approach would be to form the groups of the collection sites based on longitude range, latitude range and rainfall. Then the binary data (presence/absence) could be regressed on the factor representing the collection site groups. Thus with an objective to model presence/absence of each species with spatial groups of sites, we may prepare data file where a column may represent spatial group and a column for the presence (denoted by 1) and absence (denoted by 0) for the selected species.

Procedure: Since presence/absence is a binary variable, the regression procedures based on normal errors and linear regression function or identity link function) are not applicable. In this case, we need to use a generalised linear regression model. For the presence/absence data we

can use logit link function, and binomial distribution for error distribution. The analysis was carried out on the presence/absence data on each species. The procedure has been coded using Genstat 5 commands in the Appendix.

Firstly we form the groups using the histograms:

```
Histogram of longitudes of the collection sites:
              - 32.7  14 **************
         32.7 - 33.0  10 **********
         33.0 - 33.3  13 *************
         33.3 - 33.6   2 **
         33.6 - 33.9   4 ****
         33.9 - 34.2   8 ********
         34.2 -        0

Histogram of latitudes of the collection sites
             - 34.80  11 ***********
       34.80 - 34.92  13 *************
       34.92 - 35.04  11 ***********
       35.04 - 35.16   0
       35.16 - 35.28   2 **
       35.28 - 35.40  13 *************
       35.40 -         1 *

Histogram of annual rainfall at the collection sites:
              -  300   3 ***
          300 -  400  10 **********
          400 -  500  20 ********************
          500 -  600   6 ******
          600 -  700   4 ****
          700 -  800   5 *****
          800 -        3 ***
where, Scale:  1 asterisk represents 1 unit.
```

An examination of the above distribution, with merger of the nearby classes, resulted into the following groups of the sites.

Longitude groups:

Longitude(GrLon)	< 32.7	32.7 - 33.3	33.3 +
Number of cases	14	22	15

Latitude groups:

Latitude (GrLat)	< 34.8	34.8-35.04	>35.04
Number of cases	11	24	16

Rainfall groups:

Rainfall (GrRain)	<400	400-500	>500
Number of cases	13	20	18

The frequency of species classified in above groups are in the following. These will be used in the various sub-sections to follow.

Frequency distribution of species

Species	Longitude			Latitude			Rainfall		
	32.7	32.7-33.3	>33.3	34.8	34.8-35.04	>35.04	<400	400-500	>500
Bicornis	1	3	1	2	0	3	4	1	0
Biunciali	3	7	2	2	6	4	1	5	6
Comosa	0	0	1	0	0	1	0	0	1
Geniculat	2	6	4	3	5	4	2	2	1
Peregrina	3	2	5	2	5	3	5	5	1
Triuncial	5	4	2	2	8	1	1	1	9
Total	14	22	15	11	24	16	13	20	18

We used Genstat 5 to model the presence/absence data under the setting of Link=Logit; Distribution =Binomial. For species 2, we have the following results from regression analysis of deviance.

Longitude, lattitude and their interaction groups:

*** Summary of analysis ***

	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	6	7.50	1.250	1.25	0.277
Residual	44	48.15	1.094		
Total	50	55.65	1.113		
Change	-2	-4.88	2.441	2.44	0.087

*** Accumulated analysis of deviance ***

Change	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
+ GrLon	2	1.801	0.900	0.90	0.406
+ GrLat	2	0.818	0.409	0.41	0.664
+ GrLon.GrLat	2	4.882	2.441	2.44	0.087
Residual	44	48.150	1.094		
Total	50	55.651	1.113		

Rainfall groups

*** Summary of analysis ***

	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	2	3.19	1.596	1.60	0.203
Residual	48	52.46	1.093		
Total	50	55.65	1.113		
Change	-2	-3.19	1.596	1.60	0.203

*** Accumulated analysis of deviance ***

Change	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
+ GrRain	2	3.192	1.596	1.60	0.203
Residual	48	52.459	1.093		
Total	50	55.651	1.113		

The above results were generated for each of the four species (number) : *biunciali* (2), *geniculat* (4), *peregrina* (5), *triuncial* (6). Species *bicornis* (1) and *comosa* (3) were excluded from such an analysis due to very few number present. Probabilities of observing the deviance are summarised in the following to examine the variability between the groups.

Species	Rainfall DF=2	Longitude DF=2	Latitude DF=2	Lond x Latd DF=2	All groups DF=6
	Probability of deviance greater than observed				
2 *biunciali*	0.203	0.406	0.664	0.087	0.277
4 *geniculat*	0.010	0.610	0.899	0.251	0.681
5 peregrina	0.067	0.180	0.842	0.007	0.031
6 *triuncial*	0.001	0.318	0.205	0.261	0.227

(DF = degrees of freedom)

The above table shows statistically significant evidence (at 5% level) that the presence of species, *geniculat* (4) and *triuncial* (6) is associated with the rainfall zones, and of species *peregrina* (5) with the interaction of (or group formed by) longitude and latitude. At 10% level of significance, species 4 also appears associated with rainfall. The presence of species 2 does not show any variation with any of the above classification factors.

4.3 Distributional behaviour of abundance of species

Abundance of species is often described by a relationship between rank and frequency of the species. For example, for the data from Cyprus, we have

Species	Rank	Frequency/abundance
Ae. biunciali	1	12
Ae. geniculat	2	12
Ae. triuncial	3	11
Ae. peregrina	4	10
Ae. bicornis	5	5
Ae. comosa	6	1
(Total		51)

Such frequencies can be modelled by fitting a geometric frequency distribution function (May, 1975):

$$n_i = N\, C_k\, k\, (1-k)^{i-1}, \quad i = 1, 2, \ldots s$$

where k = the unknown parameter representing the proportion of available niche space or resource that each species occupies

n_i = the number of individuals in the ith species

N = the total number of individuals

s = number of species

$C_k = [1 - (1-k)^s]^{i-1}$

The distribution was fitted using Genstat 5 codes (Appendix) to the whole collection. The estimate of the parameter k was found as 0.1966 with standard error of 0.0213. The expected frequencies were as follows:

		Frequencies	
Species	**Rank**	**observed**	**expected**
Ae. biunciali	1	12	13.7
Ae. geniculat	2	12	11.0
Ae. triuncial	3	11	8.9
Ae. peregrina	4	10	7.1
Ae. bicornis	5	5	5.7
Ae. comosa	6	1	4.6

The chi-square statistic for the goodness of fit test was found as 4.89 on 4 degrees of freedom indicating a satisfactory fit to the observed frequencies using geometric distribution.

The above distribution was also fitted to the rank/abundance frequency classified by the longitude groups, latitude groups and rainfall groups. For each of these groups, observed frequencies fitted geometric distribution satisfactorily.

4.4 Measurement of diversity and conservation

Diversity (also rarity) is widely used (as criteria) to judge the suitability of a habitat for conservation" (Magurran 1988). The diversity,as defined by the ecologists, consists of two components: 1) variety [of entities] (richness in terms of entities) in terms of alleles/genes/varieties/populations/species/genus, and 2) relative abundance [of entities] The diversity can be measured by combining these two components togherther. Diversity has been studies have been conducted in specific regions formed along a transect or habitats (Magurran, 1998). The most popular indices to measure divesity are:

Margalef's diversity index (Clifford and Stephenson, 1975)

$$D_{Mg} = (S\text{-}1)/\ln N$$

Menhinick'index (Whittaker 1977)

$$D_{Mn} = S/\sqrt{N}$$

Where S *is* the number of species (groups/clusters) recorded, and N the total number of individual (samples) summed over all the species (groups/clusters).

4.4.1 Richness indices

Number of variants (species) in a sample. It could be a number of variants per area (density) as well as per number of individuals or biomass (numerical richness).

Expected number of species in a region

Based on an observed abundance data, one may be interested in estimating the species richness in terms of the number of species that may be expected from a given sample size. Across various regions of collection, sample sizes are not always equal. In such cases, rarefaction is a way to counter this problem. This can be done by using the rarefaction technique of Sanders and modified by Hurlbert (1971).

Expected number of species:

$$E(S) = \sum \left\{ 1 - \left[\binom{N - Ni}{n} / \binom{N}{n} \right] \right\}$$

Where,

E(S) is the expected number of species

n, a standaridised sample size (number of individual in the smallest sample)

N, the total number of individula recorded

*N*i, the number of individual in the ith species

In case of above data from the whole of Cyprus, the computed values of species expected from various sample sizes were computed as:

Standard size	Number of species
15	5.09
20	5.31
25	5.46

The number of species were also estimated using the groups of site collection classified according to longitude and latitude.

		Number of species	
Standards Size chosen	Groups	Expected	Upper limit
	Longitude		
14	<32.7	4.91	6.07 (=4.91+2*0.58)
	32.7-33.3	4.83	5.99
	>33.3	5.85	7.03
	SD	0.580	
	Latitude		
11	<34.8	4.92	6.36
	34.8-35.04	3.93	6.28
	>35.04	5.33	7.32
	SD	0.724	

In above, SD stands for standard deviation of the estimate of expected number of species and the upper limit is estimate plus two times the standard deviation. It may be noted that this upper limit has not been computed from the variance formula for the number of species. In the present case, we assumed that the variance would be by and large homogeneous across the three groups, therefore the three estimates were used to compute their SD for evaluating the upper limit.

Abundance models

Not all the variants (species) have the same degree of abundance. In general, a few variants would be very common, some commonly encounterd, while most would be represented by only a few records (accessions)

4.4.2 Shannon and Simpson indices

Shannon and Simpson indices are heterogeneity measures, they include both the richness and evens in one single figure:

Shannon:

$$H' = - \Sigma p_i \times ln\ (p_i),$$

Simpson:

$$D = \_ (p_i)^2$$

p_i is estimated as n_i/N

where n_i is the species records (specimens, records from the flora, germplasm data). N is the total number of individual records of all the species.

The Shannon index has been corrected for bias and is given as

$$H1 = - \Sigma p_i \times ln\ (p_i) - (S-1)/N + (1 - \Sigma p_i^{-1})/(12N^2) + \Sigma(p_i^{-1} - p_i^{-2})/(12N^3)$$

With variance, $Var(H1) = (\Sigma p_i \times (ln\ (p_i))^2 - \Sigma (p_i \times ln\ (p_i))^2)/N - (S-1)/(2N^2)$

(see Hutcheson 1970; Bowman et al 1971).

When applied to all the 51 samples from Cyprus, we find the following estimates of D, H' and H1:

D=0.1898 , H'=1.636, H1=1.534 and SE(H1)=0.0653

SE stands for standard error of the variable in the parenthesis.

We further pursued evaluation of the diversity with a view to see if diversity changes across the latitude , longitude and rainfall groups. The estimates are presented in the following table. Test the homogeneity of the H1 indices was carried out using a weighted analysis of variance where weights were inversely proportional to the estimated variance of H1. We got the following.

	Longitude			Latitude			Rainfall		
	<32.7	32.7-33.7	>33.3	<34.8	34.8-35.04	>35.04	<400	400-500	>500
No. of species	5	5	6	5	4	6	5	5	5
D	0.187	0.199	0.171	0.127	0.228	0.150	0.218	0.274	0.333
H	1.494	1.518	1.617	1.594	1.366	1.667	1.413	1.327	1.195
H1	1.187	1.330	1.252	1.207	1.239	1.328	1.073	1.109	0.944
SE(H1)	0.157	0.108	0.179	0.140	0.066	0.148	0.194	0.161	0.200

chisquare	0.5903	0.397	0.426
DF	2	2	2

Diversity should not be used to select sites for conservation independently from the habitats otherwise entity-poor habitats might never be considered such as those of harsh environment (Magurran, 1998). Diversity is of value if it includes habitats and if it takes into account the effects of area. For this it would be better to classify the area/entity under study into groups (forest into stands) and calculate the diversity within each group. Select representative examples of the whole range of habitats . Apply numerical classification (method of subdividing the environment) to select representative samples (Austin and Margules,1896). Use of passport data (ecological, geographical and ethno-botanical,) data as a background information for better conservation and utilization.

4.5 Spatial pattern of the species

A feature of species may be that their abundance is associate with specific location. Thus an objective may be to establish an inter-relationship between species and site of collection
The correspondence analyses could be used to study the preference of the species to prevail in certain sites in abundance relative to the other sites. Using SYSTAT, correspondence analysis was carried out on the data from Cyprus where the SITE represented the 32 sites of collections.. One species *comosa* was excluded from the analysis since it occurred at only one site. The results are:

```
Simple Correspondence Analysis

Chi-Square =      110.795.
Degrees of freedom = 116.
Probability =        0.619.

   Factor   Eigenvalue Percent Cum Pct
      1        0.773   34.87   34.87 -------------
      2        0.658   29.68   64.55 -----------
      3        0.526   23.73   88.28 ---------
      4        0.260   11.72  100.00 ----
```

Sum 2.216 (Total Inertia)

Row Variable Coordinates

Name	Mass	Quality	Inertia	Factor 1	Factor 2
bicornis	0.100	0.845	0.600	-1.734	1.438
biunciali	0.240	0.176	0.302	0.335	-0.330
geniculat	0.240	0.690	0.343	-0.049	-0.992
peregrina	0.200	0.303	0.433	-0.798	-0.139
triuncial	0.220	0.934	0.538	1.201	0.916

Row variable contributions to factors

Name	Factor 1	Factor 2
bicornis	0.389	0.314
biunciali	0.035	0.040
geniculat	0.001	0.359
peregrina	0.165	0.006
triuncial	0.411	0.281

Row variable squared correlations with factors

Name	Factor 1	Factor 2
bicornis	0.501	0.345
biunciali	0.089	0.087
geniculat	0.002	0.688
peregrina	0.294	0.009
triuncial	0.590	0.343

Column variable coordinates

Name	Mass	Quality	Inertia	Factor 1	Factor 2
1	0.020	0.474	0.063	-0.056	-1.224
2	0.060	0.828	0.029	-0.194	-0.601
3	0.040	0.432	0.102	-0.795	0.683
4	0.020	0.781	0.180	-1.972	1.773
6	0.020	0.213	0.080	-0.907	-0.171

.

30	0.060	0.803	0.026	0.564	-0.167
31	0.020	0.098	0.063	0.382	-0.408
32	0.040	0.759	0.047	0.874	0.361
33	0.020	0.886	0.071	1.366	1.129

Column variable contributions to factors

Name	Factor 1	Factor 2
1	0.000	0.046
2	0.003	0.033
3	0.033	0.028
4	0.101	0.096
6	0.021	0.001

.

.

30	0.025	0.003
31	0.004	0.005
32	0.040	0.008
33	0.048	0.039

Column variable squared correlations with factors

Name	Factor 1	Factor 2
1	0.001	0.473
2	0.078	0.750
3	0.249	0.183
4	0.432	0.349

.

.

30	0.738	0.065
31	0.046	0.052
32	0.648	0.111
33	0.527	0.360

The bi-plot which exhibits correspondence of species with sites are in the Figure 1.

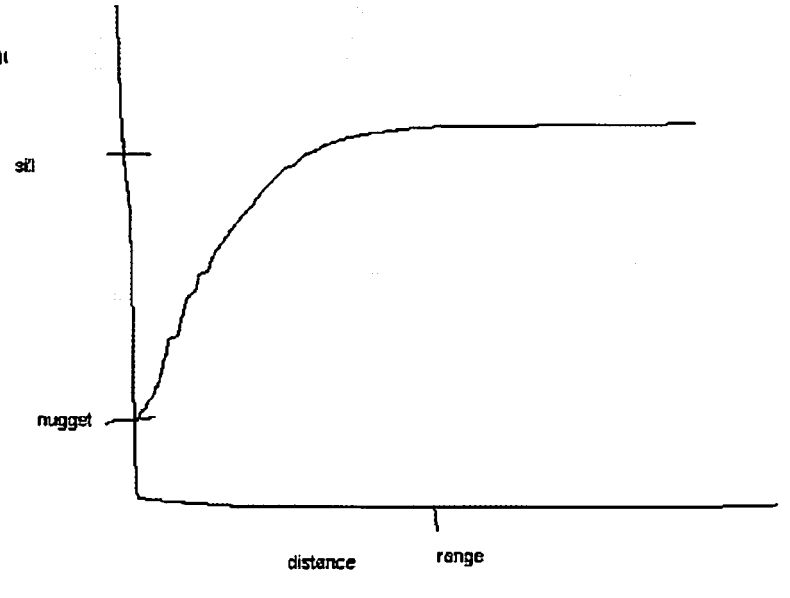
[insert Figure 1]

4.6 Spatial modeling of species abundance

Another aspects of spatial association of the abundance of species may be from a predictive view point. Here, one may be interested in developing models in terms of site co-ordinates (longitude, latitude) to predict the abundance at sites not in the collection and develop a contour over the areas of interest with regular grids. This can be attempted using spatial models (Cressie 1994). In the classical approaches, the variable being modelled are assumed independent and the predictions are made on mean irrespective of the location, or at the best a polynomial function could be fitted for accounting the variation in site. While in reality, the properties or values of a variable at nearby locations are more similar compared to locations which are far apart. The major assumption of spatial models is that the variable being modelled (Z(s)) should follow a spatial random process, where s represents the position coordinates (e.g., in x-, y-, and z-direction in a three dimensional system) . The main steps involved in spatial modelling include:

1. quantification of spatial variation using sample variogram (or semi-variogram). Method of moments or its robust estimates can be used.
2. Fitting a theoretical model of the variogram to the sample variogram and estimation of the variogram parameters. Commonly used variogram model are:
   - Spherical model
   - Gaussian model
   - Exponential model
   - Linear model
   - Hole model

A th



The main three characteristics of a variogram models are:

Nugget: measures the variation at a location. Ideally it should be zero but normally includes micro-scale variation and variation due to measurement errors.

Sill: measures the maximum variation between any two locations.

Range: the distance beyond which the variance between the two pints does not increase further. Thus the points within the range show spatial dependence and those outside are independent.

3. Once a sample variogram has been obtained, the parameters of the chosen variogram model can be estimated by various methods including maximum likelihood, least squares, generalised least squares.
4. *Spatial prediction*: Once a particular variogram model is found appropriate, it can be used to predict the variable at a given (not necessarily in the sample of observed locations) or on various grid points in a region. The spatial prediction is also named as "kriging" after D.G. Krige. A model for kriging the spatial process can be written as:

   $Z(s) = \sim(s) + \quad (s)$

- $\sim(s)$ represents the trend surface and (s) the spatial error. The model assumptions are in terms of stationarity of $\sim(s)$ and (s). The various forms of $\sim(s)$ give rise to prediction methods:
  - simple kriging:

    $$Z(s) = \_\_{i}\, Z(s_i) + [1 - \_\_{i}]\sim$$

  - ordinary kriging:

    $$Z(s) = \_\_{i}\, Z(s_i)$$

  - universal kriging:

    $$Z(s) = \_\hat{}_j f_j(s) + \quad (s)$$

where $f_j(s)$ are known functions to model trend values, $\hat{}_j$ are regression coefficients.

The abundance data (over all species) were sorted for the sites under collection. SYSTAT was used to model the variogram. Spherical form of the variogram with

nugget=0.2 , sill=0.6 and range=0.2

showed satisfactory fit to the observed variogram (Figure 2).

[insert Figure 2 here]

These values were then taken to model the abundance and the kriged values are shown in the contour Figure 3.

[insert Figure 3 here]

4.7ArEography

Rapoport (1982) used areography to study the configuration of geographical ranges of collections of species. This method helps to locate sites, transacts or areas that may yield new variants. The region under study need to be divided to equal areas (S_i), although these areas may not be uniform in terms of topography. Comparison on the basis of richness and abundance can be carried out between the areas using both the regression and cluster analysis.

5. SOFTWARE RESOURCES

The following are only few of the large resource base of various software.

Statistical packages

SAS
SPSS/PC
SYSTAT
BMDP
GENSTAT

GIS packages

MAP MAKER
IDRISI
ATLAS

ARCINFO (ESRI)
ARCVIEW GIS

Analysis of molecular data:

NTSYS pc (Numerical Taxonomy System): F.J. Rohlf; from Exeter , Software, 100 North Country Road, Setauket, NY 117@-S- SA (price: $155).

BIOSYS-1: D.L. Swofford, 1989; BIOSYS-1, a computer program for the analysis of allelic variation in population genetics and biochemical systematics, release 1.7. Illinois natural history Survey, Urbana, IL, USA.

RAPDistance version 1.03: J.A. Armstrong, R. Gibbs, R. Peakall, G. Weiller, 1995. RAPDistance, Package Manual. Australian National University, Canberra, Australia.
ftp://life.anu.edu.au/pub/molecular_biology/software/rapd103.zip

PHYLIP (Phylogeny Inference Package): J. Felsenstein; from the author, Department of Genetics, SK 50, University of Washington, Seattle, WA 98195, USA (free).
emil: joe@genetics.washington.genetics.edu
Web site:http:// evolution.genetics.washington.edu.phylip.html

MEGA (Molecular Evolutionary Genetic Analysis): S. Kumar, K. Tamura, M. Nei, 1993; from Joyce White, Institute of Molecular Evolutionary Genetics, 328 Mueller Laboratory Pennsylvania State Uhniversity, University Park, PA 16802, USA (price: $50).
Email: imeg@psuvm.psu.edu

MALIGN: W. Wheeler, D. Gladstein; from the authors, Dept of Invertebrates, American Museum of Natural History, Central Park West, 79th Street, New York, NY 10024 5192, USA (price: $50).

CLUSTAL W: J.D. Thompson, D.G. Higgins, J.T. Gibson, 1994; CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.
email: gibson@embl-heidelberg, de.: DesHiggins@ebi.ac.uk Anonymous ftp: ftp.ebi.ac.uk; and: ftp.bioindian.edu

PAUP (Phylogenetic Analysis Using Parsimony); D. Swofford. laboratory of Molecular Systematics, Smithsonian Institution, Washin ton DC 20560, USA (price: $1 00).

MACCLADE (Analysis of Phyiogeny and Character Evolution): W.P. Maddison, D.R. Maddison; from Sinauer:Associates,108 North Main Street, Sunderland, MA 01375, USA (price: $100).

HENNIG86: J.S. Farris; from the author; Molekylaersystematiska laboratoriet, Naturhistoriska riksmuseet, S 104 05 Stockholm Sweden.
DADA and CLADOS: by K. Nixon, L. H. Bailey Hortorium Cornell University, Ithaca, NY 14853, USA.

RNA (Rapid Nucleotide Analysis) by J.S. Farris, from the author, Molekylaersystematiska laboratoriet, Naturhistoriska rikmuseet, S 104 05 Stockholm, Sweden (price: $50)

NONA by P. A. Goloboff, from the author, Fundacion Miguel Lillo, Miguel Lillo 251, 400 San Miguel de Tucuman, Argentinca (price: $50)

TREEVIEW by R.D. M. Page, Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow G12 8 QQ, Scotland, UK.
Mail:dpage@udcf.gla.ac.uk

WINAMOVA program version 1.55 provided by L. Excoffier.
email=excoffie@sc2a.unige.ch
ftp://acasunl.unige.ch/pub/comp/win/amova.amova 155.zip

RAPDistance version 1.03. J. Armstrong, A. Gibbs, R. Peakal, CWeiller. 1995; RAPDisatnce, Package Manual. Australian National University, Canberra, Australia.
ftp:??life.anu.edu.au/pub/molecular_biology/software/rapdl03.zip

GDA (Genetic Data Analysis), 1996. P.O. lewis and D. Zaytin- 1996. Genetic Data Analysis: software for the analysis of discrete genetic data. Sinauer Assoc.. Sunderland, MA 01375. USA

FSTAT J. Goudet. 1995. Fstat version 1.2. A computer program to calculate F-statistics. J. Hered. 86:485-486.

RAPDFst B.L. Apostol, W.C. Black, P. Reiter, B.R. Miller. 1996. Population genetics with RAPD-PCR makers: the breeding structure of *Aedes aegypti* in Puerto Rico. Heredity 76: 325-334. Anonymous ftp: lamar.costate.edu in directory/pub/wcb4

MICROSAT and DELTAMU D.B. Goldstein, A Ruiz-Linares, M. Feldman, L.L. Cavaili.Sforza. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Nat. Acad. Sci. USA 92:6720-6727.
Email: minch@lotka.stanford.edu
Http:// lotka.stanford.edu/distance.html

GENEPOP version 1.2 M. Raymond, F. Rousset. 1995. GENEPOP (VF. 1.2): A population genetics softeare for exact tests and ecumenicism. J. Hered. 86:248-249.

REFERENCES

Barnett Vic and K. Feridun Turkman (editors) (1993).*Statistics for the Environment*. John Wiley & Sons Ltd. West Sussex, UK.

Burgess, T.M. and Webester, R. (1980). *Optimal interpolation and isarithmic mapping of soil properties. 1: The semi-variogram & ponctual kriging*: Journal of Soil science, 31, 315-331.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc. USA.

Cressie, N. (1990b). *The origins of Kriging*. Mathematical Geology, 22:239-252.

Kahn, Peter B. (1990) *Mathematical methods for Scientists and Engineers: Linear and Nonlinear Systems*. John Wiley & Sons, Inc. Singapore.

Magurran, Anne E. (1988). *Ecological Diversity and Its Measurement*. Groom Helm Australia, New South Wales, Australia.

Matheron. G. (1965). *La theorie des variables regionalisees et ses applications*. Masson, Paris.

Maurer, Brian A., (1994) *Geographical Population Analysis: Tools for the Analysis of Biodiversity*. Blackwell Scientific Publications, Alden Press, Oxford, UK.

Appendix

These contain some necessary codes of Genstat 5 commands.

1. *Creating histograms and tabulation of the abundances*

```
" Data from CYGER89.gsh "

Hist X_LON: & Y_LAT :& RAINFALL

Groups[NGroup=3] X_LON; GrLon; limi=!(32.7,33.3)
Groups[NGroup=3] Y_LAT; GrLat; limi=!(34.8,35.04)
Groups[NGroup=3] RAINFALL; GrRain; limi=!(400,500)

Tabu[Class=SP, GrLon, GrRain; prin=nobs]RAINFALL
Tabu[Class=SP, GrLat, GrRain; prin=nobs]RAINFALL

Tabu[Class=SP; prin=Nob] RAINFALL
Tabu[Class=SUB_REG; prin=Nobs] RAINFALL

Tabu[Class=SP, GrRain; prin=nobs]RAINFALL
Tabu[Class=SP, GrLon; prin=nobs]RAINFALL
Tabu[Class=SP, GrLat; prin=nobs]RAINFALL

Tabu[Class=SP, GrLon, GrLat; prin=nobs]RAINFALL

Tabu[Class=GrLon; prin=Nobs]RAINFALL
Tabu[Class=GrLat; prin=Nobs]RAINFALL
Tabu[Class=GrRain; prin=Nobs]RAINFALL
```

2. Modelling presence/absence in terms of site groups (binomial distribution and logit link)

```
Unit[51]
Vari NTot, NResp
Calc NTot=1

For i=2,4,5,6
Calc NResp=(SP==i)
Model[Link=logit; dist=bino] NResp; NBin=NTot
Terms GrLon*GrLat
Fit[prin=*]
Add[pri=*]  GrLon
Add[prin=*]GrLat
Add[Prin=m,s,a; fprob=y] GrLon.GrLat
Endf

For i=2,4,5,6
Calc NResp=(SP==i)
Model[Link=logit; dist=bino] NResp; NBin=NTot
Terms GrRain
Fit[prin=*]
Add[pri=m,s,a;fpro=y] GrRain

Endf
```

```
Stop
```

3. Expected Number of species

```
Scal S,N,NI,  hU, HL, ExpNSp , hh
Vari [valu=9,3,0,4,2,1,1,0,1,0,1,1] NS  " from the book"

Prin ' Longitude-wise'
Scal n; 14
For NS= !(1,3,2,3,5), !(3,7,6,2,4), !(1,2,1,4,5,2)

Prin ' Latitude-wise
Scal n; 11
For NS= !(2,2,3,2,2) , !(6,5,5,8) , !(3,4,1,4,3,1)

" all species  and three values of sample sizes"

Vari[values=5,12,1,12, 10, 11] NS
For n=15, 20, 25

Calc S=Nobs(NS*(NS.ne.0) ) : Calc N=Sum(NS)
Scal Ratio; 0
For i=1...#S
Calc NI=NS$[i]
Calc hU=N-NI : Calc hL=N-NI - n+1
Vari[Valu=#hL...#hU] V1
Calc hU=N : Calc hL=N - n+1
Vari[Valu=#hL...#hU] V2
Calc hh=Exp(Sum( log(V1)-log(V2) ) )
Calc Ratio=Ratio+hh
Endf
Calc ExpNSp=S-Ratio
Prin [orie=a] NS ; deci=0
Prin n, Ratio, ExpNSp
Endf
Stop

" To compute the upper limit of the number of species"
Vari [nval=3] x, y; values= !(4.912,4.832,5.874), !(4.921, 3.925,5.334)
Scal SD
Vari[nval=3]Upp
Calc SD=Sqrt(var(x)) : Calc Upp=x+2*SD : Prin SD, Upp
Calc SD=Sqrt(var(y)) : Calc Upp=x+2*SD : Prin SD, Upp
Stop
```

4. Fitting geometric distribution to rank abundance data

```
" whole of Cyprus"
Scal S; 6
Vari[Nval=S] Ns ; !(12,12,11,10,5,1)
" rainfall-wise "
Scal S; 5
Vari[nvalues=S] NSp[1...3]; !(5, 4,2, 1,1), !(9,5,4,1,1), !(9,6,1,1,1)
" Longitude wise"
Vari NSp[1...3]; !(1,3,2,3,5), !(3,7,6,2,4), !(1,2,1,4,5,2)
"Latitude -wise"
```

```
Vari NSp[1...3] ; !(2,2,3,2,2), !(6,5,5,8), !(3,4,1,4,3,1)

For Ns=NSp[1...3]
Scal S : Calc S=Nobs(Ns)
Sort[Dire=d] Ns
Vari[valu=1...S] Rank
Scal N: Calc N=Sum(Ns)
Expr  DeviSS; valu=!e(  Devi=Sum( (Ns-N*k* ((1-k)**(Rank-1) ) /(1-(1-k)**S)
)**2)   )
Model [Function=Devi]
RCycle k ; init=.5
FitN[Prin=s,e,m; Calc=DeviSS]
Vari[Nval=S] Fitted
Calc Fitted=N*k* ((1-k)**(Rank-1) ) /(1-(1-k)**S)
Scal Chisq, ChiDf, PrChiGt  : Calc Chisq=Sum(Ns*Ns/Fitted ) -N
Calc ChiDf=S-2 : Calc PrChiGt=1-Chisq(Chisq;ChiDf)

Prin Chisq, ChiDf, PrChiGt   :  Prin Rank, Ns, Fitted
Endf
Stop
```

5. Diversity indices

```
" Longitude wise"

Vari[nvalues=5] NSp[1...2]; valu= !(1,3,2,3,5), !(3,7,6,2,4)
Vari[nvalues=6] NSp[3]; !(1,2,1,4,5,2)

Scal D, H0, H1, N, SEH1
Vari[Nval=3] VecH, VecSEH, Wet

For ss=NSp[1...3];i=1...3; S=5, 5, 6  " S no. of species "
Calc N= Sum(ss)
Calc D=Sum(  ss*(ss-1) / N /(N-1)  )
Calc H0= -Sum( (ss/N)*log(ss/N)  )
Calc H1= H0-(S-1)/N + (1-Sum(N/ss) )/(12*N*N) +\
     (Sum(N/ss-(N/ss)**2))/(12*N**3)
Calc SEH1= ( Sum( (ss/N)* (log(ss/N))**2  ) - H0*H0 )/N  - (S-1)/(2*N*N)
Calc SEH1=sqrt(SEH1)
Prin D, H0, H1, SEH1, S
Calc VecH$[i]=H1 : Calc VecSEH$[i]=SEH1
Endf
Calc Wet=1/VecSEH**2
Anov[Weigh=Wet; prin=a,m] VecH

" Latitude-wise"

Vari[nvalues=5] NSp[1]; !(2,2,3,2,2)
Vari[nvalues=4] NSp[2]; !(6,5,5,8)
Vari[nvalues=6] NSp[3];  !(3,4,1,4,3,1)

" Rainfall-wise"
Vari[nvalues=5] NSp[1...3]; !(4,1,2,5,1), !(1,5,9,4,1), !(6,1,1,1,9)

" whole Cyprus"
Vari[nvalues=6] NSp[1]; !(5, 12, 1, 12,10,11)
```

Stop

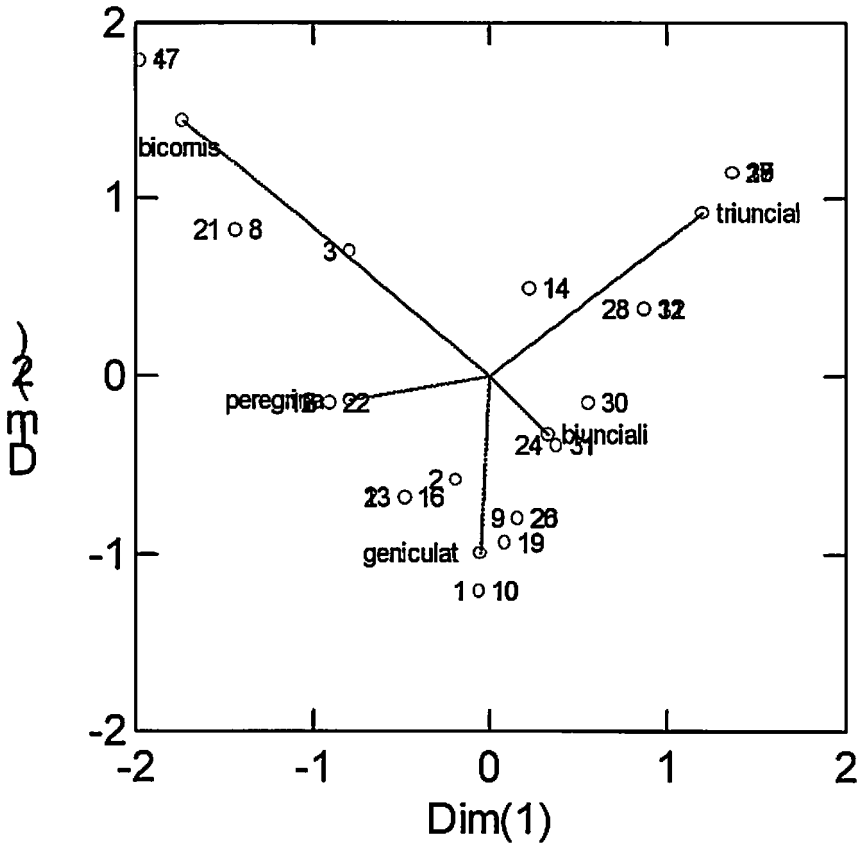


Figure 1. Biplot of species and site of collections on abundance of species

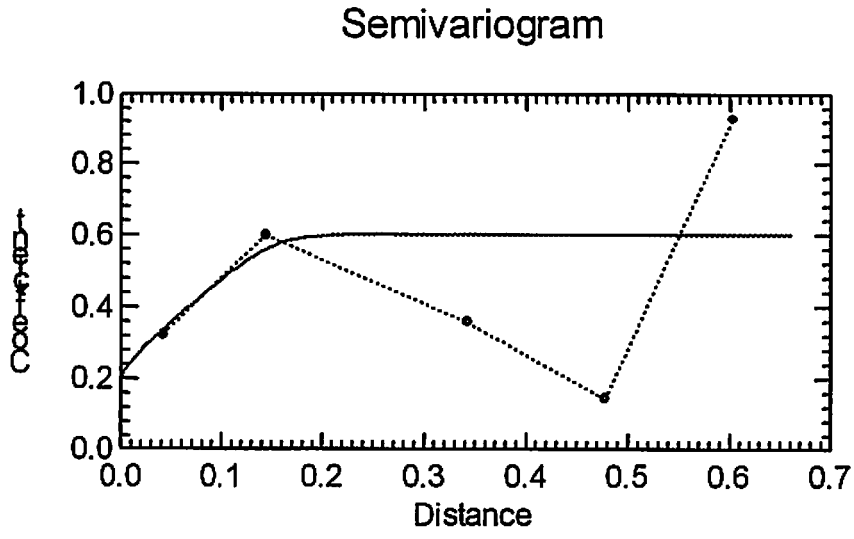


Figure 2. Semi-variogram of abundance of species

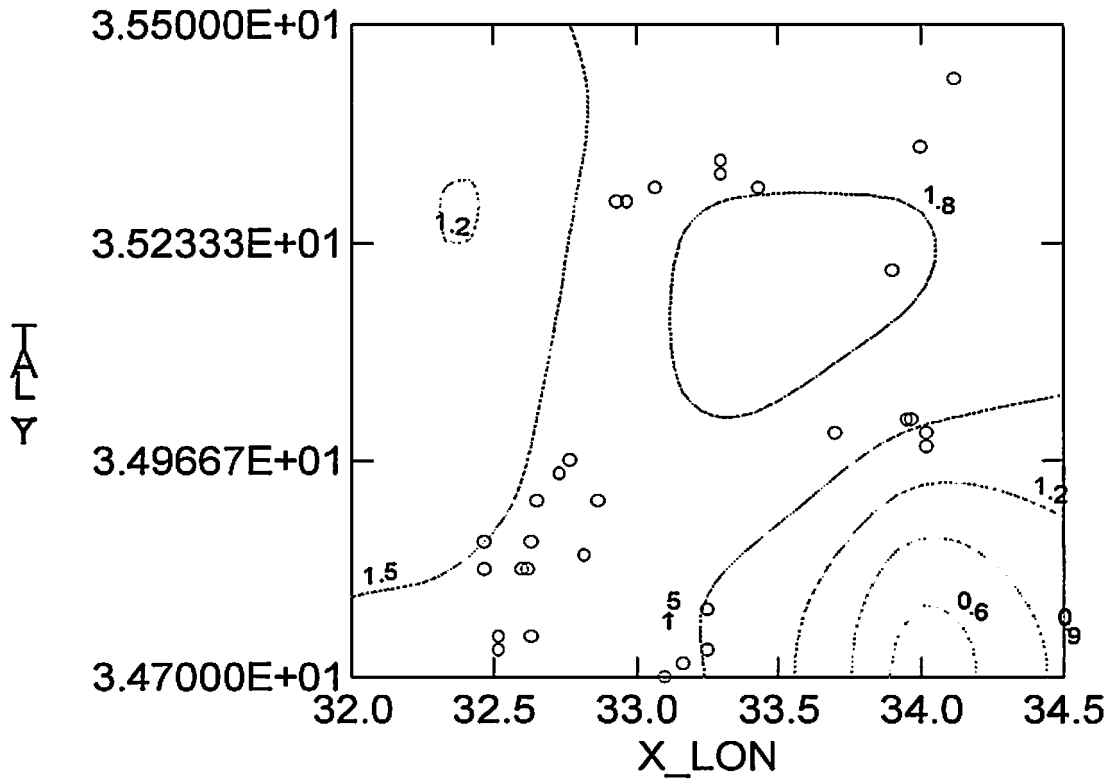


Figure 3. Kriged estimates with spherical variogram (Nugget=0.2, sill=nugget+0.4, range=0.2)