Z. Morph. Anthrop.	81	1	33–39	Stuttgart, Dezember 1995

Anthropometry & Human Genetics Unit
Indian Statistical Institute Calcutta

Spatial autocorrelation analysis reveals that A, B and O allele frequency surfaces on the Indian subcontinent are highly fractured

By **B. N. Mukherjee** and **Partha P. Majumder**

With 3 figures and 3 tables in the text

Abstract: Spatial autocorrelation analysis performed on published data pertaining to caste and tribal populations of the Indian subcontinent has revealed that the surfaces of A, B and O allele frequencies are highly fractured. The only significant spatial autocorrelation was observed in respect of the A allele frequency among caste populations.

Zusammenfassung: Räumliche Autokorrelations-Analysen an veröffentlichten, nach Kasten und Stämmen differenzierten Daten aus Indien zeigen, daß die räumliche Verteilung der A, B und O Allele starke Unregelmäßigkeiten aufweisen. Nur bei Kastenbevölkerungen konnte für das Allel A eine signifikante räumliche Autokorrelation festgestellt werden.

Introduction

During the last few decades a large number of studies pertaining to genetic diversity has been carried out among population groups residing on the Indian subcontinent. Of the various polymorphic marker systems studied, the ABO blood group system has been the most extensively studied system. To investigate patterns of variation of A, B and O allele frequencies among populations of the Indian subcontinent, some statistical studies have also been conducted (MAJUMDER & ROY 1982a, b; WALTER et al. 1991). These studies have revealed some interesting patterns in respect of variation among populations arranged in order of social hierarchy, but no major geographical patterns were discernible. In fact, MAJUMDER & ROY (1982b) found that while a broad geographical partitioning of the Indian subcontinent explained a significant proportion of the overall variation in A, B and O allele frequencies, a further consideration of latitudinal and longitudinal variation within the broad geographical zones did not further explain the variation in the allele frequencies to any significant extent. Apart from this particular study, to the best of our knowledge, no other study has been conducted in which the allele frequency data were subjected to rigorous statistical analysis with the objective of discovering spatial patterns. Of course, a cursory visual inspection of the data does not indicate that there are any clear spatial patterns. This has resulted in the notion that the surfaces of A, B and O allele frequencies over geographical space on the Indian subcontinent are highly fractured. The purpose of the present study was to test this view by a thorough statistical examination of the available data, and to

quantify the extent of fracture of the allele frequency surfaces. This has been done by performing a spatial autocorrelation analysis, the objective of which is to discover whether the observed value of a variable (in the present case, A or B or O allele frequency) in one geographical location is dependent on the values at neighbouring locations. If such a dependence exists, the variable is said to exhibit spatial autocorrelation. The use of spatial autocorrelation analysis has been popularised in anthropology by ROBERT SOKAL and his colleagues (see, for example, SOKAL & FRIEDLAENDER 1982a), and many useful conclusions have been derived by this group (see, for example, SOKAL & MENOZZI 1982b) using this statistical technique.

Materials and methods

The data analysed in the present study have been drawn from published sources. All data published on Indian populations until 1993 were compiled. Attempts were made to make this compilation exhaustive. We do not, however, claim that no published data sets have been left out. Our compilation included names of populations, place(s) of sampling, and A, B, AB and O phenotype frequencies. The classification of the populations into socio-religious categories (caste: high, middle, low; tribe; religious categories: Muslim; Buddhist; etc.) was done by use of accepted lists that are available and in consultation with other anthropologists. The latitudes and longitudes of places of sampling were also found by use of standard geographical tables. Since no uniform statistical procedure for computing A, B and O allele frequencies was followed in the publications, we recomputed allele frequencies of all compiled data sets using the maximum likelihood procedure. In the present study, only data sets on anthropologically 'well-defined' populations have been included. Thus, data sets pertaining to ill-defined populations such as 'Bengalis', 'South Indians', etc. have been excluded. Further, data sets for which locations of sampling were not available, or data sets for which significant departures from Hardy-Weinberg equilibrium were observed were excluded. A particular data set pertaining to the Shompens of Andaman and Nicobar Islands was also excluded, because all sampled individuals in this study were reported to be of the O blood group (estimated O allele frequency = 1).

For the purpose of the present analysis, hierarchical classification of populations belonging to the Hindu caste system has not been used; all caste populations have been treated as a single group. This was done to avoid vagaries of small sample sizes. The tribal populations have been treated as another separate category. Data on religious groups, such as Muslims, Christians, Parsees, etc., have not been included in the present analysis. The number of data sets on such groups was too small to perform a meaningful analysis. In all, the number of data sets pertaining to caste populations included in this study was 491, the number pertaining to tribal populations was 399; the total number of data sets in the pooled category (caste + tribal) was, therefore, 830.

For the purpose of spatial autocorrelation analysis, the Indian subcontinent was enclosed in a geographical grid of 1200 (= 40×30) cells. Each cell in this grid corresponded to a latitudinal distance of 1° and a longitudinal distance of 1°. Each population included in the data base was classified into one of these 1200 cells based on the latitude and longitude of the place of sampling of this population. The allele frequency estimate (A or B or O) of this population was then substituted as the value of the cell in the grid. When multiple populations were classified into the

same cell, the weighted average of the allele frequency estimates of these populations was computed and substituted as the value of the cell in the grid.

Spatial autocorrelation was estimated using the Moran's I statistic. This statistic is defined as follows: Suppose there are n locations (cells of the geographical grid, in the present case). Let p_i denote the allele frequency (of one of A, B or O alleles) in the i-th location. Let w_{ij} denote a weight (a positive real number) if locations i and j are 'connected'; w_{ij} is taken to be zero if the locations are 'unconnected'. Frequently, as in the present study, w_{ij} is assigned a value 1, if the locations i and j are connected. [In the present study, although all cells in the geographical grid were initially assumed to be connected, many cells eventually became unconnected because of missing data (empty cells). In the present study, spatial autocorrelations were computed using only those cells which were non-empty.] Let $z_i = p_i - \bar{p}$. Let $W = \Sigma^n \Sigma^n_{i \neq j=1} w_{ij}$. Then,

$$I = \frac{n}{W}\left[\sum_{i \neq}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j / \sum_{i=1}^{n} z_i^2\right].$$

The range of I is from -1 to $+1$. Test of significance of an observed value of I is performed by calculating the test statistic $Z = [I - E(I)]/\sqrt{V(I)}$, which follows a $N(0,1)$ distribution under the null hypothesis of no spatial autocorrelation, where the expectation [E(I)] and variance [V(I)] of I are obtained from n! I values each computed by randomly permuting the p_i values among the n localities (see CLIFF & ORD 1973 for further details). This randomization approach, albeit computer intensive, was adopted because it was felt that the alternative approach of computing the expectation and variance under the assumption that the p_i values are drawn from a Normal distribution was not suitable for the present data on allele frequencies. Spatial autocorrelations were computed using both data on adjacent cells (distance 1; latitudinal and longitudinal distance of $\approx 1°$) and on cells one apart (distance 2; latitudinal and longitudinal distance of $\approx 2°$). (Further computations at higher distance levels were not performed because of small number of non-empty cells.)

Results and discussion

The first point that was noticed after classifying populations into cells of the geographical grid was that although the grid contained 1200 cells, most cells were empty. For the data base pertaining to the caste populations, there were only 77 non-empty cells. For the data base pertaining to the tribal populations, there were only 100 non-empty cells. And, in the pooled data base, there were 143 non-empty cells. Thus, it was discovered that although the number of populations covered in the genetic polymorphism studies in India is very large, the populations have been largely sampled from a restricted number of locations. Put another way, the geographical spread of the locations of sampling of the populations is not very large. The major reason for this is perhaps that for convenience of setting up field laboratories and/or shipment of blood samples to the base laboratory, most investigators have sampled from locations that are close to cities or large towns. This feature of the data, however, is a handicap to performing a detailed geographical analysis. The results of the present study should, therefore, be interpreted bearing this fact in mind.

Before performing a spatial autocorrelation analysis, we first prepared histograms of A, B and O allele frequencies across geographical space. These are

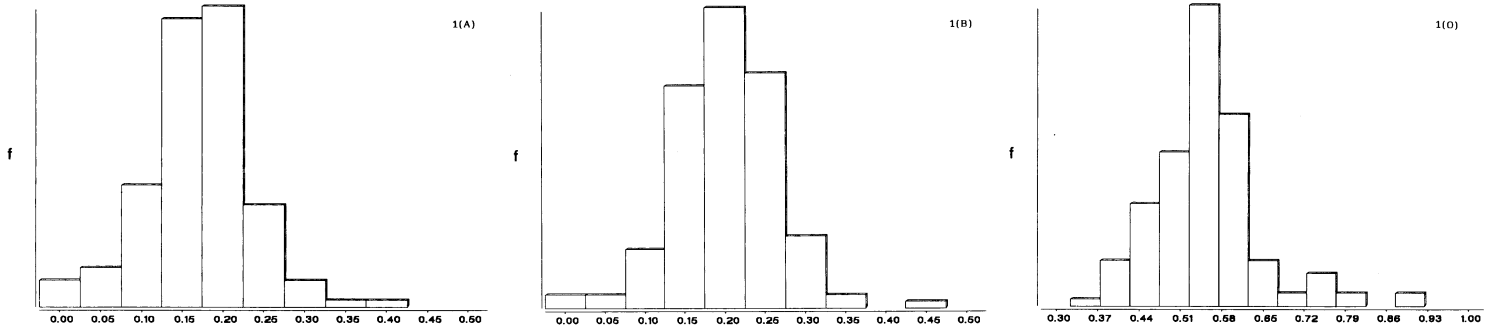

Fig. 1. Histograms of A [1(A)], B [1(B)] and O [1(O)] allele frequencies on the Indian subcontinent.

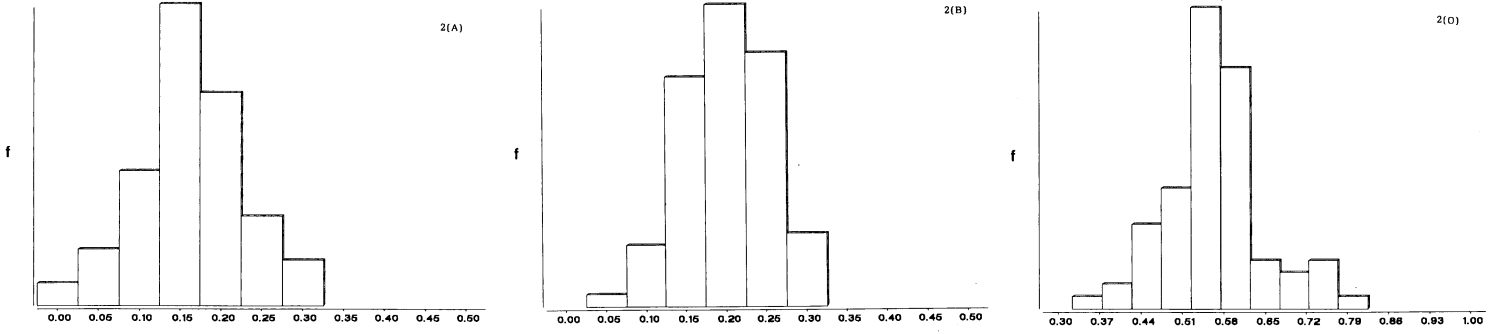

Fig. 2. Histograms of A [2(A)], B [2(B)] and O [2(O)] allele frequencies of caste populations on the Indian subcontinent.

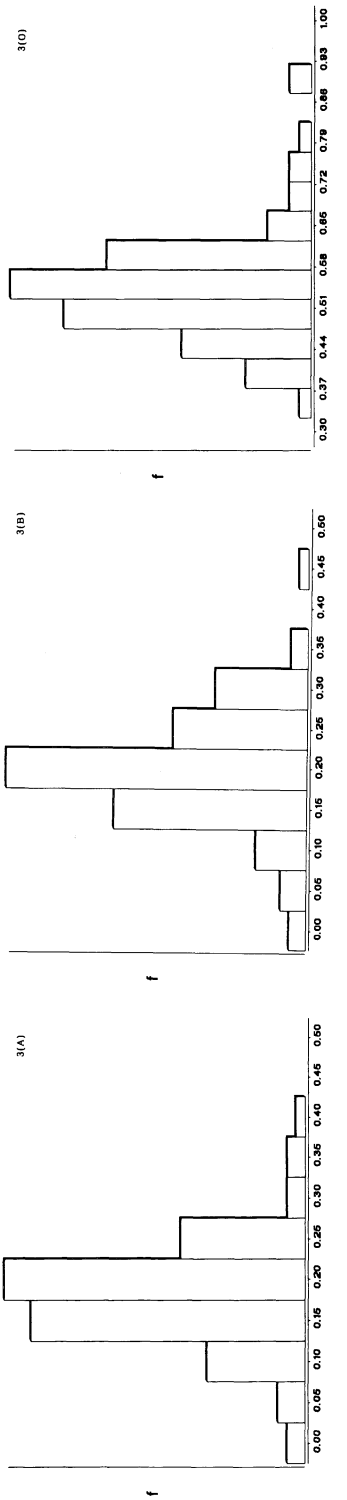

Fig. 3. Histograms of A [3(A)], B [3(B)] and O [3(O)] allele frequencies of tribal populations on the Indian subcontinent.

presented in Fig. 1 (A), (B) and (0) pertaining to the pooled data, in Fig. 2 (A), (B) and (0) pertaining to the caste populations, and in Fig. 3 (A), (B) and (0) pertaining to the tribal populations. It must be noted that in these figures the frequency (f) on the Y-axis refers to the numbers of cells in the geographical grid and not to the number of populations. It is observed from these figures that the ranges and/or frequency distributions of A, B and 0 allele frequencies are quite different between castes and tribes. However, the mean frequencies of B and 0 alleles are nearly the same between castes and tribes; the tribal populations have a higher frequency of the A allele compared to the caste populations (see second row of Tables 2 and 3).

Table 1. Results of spatial autocorrelation analysis of A, B and 0 allele frequencies in defined populations on the Indian subcontinent.

Item	A		B		0	
	Distance		Distance		Distance	
	1	2	1	2	1	2
No. of cells	143	35	143	35	143	35
Mean allele frequency	.194	.194	.228	.236	.581	.582
s.d. of allele frequency	.005	.009	.005	.013	.008	.018
Moran's I	.059	.513	.180	−.328	.117	−.002
Z	.570	2.083	1.625	−1.182	1.088	.110

Table 2. Results of spatial autocorrelation analysis of A, B and 0 allele frequencies in caste populations on the Indian subcontinent.

Item	A		B		0	
	Distance		Distance		Distance	
	1	2	1	2	1	2
No. of cells	77	18	77	18	77	18
Mean allele frequency	.186	.189	.227	.212	.587	.600
s.d. of allele frequency	.007	.012	.006	.015	.010	.020
Moran's I	.331	1.022	.235	−.466	.372	.013
Z	1.527	2.264	1.095	−.848	1.715	.150

Table 3. Results of spatial autocorrelation analysis of A, B and 0 allele frequencies in tribal populations on the Indian subcontinent.

Item	A		B		0	
	Distance		Distance		Distance	
	1	2	1	2	1	2
No. of cells	100	27	100	27	100	27
Mean allele frequency	.204	.205	.227	.228	.569	.567
s.d. of allele frequency	.007	.012	.007	.017	.010	.017
Moran's I	.114	.315	.106	−.387	.190	−.126
Z	.910	1.103	.847	−1.110	1.475	−.276

The results of spatial autocorrelation analysis are given in Tables 1, 2 and 3 pertaining to the pooled data, caste populations and tribal populations, respectively. These results indicate that except for the observed spatial autocorrelation for A allele frequency at distance level 2 among caste populations and resultantly also in the pooled data, none of the remaining autocorrelation values is significant ($Z < 1.96$). This indicates that the spatial surfaces of allele frequencies for the ABO blood group system on the Indian subcontinent are, in general, quite fractured.

The present results, which should be viewed as preliminary, indicate without any doubt the fractured nature of A, B and O allele frequency surfaces on the Indian subcontinent. However, it is interesting to note that the surfaces of the caste populations show higher spatial autocorrelations compared to those of the tribal populations. It is difficult to assign specific causes that have led to the fractured nature of the gene frequency surfaces. We suspect that the existence of local subgroups within larger population groups, and the lack of intermixture among local subgroups may be the major cause. We are now performing further analyses by forming a finer geographical grid and examining subsets of the data defined ecologically and linguistically.

Acknowledgements

We are grateful to Mr. B. N. PAL and Mr. R. N. DAS for help in data compilation and computations.

References

CLIFF, A. D. & ORD, J. K. (1973): Spatial Autocorrelation. – Pion Ltd., London.

MAJUMDER, P. P. & ROY, J. (1982a): Distribution of ABO blood groups on the Indian subcontinent: A cluster analytic approach. – Current Anthropology **23**: 539–566.

– (1982b): Distribution of ABO blood groups on the Indian subcontinent: A study of micro-geographical variation. – In: MALHOTRA, K. C. & BASU, A. (eds.): Human Genetics and Adaptation, vol. 1. – Statistical Publishing Society, Calcutta, pp. 175–186.

SOKAL, R. R. & FRIEDLAENDER, J. (1982): Spatial autocorrelation analysis of biological variation on Bougainville Island. – In: CRAWFORD, M. H. & MIELKE, J. H. (eds.): Current Developments in Anthropological Genetics, vol. 2. – Plenum Publ. Corp., New York, pp. 205–227.

SOKAL, R. R. & MENOZZI, P. (1982): Spatial autocorrelations of HLA frequencies support demic diffusion of early farmers. – Amer. Nat. **119**: 1–17.

WALTER, H., DANKER-HOPFE, H. & BHASIN, M. K. (1991): Anthropologie Indiens. – Gustav Fischer Verlag, Stuttgart.

Author's address:

PARTHA P. MAJUMDER, Anthropometry & Human Genetics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.