



Bayesian and multivariate analysis of buffaloes of Indo-Gangetic plains for revealing cryptic spatial patterns of genetic variability

UPASNA S¹, PRIYANKA BANERJEE², JYOTI JOSHI³ and R K VIJH⁴

National Bureau of Animal Genetic Resources, Karnal, Haryana 132001 India

Received: 8 October 2010; Accepted: 17 April 2011

ABSTRACT

The blood samples of 625 buffaloes collected from 34 sampling locations of Indo-Gangetic plains were collected and DNA isolated. The genotype data was generated on these animals using 11 microsatellite markers. The multilocus genotype data were analysed to decipher the genetic structures. The data analysis utilised the spatial coordinates to reveal the cryptic spatial patterns of genetic variability. The Bayesian as well as spatial multivariate analysis revealed 5 populations of buffaloes belonging to Indo-Gangetic plains. The F_{IS} and F_{ST} values were calculated for various populations or clusters formed. The 5 clusters formed using Geo-referenced multilocus genotypes are not in geographic contiguity but reveal cryptic spatial patterns.

Key words: Bayesian and multivariate analysis, Buffalo, Indo-Gangetic plains, Microsatellites

Several ecological and evolutionary processes influence genetic variation. The joint analysis of genetic and spatial information can lead to a better understanding of structuring of the populations.

Landscape genetics data, collected at a finer scale than that typical of phylogenetics, will help our understanding of micro-evolutionary processes that generate genetic structure across space. The two key steps of landscape genetics are the detection of genetic discontinuities and the correlation of these discontinuities with landscape. Indian buffaloes have been classified in 10 distinctive breeds (Sahai and Vijh 2000). There are also several populations of buffaloes which have been studied (Tantia *et al.* 2006, Vijh *et al.* 2008 and Mishra *et al.* 2009). However, there has been no defined breed of buffalo in the Indo-Gangetic plain of UP which have 23.4% of Indian buffalo population. No information on genetic structure of this large population exists.

In the present study, we attempted to identify the genetic discontinuities in the buffalo population of Ganga-Yamuna Basin using a Bayesian model implemented in a Markov chain Monte Carlo scheme that permitted inference of the location of such genetic discontinuities from individual georeferenced multi-locus genotypes, without *a priori* knowledge on population units and limits. In this method, the global set of sampled individuals was modeled as a spatial mixture of panmictic populations, and the spatial organization

of populations was modeled through Voronoi tessellation (Guillot *et al.* 2008). We also investigated cryptic spatial patterns of genetic variability using geo-referenced multilocus genotype using spatial principal component analysis (Jombart 2010).

MATERIALS AND METHODS

Blood samples of 625 buffaloes were collected from throughout the Indo-Gangetic plains. The districts from which there were very few samples were clubbed with the adjoining districts and a total of 34 conglomerates. The DNA was isolated following normal protocols (Sambrook *et al.* 2001) and 11 heterologous microsatellites were selected from cattle database. The microsatellite data was generated using fluorescently labeled primers and using ABI Automated DNA sequencer. The data generated was extracted using GeneMapper software v4.1.

We analyzed the data of 625 buffaloes belonging to 34 sampling locations (districts) of Indo Gangetic plains. Microsatellite data was generated on these buffaloes using 11 microsatellites. These animals were nearly evenly distributed over an area that corresponds to a landscape highly fragmented by two major rivers of Ganga, Yamuna and their tributaries. We analyzed the buffalo data set by processing 10 independent MCMC runs of spatial D -model as implemented in Geneland software (Guillot *et al.* 2008). We used priors on K - uniform between 1 and 10 and on $\bar{\epsilon}$ - uniform between 0 and 100. Each run included 200,000 iterations and a short burn-in period of 25000.

Present address: ¹ Research Associate (e mail: upasna30@gmail.com), ^{2, 3} Senior Research Fellow, ⁴ Principal Scientist (e mail: rameshkvijh@gmail.com).

The sPCA analysed a data matrix $\times$ which contained genotypes or populations in rows and alleles in columns. Spatial information was stored inside a spatial weighting matrix L which contained positive terms corresponding to some measurement of spatial proximity among entities. This matrix is row-standardized (i.e. each of its rows sums to one), and all its diagonal terms are zero. L can be used to compute the spatial auto-correlation of a given centred variable $\times$ (i.e. with mean zero) with n observations ($\times 2 Rn$) using Moran's I (Moran 1948, 1950, Cliff and Ord 1981):

$$I(X) = \frac{\sum \times^T L \times}{\sum \times^T \times}$$

In the present study of buffaloes belonging Indo-Gangetic plains the $\times$ contains frequencies of an allele. Moran's I was used to measure spatial structure among the values of $\times$: it is highly positive when values of $\times$ observed at neighboring sites tended to be similar (positive spatial auto-correlation, referred to as global structures), while it was strongly negative when values of $\times$ observed at neighboring sites tend to be dissimilar (negative spatial autocorrelation also called local structures).

The sPCA defined the following function to measure both spatial structure and variability in $\times$:

$$C(\times) = \text{var}(\times) I(\times) = \frac{1}{n} \sum \times^T L \times$$

$C(\times)$ is highly positive when $\times$ has a large variance and exhibits a global structure; conversely, it was largely negative when $\times$ had a high variance and displayed a local structure. This function was the criterion used in sPCA in present study, which finds linear combinations of the alleles of X (denoted $\phi = Xv$) decomposing C from its maximum to its minimum value. $C(Xv)$ was a product of variance and of auto-correlation. A structure with a low spatial autocorrelation can barely be interpreted as a spatial pattern; similarly, a structure with a low variance would likely not reflect any genetic structure. For sPCA, we utilised the software Adegenet 1.2–5 written in R (Jombart 2010).

RESULTS AND DISCUSSION

The posterior distribution gave a mode of 5 using the D-spatial model was then rerun on 50000 iterations with a fixed value of $K=5$. Maps for the posterior probability belonging to each population were then derived. The Geneland procedures are efficient methods to achieve genetic discontinuities within the landscape from the individual geo-referenced multilocus genotype data, without any a priori knowledge on the population units and limits (Guillot *et al.* 2005). The method as implemented in Geneland software is also well suited for revealing cryptic spatial genetic structure and also provides a suitable approach for the detection of migrants (individuals poorly genetically related to their spatial neighbors) and their assignment to the population of

origin. In the present study, we identified the five populations of Indo- Gangetic plains which in fact belong to spatial domains that are apparently unconnected but may belong to same population unit. The software provided a significant advantage in comparison to other methodologies that aimed to define the populations. The F model utilised in the present study gave a rather loose definition to what a population is. The 5 populations that were identified using the Geneland algorithm are depicted as Fig. 1. (a-e).

The five populations deciphered by Geneland (Fig. 2) belong to the districts of Mau, Balia and Ghazipur (Cluster 3), Fatehpur, RaeBareilly, Chandauli, Partapgarh (Cluster 2), Azamgarh, Gonda, Balrampur, Sultanpur, Gorakhpur and Sant Kabirnagar (Cluster 4), Baghpat, Meerut, Ghaziabad, Etawah, Auraiya, Jalaun, Luknow, Sultanpur, Partapgarh, Faizabad, siddharthnagar, Mirzapur (Cluster 5) and balance in Cluster 1.

The estimated F_{IS} values for the 5 clusters thus formed in the analysis were 0.3054, 0.1795, 0.0440, 0.2635 and 0.2533 respectively. Thus the population (Cluster 3) has least inbreeding coefficient. The F_{ST} values amongst the five clusters estimated are shown in the Table 1.

Jombart *et al.* (2008) proposed a new method sPCA as a tool to investigate cryptic spatial patterns of genetic variability using georeferenced multilocus genotype. The method relies on a modification of PCA so that not only the variance between the studied entities (individuals or populations), but also their spatial autocorrelation were taken into account. The main results of the analysis are maps of entities scores allowing a visual assessment of the spatial genetic structures. The obtained scores reveal 2 types of patterns, which were defined as global and local structures (Thioulouse *et al.* 1995). A global pattern would differentiate between two spatial groups or find a cline, whereas local scores would retrieve stronger genetic differences among neighbours than among random pairs of entities. As the studied entities can be genotypes or groups of genotypes (referred to as 'populations', in a broad sense), global and local structures encompass a wide range of biological situations. Global patterns of genotypes indicated population patches in an island model, as well as cline wherever isolation by distance occurred as in the case of present study of buffaloes of Indo- Gangetic plains. Local structures could arise when individuals from the same genetic pool mate in the population compared to the individuals from other genetic pools.

The procedure as implemented in software sPCA did not decompose the total variance into decreasing additive components rather the product of the variance $\text{var}(w)$ and the spatial auto-correlation $I(w)$ is separated into positive, null and negative components. It means that even if the variance is always positive, the spatial auto-correlation can be positive as well as negative. Hence, while PCA focuses on the scores associated to the highest eigen values, sPCA

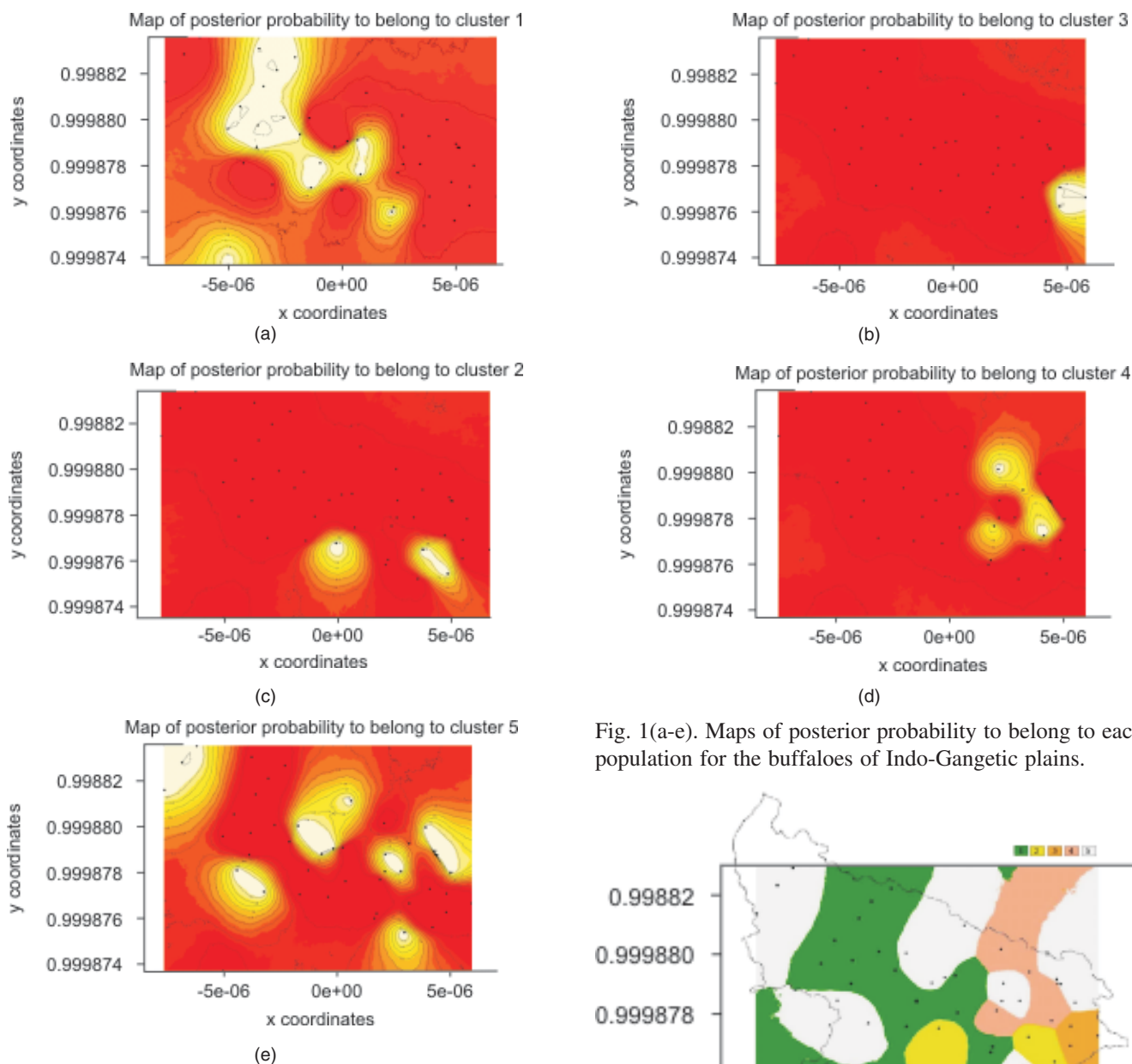


Fig. 1(a-e). Maps of posterior probability to belong to each population for the buffaloes of Indo-Gangetic plains.

encompasses two types of informative scores, both reflecting an aspect of the spatial patterning of the genetic variability. On the one hand, scores with a strong variance and a highly positive spatial autocorrelation (that is, global structures) correspond to highly positive eigen values. On the other hand, scores with a strong variance and a highly negative spatial autocorrelation (that is, local structures) correspond to highly negative eigen values. The negative eigen values are thus useful tools to detect local patterns and are not ignored during further analysis.

A simple bar plot of sPCA eigen values gave first idea of the global and local structures. Each eigen value can be decomposed into a variance and a spatial autocorrelation (Moran's I) component. The graphical representation is as follows in Fig. 3.

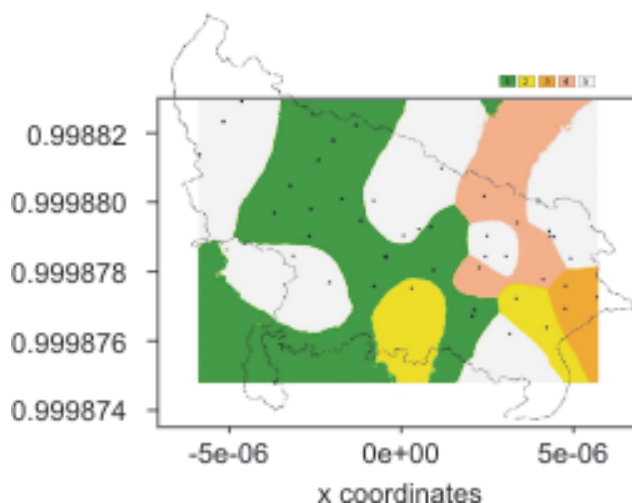


Fig. 2. Map of the mode of posterior probability to belong to each of the buffalo populations of Indo-Gangetic plains.

The resulting figure represents eigen values of sPCA (denoted $\tilde{e}_i$ with $i = 1 \text{ } n - 1$, where $\tilde{e}_1$ is the strongest global eigen value, and $\tilde{e}_{n-1}$ is the strongest local eigen-value) according to their variance and Moran's I components. These eigen values are contained inside a rectangle indicated in dashed lines. The maximum attainable variance by a linear combination of alleles is the one from an ordinary PCA,

Table 1. Depicting the F_{ST} values among the various clusters identified by Geneland analysis

Cluster	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Cluster1	0				
Cluster2	0.0883	0			
Cluster3	0.0661	0.242	0		
Cluster4	0.0336	0.0477	0.1611	0	
Cluster5	0.0499	0.0696	0.1814	0.0181	0

The maximum difference is between cluster 2 and cluster 3 and the least between cluster 4 and cluster 5. We did not encounter any ghost population in the analysis.

indicated by the vertical dashed line onto the right. The two horizontal dashed lines indicate the range of variation of Moran's I, given the spatial weighting matrix that was used.

This figure is useful to assess whether a given score of entities contains relatively enough variability and spatial structuring to be interpreted. For instance, here, $\tilde{e}_1$ clearly is the largest eigen value in terms of variance and of spatial autocorrelation, and can be well distinguished from all the other eigen values. Hence, only the first global structure, associated to $\tilde{e}_1$, was interpreted.

The first plot shows the connection network that was used to define spatial weightings. The second, third, and fourth plots are different representations of entities scores onto one axis in space, the first global score being the default argument axis. In each, the values of scores [$\$li(axis)$ component of the `spca` object] are represented using black and white symbols (a variant being grey levels): white for negative values, and black for positive values. The second plot is a

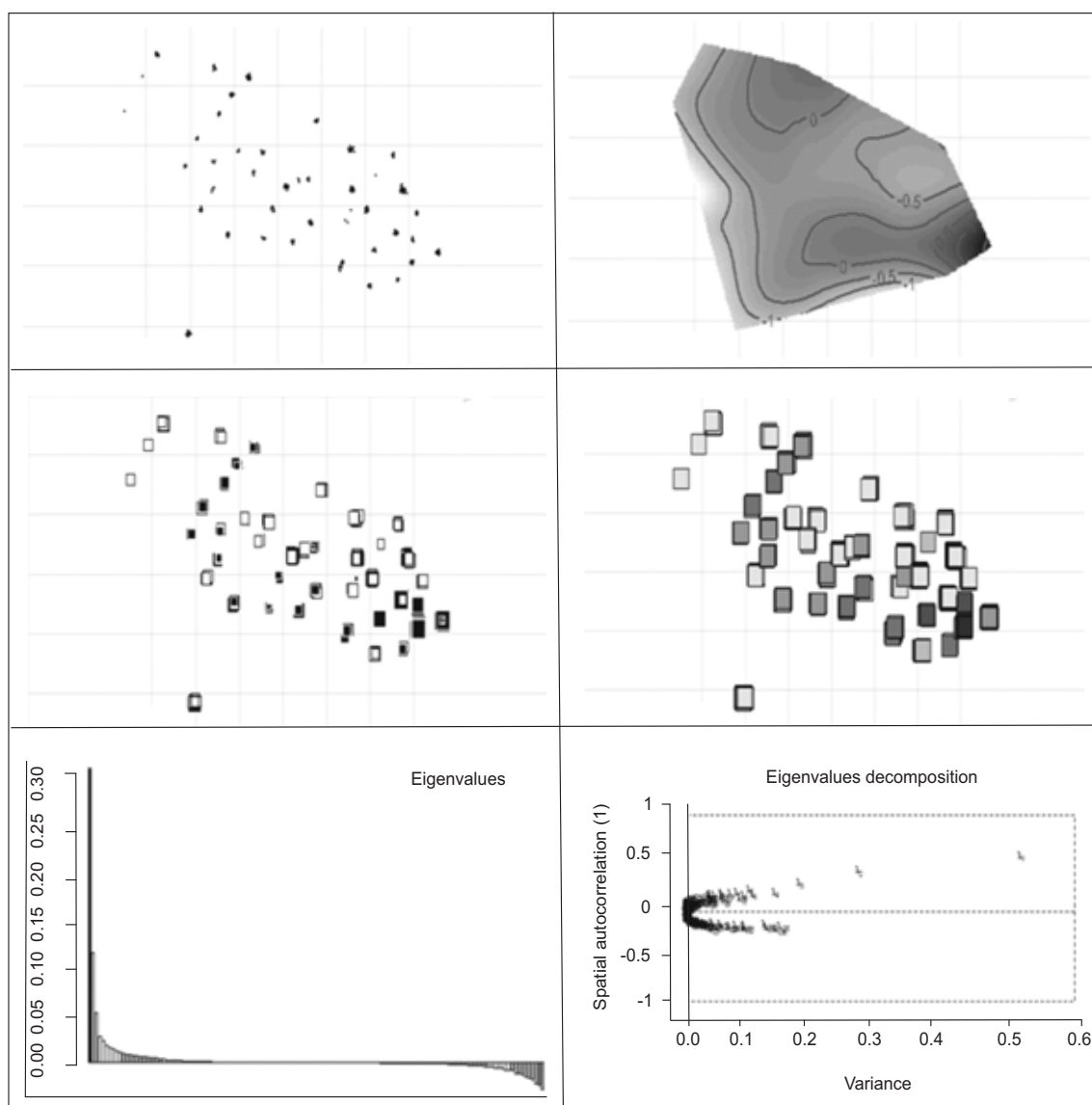


Fig. 3. Graphical representation of SPCA eigen value

local interpolation of scores (function `s.image` in `ade4`), using grey levels, with contour lines. The closer the contour lines are from each other, the steeper the genetic differentiation is. The third plot uses different sizes of squares to represent different absolute values (`s.value` in `ade4`): large black squares are well differentiated from large white squares, but small squares are less differentiated. The fourth plot is a variant using grey levels (`s.value` in `ade4`, with 'grey level' method). Here, all the three representations of the first global score show that genotypes are split in five genetical clusters, one in the west (or left) and one in the east (right), the others are intermingled amongst themselves and are not in continuity. The last two plots of the `plot.sPCA` function are the two plots of eigen values and eigen value decomposition respectively.

The color plot is another way of representing a score of sPCA. The three scores are depicted at the same time by translating each score into a channel of color (red, green and blue). The obtained values are used to compose a color using the RGB system. sPCA is an exploratory tool. The approach seems relevant as it is a reduced space ordination method; no assumptions are made about the data model. It is thus free, for instance, from modeling constraints like Hardy–Weinberg equilibrium assumptions, which are often violated when considering markers involved in selection processes. In the present study, the results obtained in both the methods are quite consistent and five clusters based on the analysis could be delineated (Fig. 4). The 5 clusters obtained had sampling locations that were not geographically contiguous. The normal multivariate analysis and other methods do not reveal the population structures which are not continuous and the inclusion of spatial coordinates not only yield global structures but also local spatial structures.

The inclusion of spatial parameters in the genetic analysis provided an effective means to decipher the population genetic structures. sPCA is a method of choice when the

genetic analysis is carried out when the assumptions of various models do not hold true. The genetic landscape of Indo-Gangetic plains shows both the global as well as cryptic genetic structures at local scale.

REFERENCES

- Cliff A and Ord J. 1981. *Spatial Processes. Model and Applications*. Pion, London.
- Crandall K A, Bininda-Emonds O R P, Mace G M and Wayne R K. 2000. Considering evolutionary processes in conservation biology. *Trends in Ecology and Evolution* **15**: 290–95.
- Guillot G, Estoup A, Mortier F and Cosson J F. 2005. A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–80.
- Guillot G, Santos F and Estoup A. 2008. Analysing georeferenced population genetics data with Geneland: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics* **24** **11**: 1406–07.
- Jombart T. 2010. *ade4*: a R package for the multivariate analysis of genetic markers. Retrieved from: <http://ade4.genet.r-forge.r-project.org/>.
- Jombart T, Devillard S, Dufour A B and Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**: 92–103.
- Mishra B P, Kataria R S, Kathiravan P, Bulandi S S, Singh K P and Sadana D K. 2009. Evaluation of genetic variability and mutation drift equilibrium of Banni buffalo using multi locus microsatellite markers. *Tropical Animal Health Production* **41**: 1203–11.
- Mishra B P, Kataria R S, Bulandi S S, Prakash B, Kathiravan P, Mukesh M and Sadana D K. 2009. Riverine status and genetic structure of Chilika buffalo of Eastern India as inferred from cytogenetic and molecular marker-based analysis. *Journal of Animal Breeding and Genetics* **126**: 69–79.
- Moran P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society* **B** **10**: 243–51.
- Moritz C. 1999. Conservation units and translocations: strategies for conserving evolutionary processes. *Heredity* **130**: 217–28.
- Sahai R and Vijn R K. 2000. *Domestic Animal Biodiversity-Conservation and Sustainable Management*. S.I. Publications, Karnal, India.
- Sambrook J, and Russell D W. 2001. *Molecular Cloning: A Laboratory Manual*. 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Tantia M S, Vijn R K, Mishra B, Bharani Kumar S T, and Arora R. 2006. Multilocus genotyping to study population structure in three Buffalo populations of India. Asian-Australas. *Journal of Animal Science* **19**: 1071–78.
- Thioulouse J, Chessel D and Champely S. 1995. Multivariate analysis of spatial patterns: a uniform approach to local and global structure. *Journal of Environmental and Ecological Statistics* **2**: 1–14.
- Vijn R K, Tantia M S, Mishra B and Bharani Kumar S T. 2008. Genetic relationship and diversity analysis of Indian buffalo. *Journal of Animal Science* **86**: 1495–1502.
- Vijn R K, Tantia M S, Mishra B and Bharani Kumar S T, Kumar P K, Vij A R, Sirothia K P, Singh V B, Kharadi P J, Gajbiye M, Iyue and Gujjar B V. 2008. Genetic relationship among buffalo breeds of India. *Indian Journal of Animal Sciences* **78**(4): 378–83.

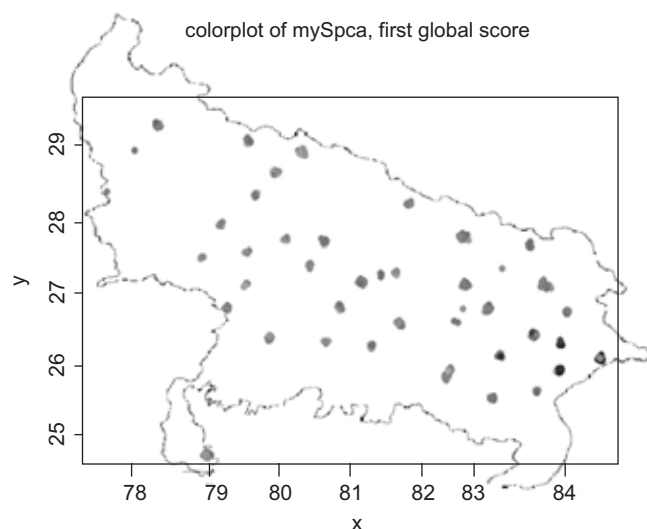


Fig. 4. Color plot of spca, first global score.