Landscape genomics and multivariate analyses: examples and prospects for poultry

D. Laloë and T. Zerjal*

INRA, UMR1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy-en-Josas, France
AgroParisTech, Génétique Animale et Biologie Intégrative, 16 rue Claude Bernard, F-75231 Paris, France
*denis.laloe@jouy.inra.fr

Multivariate Analysis or, more specifically, Geometric Data Analysis, is a statistical approach that represents multivariate datasets as a cloud of points in n-dimensional space and bases the interpretation of data on these clouds (Le Roux and Rouanet, 2004). It may be traced back to Karl Pearson, whose Principal Components Analysis (Pearson, 1901) is based on a geometric display of data. This geometric modeling allows us to consider a data table as a cloud of individuals (points representing individuals), or as a cloud of variables (points representing variables). This dual approach has been formalized by the so-called duality diagram (Cailliez and Pages, 1976; De la Cruz and Holmes, 2011), which provides a simple way to put many multivariate methods in the same framework.

Landscape genetics may be defined as an approach for describing how geographical and environmental features structure genetic variation at both the population and individual levels, highlighting the relationship existing between spatial genetic structure and the structure of landscapes. In the context of multivariate analysis, this problem may be addressed through specific methods, namely spatial multivariate analyses and redundancy analyses (a.k.a analyses with instrumental variables).

Spatial multivariate analysis

Briefly, while in the principal component analysis (PCA), the optimization criterion only deals with genetic variance (with the eigenvalue decomposition of **X'X**, where **X** is the matrix of allelic frequencies), the spatial PCA (sPCA) aims at finding independent synthetic variables that maximize the product of the genetic variance and the spatial autocorrelation. This is accomplished by the eigenvalue decomposition of the matrix **X'(L+L')X** where **L** synthesizes spatial structure among populations via a neighboring graph connecting the populations on the geographical map to model spatial structure among breeds. The resulting eigenvalues can be either positive or negative reflecting respectively a global or local spatial pattern (Jombart et al., 2008). Calculations were carried out using the *adegenet package* (Jombart, 2008) of the R software (http://www.R-project.org).

Redundancy analysis

Redundancy analysis (RDA), also known as "Principal Components Analysis with instrumental variables" (Rao, 1964) is an analysis that seeks how much of the variation in one set of variables, say **X** (e.g., landscape or climate variables) explains the variation in another set of variables, say **Y** (e.g., genetic data). RDA produces principal components that are constrained to be linear combinations of **X**. This analysis is appropriate when the number of variables in **X** is lower than the number of variables in **Y**. RDA permits variance partitioning to measure the variance explained by different sets of instrumental variables and then to sequentially test the significance of variables by an ANOVA-like permutation test (Liu, 1997; Legendre and Legendre, 2012). Calculations were carried out using the R packages *vegan* (Oksanen et al., 2013) and *ade4* (Chessel et al., 2004).

Data

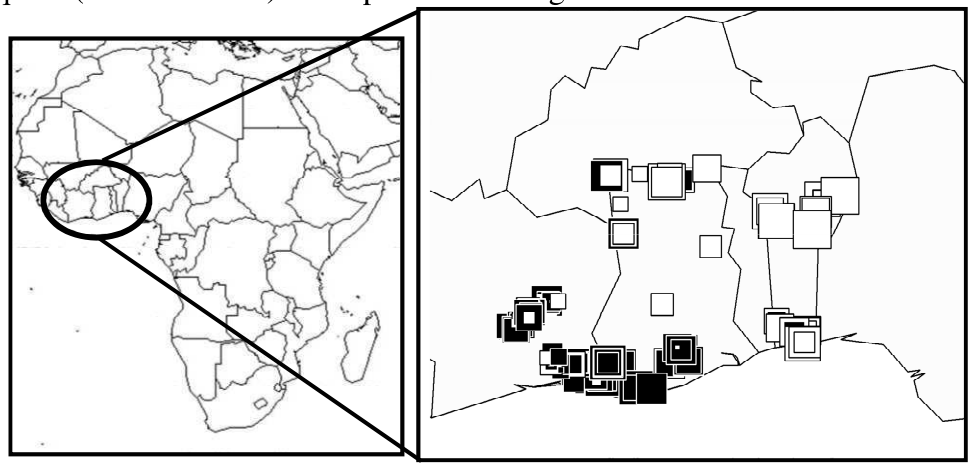
Features of these methods are illustrated through the analysis of published data (Leroy et al., 2012) consisting of 317 local African chickens sampled in an area including Ghana, Benin and the Ivory Coast, and genotyped with a set of 22 microsatellites. Geographic coordinates for each animal

were recorded, and climatic data (elevation, temperatures and rainfall) were obtained from the FAOCLIM database.

Results

The information contained inside a sPCA object can be displayed in several ways, but a frequent practice in spatial genetics is mapping the first principal components (PCs) onto the geographic space because it offers an interesting visual result of multivariate analyses. The sPCA for the African chickens is summarized in Figure 1 where individual scores are plotted on the geographical map of origin. Individuals are represented by squares. The areas of the squares are proportional to the absolute value of the score. The color of the square (black or white) corresponds to the sign of the score. Figure 1 shows the existence of a clear genetic structure opposing animals from the south-west to animals from the north-east regions. No clear geographical physical barriers separate the two areas, but they differ climatically.

Figure 1. Projection of the individual scores of the first spatial principal component onto the geographical map. The areas of the squares are proportional to the absolute value of the score. The color of the square (black or white) corresponds to the sign of the score.

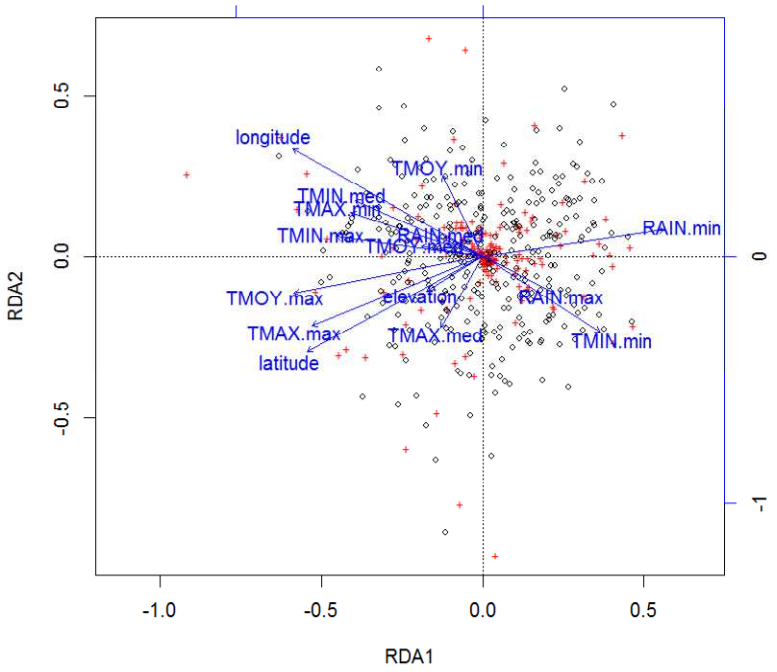


To quantify how much of the genetic variation is explained by geography and/or climatic conditions we used the RDA approach. The results are summarized in the biplot of Figure 2 that represents the correlation of each geographical or climatic variable with the two first components. Particularly, the first component (RDA1) is associated to geographical coordinates, longitude and latitude, and to a rainfall variable *RAIN.min* (minimum monthly rainfall over a year). ANOVA-like tests indicate that the effects of the geographical and climatic variables on the genetic variation are significant ($p<0.01$). Climate conditions, independent of the geography, account for 33% of the constrained total variance, while the geography, independently of the climate, accounts for 13% of it. The rest is due to a combined effect of the two.

Perspectives

The landscape genomic multivariate analysis is certainly a very promising approach for understanding the geographic effect in shaping population genetic structure. The field of landscape genomics has been largely employed in conservation and ecological studies. On the contrary, there have only been a few studies involving domesticated species (Laloë et al., 2010; Gautier et al., 2010) but the growing amount of genomic data available for domesticated animals makes this approach particularly suitable to understand the evolutionary processes and adaptive events that have shaped the genetic diversity of domesticated animals.

Figure 2. Redundancy analysis: biplot corresponding to the first two components. Points are the projection of individuals onto the components; blue arrows represent the correlations of geographical and climatic variables with the components. Climatic data gathered in form of twelve monthly averages of precipitation (RAIN), minimal temperature (TMIN), mean temperature (TMOY), maximal temperature (TMAX) were used to obtain the maximum (.max), the median (.med) and the minimum (.min) values then used in the analysis. For example, TMIN.max is the maximal value among the monthly minimal temperatures (i.e., the minimal temperature of the warmest month).



References

Cailliez, F., and J. P. Pagés. 1976. Introduction à l'analyse des données. Société de Mathématiques Appliquées et de Sciences Humaines (SMASH), Paris, 616 p.

Chessel, D., A. B. Dufour, and J. Thioulouse. 2004. The ade4 package – I: One-table methods. *R News* 4:5-10.

De la Cruz, O., and S. Holmes. 2011. The duality diagram in data analysis : Examples of modern applications. *Annals of Applied Statistics* 5:2266-2277.

Gautier, M., D. Laloë, and K. Moazami-Goudarzi. 2010. Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PLoS ONE* 5:e13038.

Jombart, T. 2008. *adegenet*: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403-1405.

Jombart, T., S. Devillard, A.-B. Dufour, and D. Pontier. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* 101:92-103.

Laloë, D., K. Moazami-Goudarzi, J. A. Lenstra, P. Ajmone-Marsan, P. Azor, R. Baumung, D. G. Bradley, M. W. Bruford, J. Cañón, G. Dolf, S. Dunner, G. Erhardt, G. Hewitt, J. Kantanen, G. Obexer-Ruff, I. Olsaker, C. Rodellar, A. Valentini, P. Wiener, and the European Cattle Genetic Diversity Consortium and Econogene Consortium. 2010. Spatial trends of genetic variation of domestic ruminants in Europe. *Diversity* 2:932-945.

Le Roux, B., and H. Rouanet. 2004. Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. 3rd English edition, Elsevier Science B.V., Amsterdam, The Netherlands.

Leroy, G., B. B. Kayang, I. A. K. Youssao, C. V. Yapi-Gnaoré, R. Osei-Amponsah, N. E. Loukou, J.-C. Fotsa, K. Benabdeljelil, B. Bed'hom, M. Tixier-Boichard, and X. Rognon. 2012. Gene diversity, agroecological structure and introgression patterns among village chicken populations across North, West and Central Africa. *BMC Genetics* 13:34.

Liu, Q. 1997. Variation partitioning by partial redundancy analysis (RDA). *Environmetrics* 8:75-85.

Oksanen, J., F. Guillaume Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens, and H. Wagner. 2013. Community Ecology Package, version 2.0-8. *http://CRAN.R-project.org/package=vegan*

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2:559-572.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya, A* 26:329-359.