Ascertainment of evolutionary processes from the genetic variation associated to each geographic point in a map

J. R. Quevedo,[*] E. Fernández-Combarro,[*] L. J. Royo, [†] I. Álvarez, [†]A. Beja-Pereira, [‡] I. Fernández,[*] A. Bahamonde[*] and F. Goyache[†,1]

[*]*Centro de Inteligencia Artificial. Universidad de Oviedo at Gijón, Campus de Viesques, E-33271 Gijón (Asturias), Spain.*
[†]*Área de Genética y Reproducción Animal, SERIDA, E-33203 Gijón, Asturias, Spain.*
[‡] *CIBIO-UP, Faculdade de Ciências, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão (VCD), Portugal;* [1]*Correspondence to*: fgoyache@serida.org

*Keywords***:** Gene variation, map projection, interpolation, diversity, uniqueness.

Introduction

The study of spatial patterns of genetic variation is useful to infer evolutionary processes (Cavalli-Sforza et al., 1993). However, appropriate techniques for spatial analysis are scarce; they usually highlight the existence of nonrandom patterns of genetic variation but do not provide information on the source and direction of the evolutionary processes we are interested in assessing.

Yet again, the techniques available for visualizing genetic variables on physical maps are limited and are based on the simple interpolation of the genetic information between two given locations (Menozzi et al., 1978). Interpolation can be a robust tool, since we can always obtain a vector connecting two distant places when genetic information from areas located between them is not available, and has consequently been widely used (Cavalli-Sforza et al., 1993; Beja-Pereira et al., 2002; Hanotte et al., 2002).

However, we may be interested in assessing geographical patterns of genetic variation taking into account the fact that not each and every one of the alleles carried by a population may be found at a single geographical coordinate. Individuals are usually more or less uniformly distributed in an area and the genetic variability we may find in an area around a geographic point will in fact be formed by the presence of individuals from different populations. In addition, if the variable to be mapped at two neighboring geographic points has similar values (despite the allelic frequencies of the population being different), the interpolation between them will tend to be constant.

Another problem is the selection of the genetic variables to be projected. Menozzi et al. (1978) proposed the use of principal component analysis to obtain (at least) one new synthetic variable summarizing most of the available genetic information. This technique has a number of limitations: genetic information may be partially lost in the computation of the factors and the number of components we can obtain will depend on the data set. However, its major limitation is that the computed synthetic variables are not directly informative and must be interpreted. Of course, each component provides a different picture on the same geographic map. Clearly, the procedure will be more useful if a large amount of the information can be summarized using only a few variables that need not be interpreted. Recent papers (see Widmer and Lexer, 2001) have highlighted the need to assess both the observable frequencies of marker alleles

(i.e. Nei's gene diversity; Nei, 1987) and the distribution of the alleles (i.e. allelic richness; Hurlbert, 1971) in order to ascertain how populations evolve.

In this presentation we propose the use of a single approach to assess the geographical patterns of gene variation from multilocus information. Our aim is helping the major drawback shared by the previously available methodologies: the assumption that all the samples obtained for a given population come from a predetermined location (Wasser et al., 2004). We advocate summarizing multilocus genetic information in just two clearly defined and easily interpretable parameters that will allow us to estimate gene flow patterns, taking into account both the highly stochastic nature of gene variation (diversity) and the distribution of rare alleles, as an estimation of the uniqueness of the populations. The model we have in mind takes into consideration that the sampled populations can have a particular geographical spreading and its individuals can be sampled in any location in the distribution area of the population to which belong. In consequence we advocate for the assessment of the genetic variation associated with each geographical point on our map to infer evolutionary patterns rather than that associated with a previously defined population. The methodology described below has been implemented in the program GeoGen, which can be freely downloaded from the World Wide Web at http://www.aic.uniovi.es/GeoGen/.

Materials and Methods

Throughout this work we will manage two concepts: a) population, which is here a whole of sampled individuals sharing historical, geographical or genetic characteristics that is described by 3 main attributes: size, distribution area (with center in a point called node following Menozzi et al., 1978) and a set of allelic frequencies; and b) deme, which is here the whole individuals (or better the whole set of alleles) belonging to any 'real' population located in an area around a given geographical point in a map here notated as (x,y). In other words, here population is the input from which we will try to infer the allelic distribution in a geographical map and demes are the 'working populations' from which we will compute the genetic variables associated to each geographical point. We will start from actual information at a population level (using at least the 3 attributes listed above) to compute the most likely distribution of the alleles represented in our population's set in an area around each point (x,y) in our map (deme). We will compute various genetic variables on each deme and, finally, the computed genetic variables will be graphically represented onto the map to allow visual inspection of possible patterns of variation.

Our methodology assumes that genetic information is only available at a population level and that we do not know the exact location at which individuals belonging to each population have been sampled. However, we can estimate the most probable distribution of the individuals of a population in a given area. To do so, we will need the following parameters: a) a node (main geographical location for a given sampled population); b) a radius from the node to define an area in which we can expect to find individuals belonging to a given population; c) an arbitrary definition as to the way in which the individuals are distributed in the area assigned to their population (basically assuming the same density of individuals throughout the area or a uniform loss of density from the data point to the periphery of the distribution); and d) an estimate of the actual population sizes. A number of individuals like that of the size of the population are distributed in the area around each node. Typically, we can expect a non-

uniform distribution of the genotypes in the area of population spread. In such a situation, our methodology assumes that the density of individuals decreases with the distance to the node: the more distant the geographical point, the lower the probability of finding an individual belonging to the population.

Being available the information above for each population we will first determine the number of individuals of each population that might be found in an area surrounding (x, y) and that hence have some influence on the genetic variation at this point (see please Figure 1). All the individuals belonging to any population that can be found in the area fitted around (x, y) will form the deme affecting genetic variation at this point. To implement this step, we need a radius (r_d) that defines the deme's area, which is the area of influence of each geographical point (x, y). Then, the number of individuals belonging to the population *s* that are likely to be present in the deme's area can be computed from the proportion of individuals of the population present in a circle of centre (x, y) and radius r_d as $N(x,y,s) = int\left[T_s * \int_{C(x,y)\cap C(s)} D_s(x_1,x_2)\, dx_1 dx_2 \right]$, where $C(x, y) \cap C(s)$ is the intersection between the circle of radius r_d with centre in (x, y) and the circle of radius r_s with centre in the node p_s; D_s is the density function of population; and *int* stands for the function that computes the integer part of a real number. The whole individuals belonging to any population located in the circle of radius r_d and centre in (x,y) is the deme associated to this geographical point on which we will compute the genetic variables to be represented onto a map.

Further, to compute the genetic variables associated to the point (x, y) we have to read the individuals of each population forming the deme as sets of alleles taking into account the available allelic frequencies for each population. To discuss this point, let us fix a population *s* and a locus *l* with alleles in {1, …, m_l}. A first attempt would be to consider that we have in $C(x, y) \cap C(s)$ each of the alleles present in the population s, and that the number of such alleles would be proportional to the number N(x,y,s) of individual of *s* around (x, y). However, this is not a realistic assumption, since rare alleles are not likely to be present if N(x,y,s) is small. Therefore, we opt to distribute the twice N(x,y,s) alleles according to their relative frequency in the population *s*. So, given a pair (a_1,a_2) of different alleles on the locus *l* and assuming the independence of their apparition in each individual, we will simply have: $Pr(a_1,a_2 \mid s,l) = f(s,l,a_1) * f(s,l,a_2)$

Genetic variables to be mapped

Once we known the allelic frequencies associated to each point in map we will compute two genetic measures computed following the methodology proposed by Shannon (1948): a) diversity $D_{(H)}$ which is defined as the entropy associated with the probability distributions of the allelic frequencies in a set of loci; and b) uniqueness (*U*) which is defined as the entropy associated with the probability distribution of an allele across loci and populations. The Shannon's formula defines the entropy *H* of the variable *X* is defined by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_i is the probability associated with a certain value of a random variable *X*. Since allelic frequencies at a given locus represent a probability distribution with associated entropy, *diversity* ($D_{(H)}$) would simply be the average entropy associated with multilocus genotype information; the more polymorphic and the more uniform the alleles are distributed in the sampled loci, the higher the value of associated entropy will be.

On the other hand, we use Shannon's formula to assess *genetic uniqueness* (*U*). In terms of entropy, the rarity of a given allele can be assessed with respect to the population in which it is found and across the entire set of sampled populations. The idea is that rare alleles (those present in one or just a few populations in low frequency) produce higher *U* values than those uniformly distributed across populations. Across populations, we have a discrete distribution for each allele *a* on loci *l* in each population *s*. However, the rare alleles will yield values of entropy (*H(l,a)*) near 0, while the highest values will be reached by those alleles uniformly spread in all populations, with values near $\log_2$ of the number of populations (*N*). Since we are interested in emphasizing the rarity of the alleles, we have to invert the computed values of entropy in order to obtain a first approach to the uniqueness (U) of a given allele across populations. Thus, we define $w(l,a) = \log_2(N) - H(l,a)$ and then we use $w(l,a)$ to weight the absolute frequency (n) found for the allele *a* on locus *l* for the individuals belonging to the population *s* in the deme's area around (x, y), and finally, we define

$$U = \max_{l=1}^{l}\left(\max_{a=1}^{a}\left(w(l,a)\sum_{N=1}^{N} n\right)\right)$$

which is a good approach to our requirements; the less frequent and the less smoothly distributed allele across populations, the higher the value of associated *U* will be.

Data sets

The methodology is demonstrated by applying it to: a) two microsatellite data sets for livestock populations of the ovine (unpublished) and bovine (Beja-Pereira et al., 2003; Jordana et al., 2003) species; and b) a published microsatellite data set for Eurasian otter (*Lutra lutra*) in Scotland (Dallas et al., 1999; 2002).

The first one livestock dataset genetically describes the rare Xalda sheep breed (including less than 600 live individuals) located only in Asturias (Northern Spain) (Goyache et al., 2003). For this study, 160 individuals were genotyped for a previously published set of 14 microsatellites (Álvarez et al., 2004; 2005). These 160 individuals can be classified in 14 different herds (population sample size ranging from 4 to 33). Allelic frequencies, population sizes and geographical locations for the described dataset are given as example data with the program GeoGen.

The second one livestock dataset consists of a total of 749 unrelated individuals from 10 Spanish and 5 Portuguese cattle breeds that were genotyped for 16 microsatellites in an

extensive study of genetic diversity in the local bovine populations of the Iberian Peninsula (Beja-Pereira et al., 2003; Jordana et al., 2003). Patterns of genetic variation for this dataset were recently assessed using Principal Component Analysis and interpolation (Beja-Pereira et al., 2003), thus being useful to distinguish some interesting differences between methodologies.

Both datasets are valuable tests for our method, as the amount of differentiation between populations was fairly wide (observed pairwise F_{ST} values ranging between 0 and 0.186 for the ovine dataset and from 0.032 to 0.148 for the bovine dataset - Jordana et al., 2003). In the ovine dataset, geographical distances between the populations all occur in a limited area. In the bovine dataset, however, some populations are fairly close to one another, with no obvious physical barriers to gene flow between them (e.g. those settled in the Cantabrian Range), while others are separated by large geographical distances and management systems (e.g. the three Southern Iberian breeds included in the dataset and those breeds settled in the Pyrenean Mountains).These samples allow us to determine the effect of differences in genetic differentiation on our method's ability to obtain reliable geographical patterns of gene variation when this is more likely to be the result of different selection forces combined with genetic drift.

The Scottish otter dataset has been used as a useful test bench for our method, as the genetic differentiation between geographical populations is not dependent on human management. This natural population has been extensively characterized (Dallas et al., 1999, 2002) showing a strong, unexpected pattern of isolation by distance in the continuous otter population on the mainland, mainly affecting southern *versus* northern populations. Significant differences in genetic composition were found among groups of otters in areas between which no obvious barriers to movement exist.

Results and Discussion

The projections of diversity and uniqueness for the *Xalda* datasets using both our methodology and interpolation are shown in Figure 2. We can only clearly detect a gradient from west to east for both diversity (*D*) and uniqueness (*U*) using our new methodology, thus leading us to consider the western *Xalda* populations as the most important for the breed, what is consistent with the history of the recovery of the breed (Álvarez Sevilla et al., 1982; Álvarez Sevilla et al., 2003). The spot identified for uniqueness in the easternmost area of distribution of the *Xalda* population is more likely to be the result of finding a contact zone in which some rare alleles that are non-uniformly distributed in the populations of the area would be ideally represented.

On the other hand, the maps built from the bovine dataset (Figure 3) can be assessed taking into account the previous work by Beja-Pereira et al. (2003). These authors identified two main principal components respectively explaining 16 and 12% of the total genetic variation in the dataset. Synthetic maps constructed interpolating the factors (Menozzi et al., 1978) clearly distinguish four Northern Iberian breeds from the others (first factor) and gave an east–west gradient of genetic variation (second factor) that was tentatively explained by the authors as consistent with the model of demic diffusion of agriculture suggested by Cavalli-Sforza et al. (1993). Although the interpolation of diversity (*D*) and uniqueness (*U*) do not give plots consistent with those provided by Beja-Pereira et al. (2003), our methodology is useful in explaining that the first factor of these authors is a variable combining both diversity (*D*) and uniqueness

(*U*). The Northern Iberian area identified as a hot spot of variability by Beja-Pereira et al. (2003) includes the breeds in which we detected the highest diversity and uniqueness. In contrast with our methodology, interpolation does not give images that are consistent with those previously reported by Beja-Pereira et al. (2003). The analysis of this dataset highlights the importance of using well-defined and easily interpretable variables to ascertain geographical patterns of gene variation.

Projections of diversity (*D*) and uniqueness (*U*) for the Scottish otter dataset using both our methodology and interpolation are shown in Figure 4. When diversity is projected using our methodology, we can identify three spots of variability, the main one based on the Tayside population and two less important ones for the Inverness and Strathclyde populations. In addition, there seems to be continuity in diversity in the eastern population with its centre in Tayside and extremes in the Dee River and the Borders population. When interpolation is used, however, only one centre of variability can be found (Tayside), spreading mostly in a northerly direction and showing non-consistent geographical continuity with the Inverness population. When uniqueness is projected, both our methodology and interpolation identify two hot spots associated with the Tayside and Dumfries populations. With our methodology, however, the uniqueness of the population of eastern Scotland seems to have similar patterns of variation as diversity.

The two livestock datasets analysed here show that our methodology is useful to identify patterns of genetic variation resulting mainly from genetic drift (genetic isolation in many small herds characteristics of the Northern Iberian Peninsula).Moreover, the maps obtained using our methodology do not need interpretation and we know their meaning exactly. Thus, our methodology can be compared with principal component analysis in detecting patterns of gene variation but is unquestionably more useful in detecting geographic areas acting as reservoirs of diversity and which are probably important targets for conservation purposes. On the other hand, the analysis of the Scottish otter dataset shows that our methodology may be of interest for population management and conservation purposes. As regards both diversity and uniqueness, up to three different programs should be implemented: one including the northernmost Scotland populations with its centre in Tayside in which dispersal should be facilitated; a second affecting the southernmost otter populations, in which maintaining the present levels of diversity and uniqueness should be the main goal; and finally, a conservation program for the autonomous Inverness otter population.

Acknowledgements

This research was partially funded by a grant from the Regional Government of the *Principado de Asturias* No. PC-REC04-27. A.B-P. is supported by a research grant from the Fundação para a Ciência e Tecnologia (SFRH/BD/2746/2000) through the Graduate Programme in Areas of Basic and Applied Biology from the University of Porto (GABBA), Portugal. The authors are indebted to Dr John Dallas and the Department of Zoology of the University of Aberdeen (UK) for providing geographical and genetic data from Scottish otter.

References

Álvarez I., Gutiérrez J.P., Royo L.J., Fernández I., Gómez E., Arranz J.J., Goyache F. (2005) Testing the usefulness of the molecular coancestry information to assess genetic relationships in livestock using a set of Spanish sheep breeds. *Journal of Animal Science*, **83**,737-744.

Álvarez I, Royo LJ, Fernández I, Gutiérrez JP, Gómez E, Goyache F (2004) Genetic relationships and admixture between six Northern Spain sheep breeds assessed through microsatellites. *Journal of Animal Science*, **82**, 2246-22520.

Álvarez Sevilla JA, García Peláez A, Cortés Pérez J (1982) Descripción de la oveja de raza Asturiana. *Boletín de Ciencia Naturales del I.D.E.A.*, **30**, 147-157.

Álvarez Sevilla A, Gutiérrez JP, Fernández I, Royo LJ, Álvarez I, Gómez E, Goyache F (2003) Conservación de la oveja Xalda de Asturias. *AGRI*, **34**, 41-49.

Barbujani G, Bertorelle G, Capitani G, Scozzari R (1995) Geographical structuring in the mtDNA of Italians. *Proceedings of the National Academy of Science of the USA*, **92**, 9171-9175.

Beja-Pereira A, Alexandrino P, Bessa I, Carretero Y, Dunner S, Ferrand N, Jordana J, Laloe D, Moazami-Goudarzi K, Sánchez A, Cañón J (2003) Genetic characterisation of South Western European bovine breeds: an historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity*, **94**, 243-250.

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. *Science*, **259**, 639-645.

Dallas JF, Marshall F, Piertney SB, Bacon PJ, Racey PA (2002) Spatially restricted gene flow and reduced microsatellite polymorphism in the Eurasian otter *Lutra lutra* in Britain. *Conservation Genetics*, **3**, 15-29.

Dallas JF, Bacon PJ, Carss DN, Conroy JWH, Green R, Jefferies DJ, Kruuk H, Marshall F, Piertney SB, Racey PA (1999) Genetic diversity in the Eurasian otter, *Lutra lutra*, in Scotland. Evidence from microsatellite polymorphism. *Biological Journal of the Linnean Society*, **68**, 73-86.

Goyache F, Gutiérrez JP, Fernández I, Gómez E, Álvarez I, Díez J, Royo LJ (2003). Monitoring pedigree information to conserve the genetic variability in endangered populations: the Xalda sheep breed of Asturias as an example. *Journal of Animal Breeding and Genetics*, **120**, 95-103.

Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Edward J, Rege O (2002) African Pastoralism: Genetic Imprints of Origins and Migrations. *Science*, **296**, 336- 339.

Hurlbert S H (1971) The non concept of species diversity: a critique and alternative parameters. *Ecology*, *52*, 577-586.

Jordana J, Alexandrino P, Beja-Pereira A, Bessa I, Cañón J, Carretero Y, Dunner S, Laloe D, Moazami-Goudarzi K, Sánchez A, Ferrand N (2003) Genetic structure of eighteen local south European beef cattle breeds by comparative F-statistics analysis *Journal of Animal Breeding and Genetics*, **120**, 73-87

Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press. NY, USA.

Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, **201**, 786-792.

Sambrook, J., E. F. Fritsch, and T. Maniatis. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.USA

Shannon C E (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423 and 623-656.

Widmer A, Lexer, C (2001) Glacial refugia: sanctuaries for allelic richness, but not for genetic diversity. *TREE*, **16**, 267-269.

Figure 1: This figure illustrates the assumptions of our methodology. Given three populations (pop1, pop2 and pop3) with known distribution the areas Cpop1, Cpop2 and Cpop3 with centre in, respectively, node1, node2 and node3 and characterized by different allelic frequencies we can know how many individuals from each population are in an area around a given geographical point in our map (C(x,y); in dark) and convert this deme in a set of alleles to compute genetic variables associated to the coordinates (x,y).

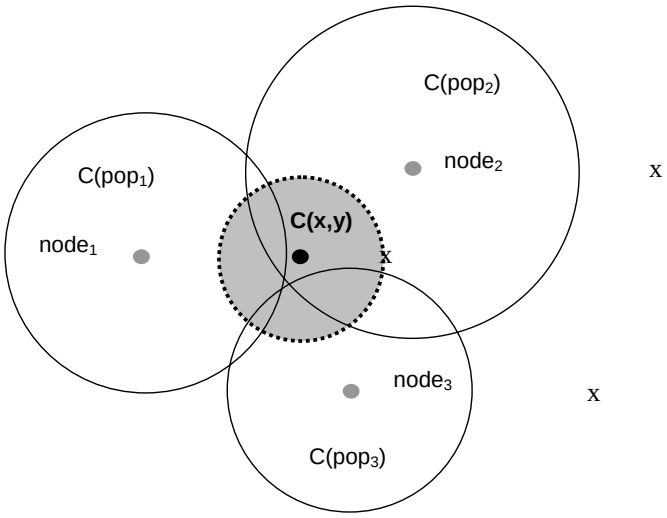

Figure 2: Projection of a microsatellite dataset corresponding to 160 individuals of the rare Xalda sheep breed of Asturias distributed in 14 different herds (sample size ranging from 4 to 33). The variables projected are diversity (*D*) (Figures 1a and 1b) and uniqueness (*U*) (Figures 1c and 1d) using both the methodology presented here (Figures 1a and 1c) and interpolation (Figures 1b and 1d). Names of the populations are presented in the figures.

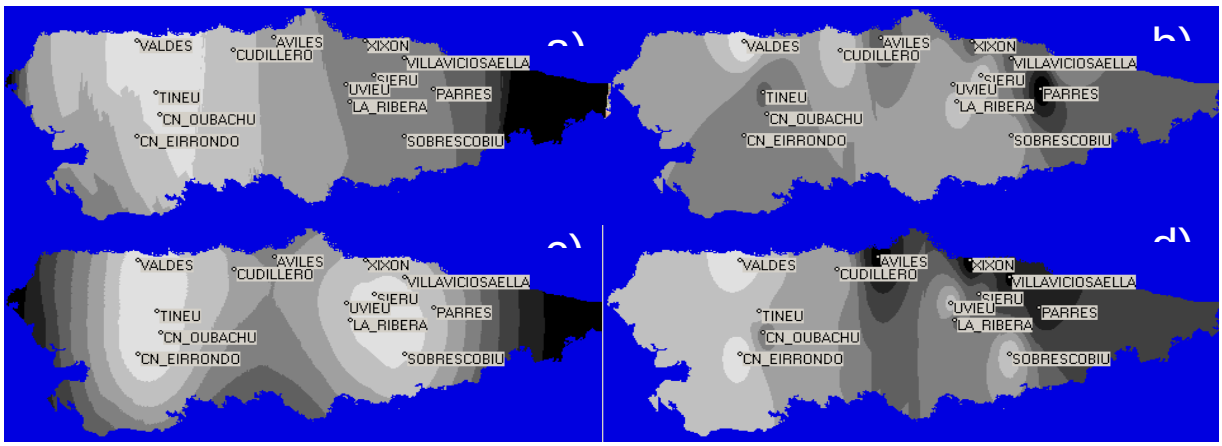

Figure 3: Projection of a microsatellite dataset corresponding to 749 unrelated individuals from 10 Spanish and 5 Portuguese cattle breeds (sample size ranging from 49 to 50). The variables projected are diversity (*D*) (Figures 2a and 2b) and uniqueness (*U*) (Figures 2c and 2d) using both the methodology presented here (Figures 2a and 2c) and interpolation (Figures 2b and 2d). Nodes of the sampled populations are represented in the figures.

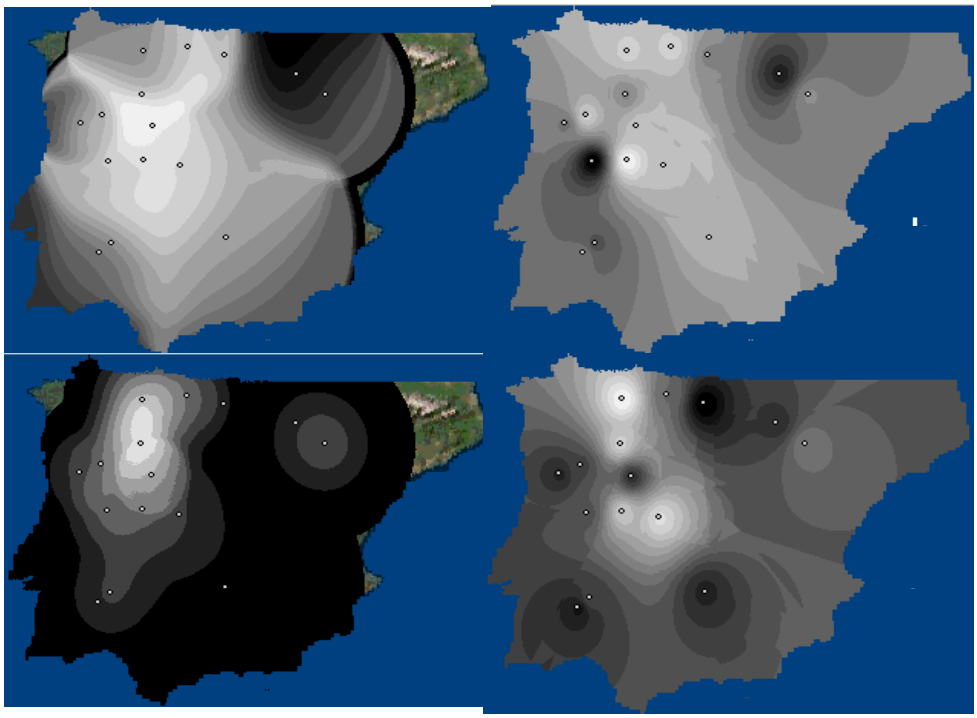

Figure 4: Projection of a microsatellite dataset corresponding to 213 individuals belonging to 9 Scottish populations of Eurasian otter (Lutra lutra) (sample size ranging from 12 to 47). The variables projected are diversity (*D*) (Figures 3a and 3b) and uniqueness (*U*) (Figures 3c and 3d) using both the methodology presented here (Figures 3a and 3c) and interpolation (Figures 3b and 3d). Abbreviations correspond to the following geographical populations: ARG (Argyll), BOR (Borders), DEE (Dee River), DON (River Don), DUM (Dumfries), GAL (Galloway), INV (Inverness), STR (Strathclyde) and TAY (Tayside).