

# Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by "Monmonier's algorithm"

Franz Manni, Etienne Guérard, Evelyne Heyer

Département Hommes, Natures, Sociétés MNHN  
Human Population Genetics Group, CNRS UMR 5145  
Musée de l'Homme – 17, Place du Trocadéro, 75016 Paris.

**THIS IS A PRECIRCULATING PAPER DISTRIBUTED BY THE AUTHOR  
THE ARTICLE HAS BEEN ACCEPTED FOR PUBLICATION IN "HUMAN BIOLOGY"  
AND IS EXPECTED TO APPEAR IN EARLY 2004.  
THE SOFTWARE WILL BE RELEASED IN JANUARY**

## *Abstract*

*When sampling locations are known, the association between genetic and geographic distances can be tested by spatial autocorrelation or regression methods. These tests give some clues to the possible shape of the genetic landscape. Nevertheless, correlation analyses fail when attempting to identify where genetic barriers may exist, namely the areas where a given variable shows an abrupt rate of change. To this end, a computational geometry approach is more suitable since it provides the locations and the directions of barriers and it can show where geographic patterns of two or more variables are similar. In this frame we have implemented the Monmonier's (1973) maximum difference algorithm in a new software in order to identify genetic barriers.*

*To provide a more realistic representation of the barriers in a genetic landscape, a significance test was implemented in the software by means of bootstrap matrices analysis. As a result, i) the noise associated in genetic markers can be visualized on a geographic map and ii) the areas, where genetic barriers are more robust, can be identified. Moreover, this multiple matrices approach can visualize iii) the patterns of variation associated to different markers in a same overall picture.*

*This improved Monmonier's method is highly reliable and it can be applied to a wider range of data than genetic: whenever sampling locations and a distance matrix between corresponding data are available.*

## **1 Introduction**

The classical way to portray genetic variability is to visualize DNA sequences or populations in dendrograms, Multidimensional Scaling (MDS) (Seber 1984; Torgerson 1958) or Principal Components Analysis (PCA) (Gabriel 1968)) plots. These methods enable the recognition of similarities and dissimilarities on a virtual space that corresponds to the plot itself. This approach is particularly suitable to the identification of clusters or of outliers that are informative of the kind of differentiation underlying the genetic variability. On such bases

different genes, species, populations, etc. can be putatively identified. Nevertheless, when the analysis is put forward to the recognition of genetic spatial patterns, related to geography, these methods of analysis can be no more appropriate.

One of the first attempts to study genetic differences in the light of geographic distances was published by Malécot (1948). Their association was formalized in the so called model of “Isolation by Distance” (IBD), meaning that a double logarithmic regression can be expected between a matrix of genetic distances and a matrix of corresponding geographic distances. This method has been proved effective in a number of studies and is often assumed as zero hypothesis of population genetics studies. When a significant regression can't be computed, the association between the two matrices can still be assessed by the Mantel test (Mantel 1967; Manly 1997) which computes the significance by a permutation approach.

Between general methods of geographic analysis there is spatial-autocorrelation, it is used to assess if the values are interrelated, and if there is a spatial pattern to the correlation. Spatial-autocorrelation measures the level of interdependence between the variables, the nature and the strength of the interdependence. Spatial-autocorrelation can be either positive or negative; in positive spatial-autocorrelation all similar values appear together, while in negative spatial autocorrelation dissimilar values appear in close association. In geographic applications there is usually positive spatial autocorrelation. This statistics is generally computed by *Moran's I* by dividing the spatial covariation by the total variation (several other methods exist). The use of spatial autocorrelation analysis of trend residuals has often been advocated to distinguish IBD from regional or long distance process. While the correlation methods described above can assess if there is an association between genetic and geographic distances they convey no information concerning the real patterns of variation of a given variable in bidimensional space. The correlation can be very high in a certain area and much less significant in another one and these differences can be hardly visible in a regression chart that never says “where” the genetic structures are.

To provide a real graphical representation of genetic differences in the geographic space, alternative approaches have been suggested. A first method consists in performing a PCA analysis and then to plot separately the first three principal components on the  $z$  axis of maps where the  $x$  and  $y$  axes are used to plot spatial coordinates (Menozzi et al. 1978). This method is still popular in population genetics even though it doesn't provide a statistical analysis of the pattern of change of genetic frequencies, because it just portrays an interpolated genetic landscape in the geographic space. As an historical note we would like to point out that genogeography, as a field of interdisciplinary investigation, was introduced into science in 1928 by the Russian geneticist A. S. Serebrovsky. Today this field of investigation is better known as gene geography. The method differs from the one of Menozzi and colleagues since the two key points of the representation are the principle of fusion-fission of genes in the homogeneous geographical space (equally free-for-all human genes), and the principle of local-linear (but not of the high-orders) interpolation of gene frequencies onto spherical surface of geographical space (Rychkov et al. 1990).

A third method of geographic visualization of patterns was originally published by Womble (1951) and rediscovered by Barbujani et al. (1989). This methodology focuses on a detailed visualization of the geographic areas associated to a considerable genetic change, what we are going to call from now on “boundaries” or “barriers”. The methodology allows different variables, within the same landscape, to be considered together. First, individual surfaces are differentiated such that steep slopes become peaks and flat ‘plains’ fall to zero. Secondly, the magnitudes of the derivatives of surfaces from different variables can be added together to get a composite picture of barriers derived from all variables. This method was further developed (Barbujani et al. 1989; Oden et al. 1993; Bocquet-Appel & Bacro 1994) and, within a continuous landscape, a consideration of significance was introduced answering

to how 'high' is the cut-off between being a barrier and not being one. By introducing a percentile consideration of significance (in other words considering values in the top X% to represent barriers) values within the landscape are compared one against another thus controlling for the effect of IBD model, if it may be applied. A limitation of the Womble procedure, as well as of the cited PCA method (Synthetic maps of human gene frequencies) already described, is that they imply the interpolation of the landscape leading to potential artefactual continuities or discontinuities (Sokal et al. 1989; 1999a; 1999b).

Since genetic structures may show correspondence with geography, general methods of geographic analysis can be successfully applied to population genetics. This implies the computation of neighbouring problems (computational geometry). This paper is meant to discuss a fourth method, the Monmonier's maximum difference algorithm (Monmonier 1973), designed for the visualization, on a geographic map, of the trend data contained in matrices. The algorithm finds the edges associated to highest rate of change in a given distance measure that can be a genetic one, a morphologic one, etc. The algorithm is applied to a geometric network connecting all the populations (sampled locations) by a Delaunay triangulation (Brassel & Reif 1973). Even though we contributed to popularize the Monmonier method in genetic studies (Manni & Barrai 2000, 2001; Manni & Pagnoni 2001; Manni et al. 2002; Palmé et al. 2003) after the early publication of Barbujani et al. (1996), a detailed discussion of Monmonier's method was never provided. To this end, we have studied the geometric constraints of the triangulation on the Monmonier's algorithm and introduced a method to test the significance of boundaries by the analysis of bootstrap matrices, since no suitable statistics has been proposed to date. This improved version of the method has been implemented in a software, available upon request, called BARRIERS. The identification of the most significant barriers can be generalized to all the cases where *i*) a distance matrix between items is available and *ii*) their sample location is known.

## 2 Materials and methods

### 2.1 The triangulation

Delaunay triangulation (Brassel & Reif 1979) is the fastest triangulation method to connect a set of points (localities) on a plane (map) by a set of triangles (Fig. 1A). It is the most direct way to connect (triangulate) adjacent points on a map. It must be noted that the Delaunay triangulation is the dual structure of the Voronoi diagram (Voronoi 1908) and one can be derived from the other (Fig. 1). Given a set of populations whose geographic locations are known, there is an only possible Delaunay triangulation. Voronoi diagrams, as defined by the author, imply that all possible points inside a polygon are closest to its centroid (the location of the sampled population) than to any other (Fig. 1B). It means that we divide the geographic space  $S$  in  $m$  subspaces  $S_i$  satisfying the following properties:

$$\begin{aligned} \cup_i S_i &= S \\ S_i \cap S_j &= \emptyset & \forall i \neq j \\ \text{DIST}(x_k, w_i) &< \text{DIST}(x_k, w_j) & \forall i \neq j, x_k \in S_i \end{aligned}$$

with  $w_i$  = centroid of  $S_i$ .

Once a network connecting all the localities is obtained, each edge of the network is associated with its distance value from a matrix as shown in figure 3.

## 2.2 *The algorithm*

Monmonier's maximum-difference algorithm (1973) is used to identify boundaries, namely the areas where differences between pairs of populations are largest. The first boundary is traced perpendicular to the edges of the network (Fig. 2). Starting from the edge for which the distance value is maximum and proceeding across adjacent edges, the procedure is continued until the forming boundary has reached either the limits of the triangulation (map) or closes on itself by forming a loop around a population. In case of multiple barriers, that are constructed one after another in a hierarchical order according to user settings, they can stop at a previously computed boundary (Fig. 2). Note that, when two edges have the same value, the one followed by a triangle with higher values is included in the boundary.

## 2.3 *Sample coordinates*

When populations are sampled on a flat surface they can easily be connected by a Delaunay triangulation; on the contrary some problems may arise when samples lie on a curved surface, as the earth surface, since it is not possible to project the position of these samples on a plane without some kind of error, whatever the kind of projection. As an example, since the geographic representation of geographic maps is more exact in its central part, the measuring of  $x$  and  $y$  coordinates of samples located near the borders will be affected by a considerable error. These distortions, related to projection of a curved surface on a plane, can result in a triangulation different from the one obtainable in a curved space, thus the geometry of barriers may be affected. A possible way to overcome this limitation is to compute a matrix of the real geographic distances (on the curved surface) between the points, then to project the position of samples on a MDS plot, and finally by using these new coordinates to compute the Delaunay triangulation. In this way the topological errors in the definition of samples location will be "averaged" giving, as a result, a more accurate triangulation.

## 2.4 *Testing the significance of barriers by a multiple matrix approach*

The definition of the Monmonier's algorithm reminds the dicotomic process of arborescence of phylogenetic trees since, once a barrier passes across the edge of a triangle, it can be extended only across one of the two remaining edges, in what we will define a "right or left" decision. To assess the statistical significance of computed barriers, we have implemented a test that is based on the analysis of resampled bootstrap matrices (for example from molecular sequences). As with bootstrap phylogenetic trees, a score will be associated to all the different edges that constitute barriers indicating how many times each one is included in one of the  $N$  boundaries computed from the  $N$  matrices (typically  $N = 100$ ). The scores are visualized by the software BARRIERS by representing the thickness of each edge proportionally to its bootstrap score (Fig. 6).

## 2.5 *The software BARRIERS*

The software we have developed computes barriers on a Delaunay triangulation using the Monmonier's algorithm and runs under MS Windows (2000 or higher recommended). Minimal systems requirements are 128 MB of RAM and a Pentium II processor. A good video card is recommended to get a fast visualization of ongoing analyses. The program has a graphic clicky interface.

Input files (data coordinates and the distance matrix) can be imported as text files by an interactive window that provides a preview of the resulting file in order to correctly set the line and column from which the import has to be started. Computed triangulation and barriers are saved as *specif* files with the extension *.dvb* (as a contraction of the names Delaunay / Voronoi / Monmonier). Results can be exported in MS Windows BMP format (\*.bmp) and in

Postscript vectorial format (\*.ps). At each run of barrier analysis, a report file (in text format) is generated giving all the details of the different steps of the algorithm in constructing the barrier, edge by edge. Setting options as the color, the thickness of Voronoi tessellation, Delaunay triangulation, Barriers and sample points are available.

The most common bug is related to the presence of samples listed with identical  $x$  and  $y$  coordinates. If this happens the triangulation between these sample points can't be computed. Another bug is related to matrices with identical distance measures since, in this case, the cited "right of left" decision can't be made. The way-out suggested by Barbujani et al. (1996.), to include in the barrier the edge associated to the shortest geographic distance, may be not appropriate since it implicitly assumes the IBD model, even when it was not tested. We undertook a more conservative approach by including, in the forming barrier, that edge that drives the boundary towards that triangle (of the two possible ones) that is associated to higher distance measures.

## 2.6 *Testing barriers : two experimental examples*

To illustrate the wide range of application of the Monmonier's algorithm we will discuss the method on genetic and surname data. Real experimental examples were chosen to avoid the oversimplifications of simulated ones.

### 2.6.1 *Genetic data*

We present the results of the typing of Unique Event Polymorphisms (UEPs) on the non recombinant region of human Y chromosome (Manni et al. 2002). The 17 populations were sampled around the Mediterranean basin accounting for 650 individuals (Fig. 3). An  $F_{ST}$  distance matrix between these population was then computed. These markers enable the definition of haplotypes whose frequencies can be helpful to identify genetic differences between populations, differences that are related to their past demographic history.

### 2.6.2 *Surnames*

Surnames can be considered as one locus on the non-recombining portion of Y chromosome and their analysis enables to get inferences on genetic structures of populations. We studied 2,4 millions of Dutch surnames (Manni 2001) and computed a Lasker's distance matrix. The significance of barriers was further tested by resampling original surnames and then recomputing barriers from 100 bootstrap matrices (Fig. 6). This example was chosen since *i*) the huge sample and *ii*) the geometry of the sampling grid provide an excellent case study.

## 3. **Results**

### 3.1 *Y chromosome variation*

The first barrier, computed on the  $F_{ST}$  distance matrix, divides western European populations from the surrounding ones (top of Fig. 3). In this example we have plotted the thickness of the barrier proportionally to the inverse ratio between the higher  $F_{ST}$  distance (0.603) and  $F_{ST}$  values associated to crossed edges (0.603/0.603; 0.389/0.603; 0.376/0.603; etc.). This method provides a very basic significance test that can be applied to barrier analysis when bootstrap matrices are not available. Results show that the thickness of the barrier decreases from the Gibraltar strait to southern Italy, thus suggesting that the genetic barrier is very strong only between north western Africa and the Iberian peninsula. These results are in agreement with several previously published papers that we will not quote since only the method is here discussed.

We can get a deeper insight in the Monmonier's method by comparing barrier analysis and a MDS plot built of the same  $F_{ST}$  distances. MDS analysis (bottom of Fig. 3) indicates that there is no perfect association between genetic and geographic distances since populations are unequally spaced on the plot, thus suggesting that genetic barriers may exist. Since in the MDS plot there is an only well-defined cluster, formed by the two Moroccan samples, a genetic barrier separating them from all the other populations could be expected, in contrast with obtained results (top of Fig. 3). This apparent discrepancy clearly points to the differences between a classic multivariate analysis and a geographic one that have been graphically illustrated in figure 4. The computed barrier doesn't enclose Moroccan populations since the Algerian sample, geographically contiguous to them, is associated to genetic distances that are higher regarding Spanish populations than Moroccan ones. As a consequence, the geometry of the boundary mirrors the differences between Iberian peninsula and North-Western Africa. Besides this differentiation, the remaining European and Middle-Eastern populations are closer on the MDS plot as it is indeed suggested by the decreasing thickness of the barrier in its final part (top of Fig. 3).

### 3.2 *Genetic structures of the Netherlands inferred by surnames*

We applied the Monmonier's method to identify surname's barriers (the zones where there are the maximum differences in the distribution of surnames) from a matrix of Lasker's distances among 226 sampled localities. We computed the first five barriers that were found contiguous one to each other (Fig. 5A). The geometry of these barriers suggests a main differentiation zone between the North and the South of the Netherlands. The significance of boundaries was tested by multiple analysis of 100 bootstrap matrices. The results confirm that the main differentiation, in the distributions of Dutch surnames, occurs along a South-Western? North-Eastern direction (Fig. 6).

It should be noted that bootstrap analysis conveys additional information since it shows that fragmentation patterns are more complex than it was suggested by the previous analysis of a single overall matrix, particularly in the South-Western area (Fig. 6). To satisfy the curiosity of the reader we will say that obtained barriers almost totally overlap the boundary between Roman Catholics and Protestants. Very few mixed marriages between the two religious groups took place in the past centuries, an occurrence that is mirrored by surnames' distribution.

### 3.3 *Editing the triangulation*

It is possible to edit the triangulation by adding "virtual points" (Fig. 5A/B), they locally modify their neighborhood being interpreted as the borders of the triangulation. This facility is of great importance since the more external links of a Voronoi tessellation, by definition, tend to infinite (Fig. 5B). It often happens that one of these borders is coupled with the highest genetic distance value of the matrix, therefore the origin of the barrier will take place outside the triangulation itself. In Fig. 5 we compare the first five barriers before (A) and after (B) adding the virtual points that close the Voronoi diagrams. Any further triangulation program that doesn't correct this geometric property of Voronoi diagrams is likely to drive to fake barriers when the Monmonier's algorithm is applied to it.

Concluding, the editing of triangulation enables the user to adapt the network to specific features of the geographic space as, for example, in the case of deserts or internal lakes. Moreover, this tool can be useful to delete some long links between distant populations; this is often the case with external samples when the general shape of the triangulation is not convex, since they will be visualized as adjacent by a Voronoi tessellation. An extreme case of external links removal is provided in figure 5A, where all the triangulation lies inside the administrative borders of The Netherlands and only links between very close neighbors were

preserved. A further example is shown in figure 4, here the long link between the French and the Georgian samples was removed (in this case its presence or absence doesn't change the geometry of the barrier, data not shown).

### 3.4 Redundancy of data

The possibility to analyze multiple matrices enables the separate analysis of single markers, thus visualizing the degree of redundancy in the data. Two scenarios may arise: *a)* a large proportion of the markers may exhibit the same geographic pattern of variation or *b)* almost each different marker (sequence) may show a different geographic pattern. This multi-matrix analysis gives a more realistic view of the “noise” associated with each marker and enables to get estimates about their informative power. The separate plot of barriers obtained from different matrices can be helpful in the decision on the number of different markers to be analyzed to get accurate results. If all patterns are different, that could mean that each marker add some information to an overall distance matrix and new markers may lead to more detailed results. If recurrent patterns are observed, it means that studied markers are sufficiently large and no ones are needed to improve barrier analysis, meaning that there is redundancy in the data. This case applies to the example on Dutch surnames (Fig. 6), since the patterns of resampled barriers are geographically quite stable meaning that different surnames (randomly resampled in bootstrap matrices) drive to the same boundary shapes. This conclusion was expected since 2,4 millions of different surnames were studied. Anyhow, even when dealing with very large samples some differences, among the computed patterns, can still be observed. They give a more realistic vision of geographic patterns of variability if compared with the analysis of a single matrix (Fig. 5A).

## 4. Discussion

### 4.1 The ideal case

When attempting to identify boundaries with the Monmonier's algorithm, the best results will be obtained with regularly spaced populations, where the area under investigation approximates a convex hull. Irregularly spaced populations may lead to some ambiguous results since barriers may exhibit a tendency to fall between the most spaced populations that, under an IBD model, are logically expected to be significantly different one from another. In such case it will be impossible to discriminate whether there is a real change in genetic features, among populations separated by the boundary, or if there is a regular genetic pattern of change (related to gene flow) that is not detected since intermediate populations were not sampled. This issue points to the crucial definition of the sampling grid that should be considered before the sampling is performed, in order to get more accurate analyses. As a general rule, large unsampled areas should be avoided when possible. One of the merits of a triangulation-based method is that the irregular distribution of samples is not masked by interpolation, differently from geographic PCA (Menozzi et al. 1978), Womble's method (Womble 1951) or genogeographical approaches (Rychkov et al. 1990).

Some topological configurations of sampled points are not suitable for Monmonier's barrier analyses. A good example of inappropriate data is represented by transects since sampled points are monodimensionally distributed. When the algorithm is applied to such cases the barriers will be forced to “cut” the triangulation vertically, where there are the highest pairwise distances among contiguous populations. Moreover, caution should be adopted in concave shaped triangulations where some links connect very distant populations. These long links may result in factitious barriers because they are probably associated with

the highest distances in the matrix, in the case of an IBD model, as it was discussed above. In such cases it may be useful to test alternative scenarios by editing the triangulation with “virtual points” to the effect of removing these links. The possibility to “close” Voronoi tessellation, implemented in the software, seems necessary since it is the only way, besides computing barriers by hand, to correctly apply the Monmonier’s method without obtaining barriers originating outside the triangulation (Fig. 5B).

#### 4.2 *Population genetics and Isolation by Distance*

Barriers represent zones of abrupt change in the pattern of genetic variation. This means that when populations fit the Isolation by distance model, the chances to find discontinuities in the pattern of change are dwindling. On the other side, the presence of isolating factors (cultural, geographic, morphologic ones) are likely to weaken the gene flow by increasing the chances to find significant barriers. As a general rule, it could be stated that there are as much barriers as the association between genetic and geographic distances is low and vice versa. This does not mean that the algorithm can’t be applied to those cases where genetic distances are demonstrated to increase according to geographic distances. When this dependence can be statistically tested as significant, a good strategy can be *i*) to compute a matrix of expected genetic distances (according to the kind of regression computed) and *ii*) to subtract it from the original matrix thus obtaining a matrix of residuals. The resulting new matrix will enable us to get a further evaluation of the pattern of barriers since it will portray the variation not related to migration phenomena.

#### 4.3 *The range of application of the method*

Since only spatial coordinates and a distance matrix (whatever the kind of measure) are needed to run the Monmonier’s algorithm, the areas of application can be as wide as the computational sciences. We may here recall that when a distance matrix can’t be obtained the method is not applicable. In this respect, the Monmonier method (1973) differs from the Wombling approach (1951) since the latter does not handle distance matrices and enables the analysis of frequency vectors (alleles, haplotypes, linguistic features, morphologic traits) only. We applied the Monmonier’s algorithm to linguistics (Manni 2001), to anthropological data (submitted), to surnames (Manni 2001; Manni & Barraï 2000, 2001), to genetics (Manni 2001; Palmè et al. 2003; Manni et al. 2002). This wide range of possible applications enhance the chances to undertake multidisciplinary and comparative studies that may be impossible with other approaches. Moreover, the application of this method can be useful in a number of biological and medical problems for instance in the case of the identification of the areas where disease rates change rapidly or in the case of the identification of different expression rates of genes.

A further, indubitable merit of the Monmonier’s algorithm is its simplicity that enables the user to supervise the details of ongoing barrier analysis. This simplicity goes together with reliable results as it was recently shown by Dupanloup et al. (2002) when they compared the Monmonier’s algorithm with a new method, to identify highly differentiated populations on a geographic landscape, called SAMOVA (Spatial Analysis of Molecular Variance). The SAMOVA method, an addendum to the Arlequin package (Schneider et al. 2000), is intended to identify maximally differentiated groups of populations without the *a priori* definition needed before (for example according to genetic, morphological, linguistic classifications). In this respect SAMOVA, since it doesn’t deal with a truly geometric approach, is able to identify maximally differentiated populations, whereas the Monmonier’s approach is better at finding genetic barriers between sets of populations (Dupanloup et al. 2002). This latter property of Monmonier’s algorithm seems particularly interesting since maximally

differentiated samples may be already visible in multivariate analyses plots, as PCA or MDS, whereas barriers are not (Fig. 4).

Another advantage of our improved version of the Monmonier's algorithm is that the SAMOVA method, in its currently available version, is only applicable to genetic data (sequences, alleles) whereas the Monmonier's approach can be applied to any kind of distance matrices. Moreover, the geometric properties of Monmonier's method are of particular interest since they make possible an estimate of the robustness of the different edges of the barriers it computes. In this sense, we advocate in resampling techniques (bootstrap, jackknife, etc) a more appropriate way to obtain a realistic picture of the significance of barriers since results can illustrate the differential robustness of the different edges of the formed. According to their differences, the two methods were proved to behave differently in population genetics studies. In a simulation based approach (Dupanloup et al. 2002) it was shown that the Monmonier's algorithm works better than SAMOVA to the effect of finding highly differentiated population groups; especially when the amount of gene flow within-groups is small. In addition, it has been suggested that the performance of both methods decreases when there is a high level of gene flow between-groups and when gene flow within-groups is very low if compared with the gene flow between-groups (Dupanloup et al. 2002). These findings can be considered a different way to state the already cited antagonism between the presence of genetic barriers and an isolation-by-distance scenario since it is trivial that no significant barriers can exist in a perfect cline of frequencies. In this perspective SAMOVA, more than a method to trace boundaries, can be considered a test of significance for population clusters. Concluding, a Monmonier's boundary plot should always be compared with the corresponding MDS or PCA analysis since the two approaches are complementary, being the differences between them the additional information conveyed by Monmonier's analysis, and they can drive to cogent conclusions only when undertaken and discussed together (Figs. 3; 4).

## 5. Conclusions

The Monmonier's maximum difference algorithm enables a better interpretation of microevolutionary processes such as gene flow, genetic drift and selection. It will also help in the identification of hidden boundaries resulting from secondary gene flow among previously isolated populations. The application of Monmonier's algorithm will lead to the understanding of those processes that caused the patterns. The application to population genetics of geographic methods and techniques of analysis seems necessary to a better understanding of the environmental constraints on human demography. Similarly, in ecological studies, philogeographic approaches progressively shifting towards a landscape genetics approach (see Manel et al. 2003). There is an increasing interest in those statistical investigations focused towards the detection of selection processes across geographic space (Nielsen 2001) and the popularization of the Monmonier's algorithm is likely to put forward the understanding of such phenomena. Future developments will probably be directed towards the comparison of different boundary maps, thus enabling the geometric assessment of the similarities/dissimilarities between the geographic patterns of variation of different variables (see Jacquez 1995). The Monmonier's algorithm, since it can be applied to any kind of distance matrix and it is independent from the metrics adopted, constitute an promising method to directly compare the patterns of genetic, cultural or ethnological differentiation.

## Cited references

- Barbujani, G., Jacquez G.M., Ligi L. 1990. Diversity of some gene frequencies in European and Asian populations V. Steep multilocus clines. *American Journal of Human Genetics*, 47: 867-875.
- Barbujani, G., Oden, N.L., Sokal R. R.. 1989, Detecting areas of abrupt change in map of biological variables. *Systematic Zoology*, 38, 376-389.
- Barbujani, G., Stenico, M., Excoffier L., Nigro L. 1996. Mitochondrial DNA sequence variation across linguistic and geographic boundaries in Italy. *Hum. Biol.* 68:201-205.
- Bocquet-Appel, J. P. and Bacro, J. N., 1994, Generalized wombling. *Systematic Zoology*, 43: 442-448.
- Brassel, K.E., and D. Reif 1979. A procedure to generate Thiessen polygons. *Geogr. Anal.*, 325:31-36.
- Dupanloup I, Schneider S, Excoffier L. 2002. A simulated annealing approach to define the genetic structure of populations. *Mol Ecol.*, 11:2571-81.
- Epperson B. and Li T. 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc. Natl. Acad. Sci. USA*, 93: 10528-32.
- Gabriel K.R. 1968. The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58: 453-67.
- Jacquez G.A. 1995. The map comparison problem: tests for the overlap of geographic boundaries. *Stat. Med.*, 14: 2343-61.
- Malecot G. 1948. *Les mathematiques de l'hérédité*. Paris, Masson.
- Manel S., Schwartz, M.K., Luikart G., Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *TRENDS Ecol. Evol.*, 18: 189-97.
- Manly, B.F.J. 1997. *Randomization, bootstrap and Monte Carlo methods in biology*. 2nd ed, Chapman and Hall.
- Manni F., Barraï I. 2001. Genetic structures and linguistic boundaries in Italy: a microregional approach. *Hum. Biol.*, 73: 335-347.
- Manni F., Leonardi P., Barakat A., Rouba H., Heyer E., Klintschar M., McElreavey K., Quintana-Murci L. 2002. Y-chromosome analysis in Egypt suggests a genetic regional continuity in Northeastern Africa. *Hum. Biol.* 74:645-58.
- Mantel N.A. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27:209-20
- Menozzi P., Piazza A., Cavalli-Sforza L.L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*, 201: 786-92.
- Monmonier, M. 1973. Maximum-difference barriers: an alternative numerical regionalization method. *Geogr. Anal.*, 3: 245-61.
- Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86: 641-7.
- Oden N.L., Sokal R.R., Fortin M.J., Goebel H. 1993. Categorical wombling: detecting regions of significant change in spatially located categorical variables. *Geogr. Anal.*, 25: 315-336.
- Palmé A.E., Su Q., Rautenberg A., Manni F., Lascoux M. 2003. Postglacial recolonization and cpDNA variation of silver birch, *Betula pendula*. *Mol Ecol.* 12:201-12.
- Rychkov IuG., Rychkov A.V., Balanovskaia E.V., Batsuur' Zh., Belkovskii A.N., Budilova E.V., Terekhin A.T. 1990. [Genogeography of human populations: computer mapping of population genetics data]. *Genetika*, 26: 332-40
- Schneider, S., Roessli D., Excoffier L. 2000. Arlequin ver 2.0: a software for population genetics data analysis. Genetics and Biometry Laboratory. University of Geneva, Switzerland.
- Seber, G.A.F. 1984. *Multivariate Analysis*. New York (NY): Wiley.
- Sokal R.R., Harding R.M., Oden N.L. 1989. Spatial patterns of human gene frequencies in Europe. *Am J Phys Anthropol* 80: 267-94.
- Sokal, R. R., N. L. Oden, B. A. Thompson, J. Kim. 1993. Testing for regional differences in means: Distinguishing inherent from spurious spatial autocorrelation by restricted randomizations. *Geogr. Anal.*, 25:199-210.
- Sokal RR, Oden NL, Thomson BA. 1999a. A problem with synthetic maps. *Hum. Biol.* 71: 1-13.
- Sokal RR, Oden NL, Thomson BA. 1999b. Problems with synthetic maps remain: reply to Rendine et al. *Hum. Biol.* 71: 447-53.
- Torgerson, W.S. 1958. *Theory and Methods of Scaling*. New York (NY): Wiley.
- Voronoi M.G. 1908. Nouvelles application des paramètres continus à la théorie des formes quadratiques, deuxième mémoire, recherche sur le paralléloèdres primitifs. *Journal Reine Angew. Math.*, 134: 198-207.
- Womble W.H. 1951. Differential systematics. *Science* 114:315-322.

## CAPTIONS TO FIGURES:

### Figure 1

**A:** When populations (solid dots) are located on a surface, a Voronoi tessellation (solid lines) and the corresponding Delaunay triangulation (dotted lines) can be computed. Because of their geometrical properties one can be obtained from the other. The Monmonier's algorithm, described in the paper, computes barriers that lie on Voronoi tessellation since their edges are equidistant from pairs of populations.

**B:** The geometrical definition of the Voronoi diagrams implies that all points inside a polygon are closer to its centroid (black dots) than to any other. This implies that any triangle in the Delaunay triangulation (dotted lines) contains no populations and, therefore, can be inscribed in an empty circle. Populations (solid dots) lie on the circumference of the circle (circumcircle property).

### Figure 2

Example of computation of a barrier with the Monmonier's algorithm. Once a triangulation between populations (solid dots) is obtained (solid lines), the edges are associated with pairwise distance measures according to the used distance matrix (genetic, morphologic, etc.). Then the highest distance measure associated with the triangulation ("95" in the example) is taken as the starting edge of the first computing barrier. The barrier is extended always across the edge associated to the highest distance, "75" instead of "45"; then "78" instead of "65"; after that "56" instead of "40"; etc. The procedure is continued until the forming boundary has reached either the limits of the triangulation or closes on itself by forming a loop around a population. In case of multiple barriers, that are constructed one after another in a hierarchical order according to user settings, they can stop on a pre-existing barrier as does the barrier '2' of the figure.

We remind that the starting edge of a barrier can be either at a border of the triangulation (barrier '1') or inside it (barrier '2'); in the first case the extension takes place in only one direction, in the second case the extension takes place in two different directions as the arrows show.

### Figure 3

Human Y-chromosome differences around the Mediterranean basin.

**Upper:** A Delaunay triangulation (thin dotted lines) and the first genetic barrier (solid line) computed on a  $F_{ST}$  distance matrix between populations. (Redrawn from Manni et al. 2002)

**Bottom:** The Multidimensional Scaling (MDS) analysis representing the same  $F_{ST}$  distance matrix mentioned above. MDS takes a set of dissimilarities (as in a distance matrix) and returns a set of points such that distances between the points in the plot are approximately equal to the dissimilarities. (Redrawn from Manni et al. 2002)

### Figure 4

In this figure we present three different scenarios corresponding to a same multivariate analysis plot (ACP, MDS, etc.) where 18 populations belong to two different and non overlapping groups ("black" and "white"). The examples are meant to illustrate the essence of Monmonier's algorithm, that is to find barriers between populations. In the first example (**A**) all the populations belonging to the two clusters are spatially contiguous, therefore one main genetic barrier is expected to be present between them. In the second example (**B**) the two clusters can still be identified but only "white" populations are geographically contiguous, therefore two main genetic barriers are expected. The third example (**C**) illustrates the case of intermixed population where there is no correspondence between the geographic location of

population and their genetic (morphologic) differences. As a conclusion no main barriers are expected to occur.

### **Figure 5**

The more external part of a Voronoi tessellation (in blue), by definition, tend to infinite (B). It often happens that one of these borders is coupled with the highest genetic distance value of the matrix, therefore the origin of the barrier will take place outside the triangulation itself. We compare the first five barriers after (“A”) and before (“B”) adding the virtual points (small blue dots in “A”) that close the Voronoi tessellation. Virtual points can be considered as virtual populations that locally modify the neighborhood of the triangulation (not shown), thus being interpreted as the borders. In “B” barriers originate outside the triangulation in contrast with the definition of the method.

The example refers to surname differences in The Netherlands (see text and Fig. 6).

### **Figure 6**

Surname differences in The Netherlands. The analysis refers to the surname distribution of 226 localities. Differences were summarized in an overall distance matrix used to compute the first five barriers with the Monmonier’s method (yellow lines).

We also analyzed the first five barriers (in red) on 100 bootstrap matrices obtained by randomly resampling original surnames. The thickness of each edge of the barriers is proportional to the number of times it was included in one of the 500 computed barriers. In green the Delaunay triangulation, in blue the Voronoi tessellation. Small blues dots outside the triangulation are the “virtual points” used to close the Voronoi tessellation (see text).

Fig. 1

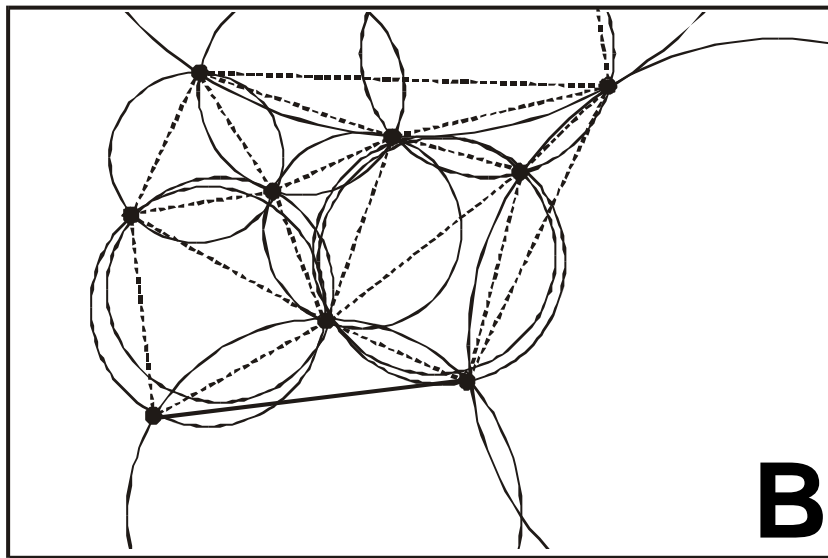
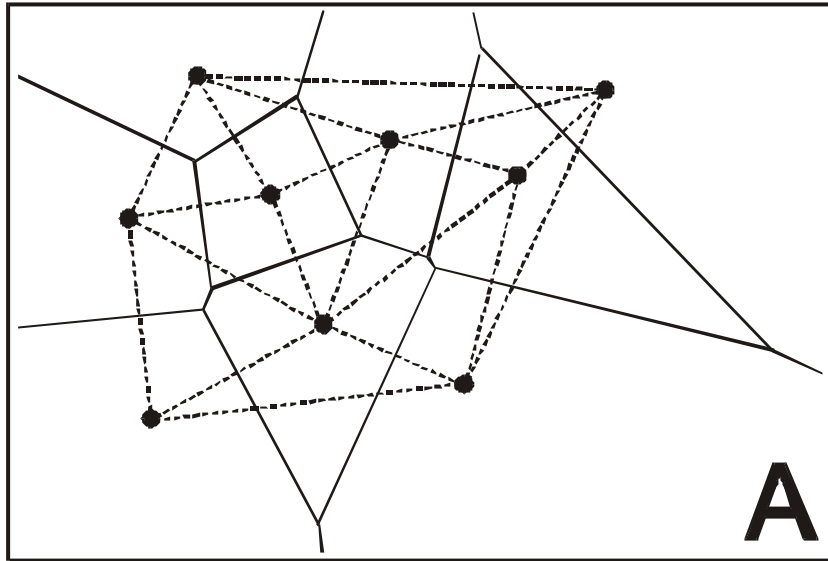
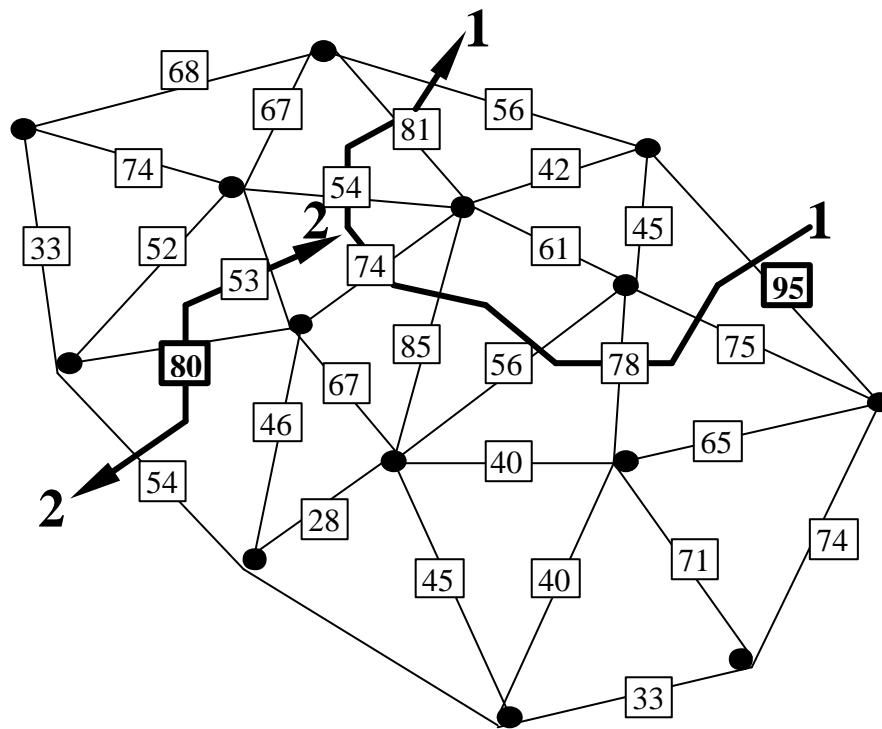
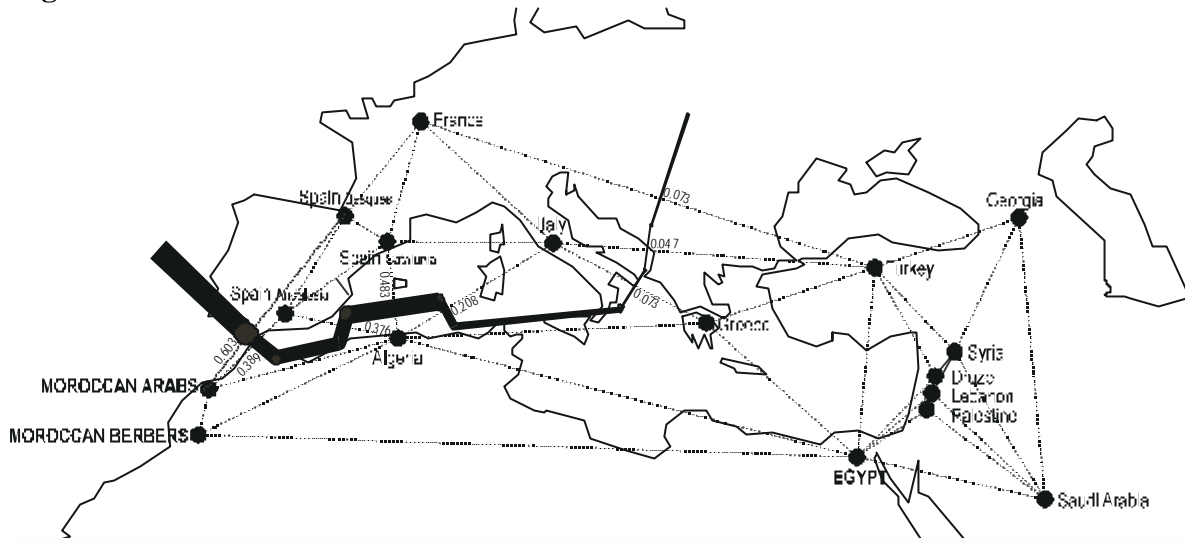


Fig. 2



**Fig. 3**



**Multidimensional Scaling scatterplot**

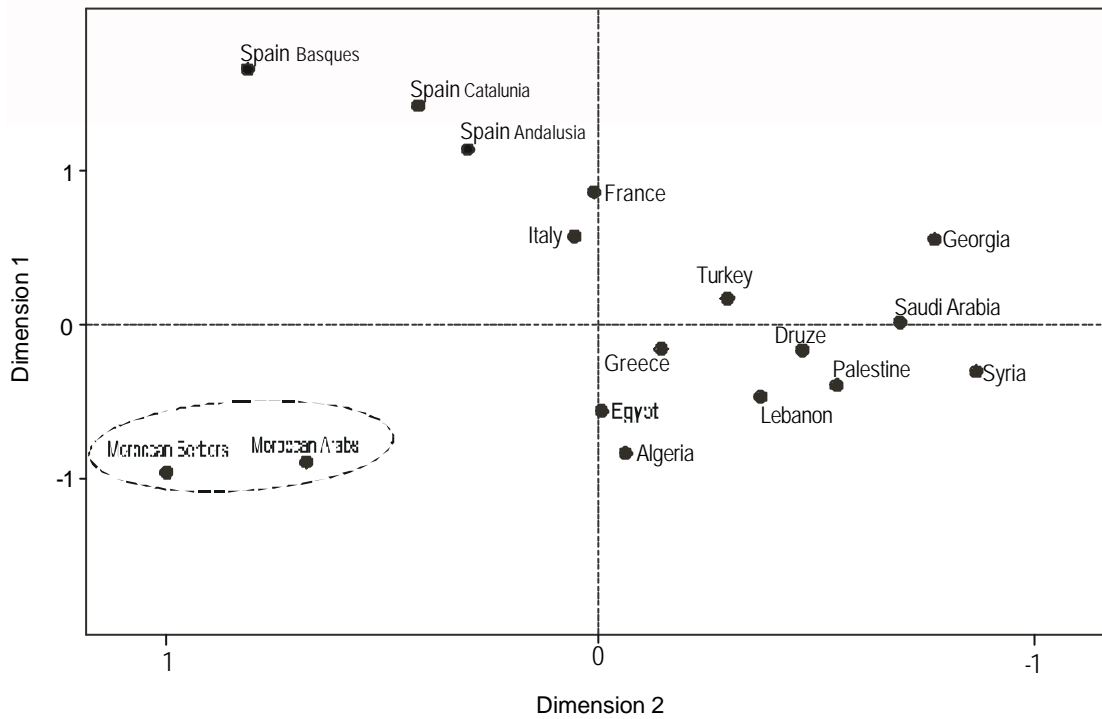


Fig. 4

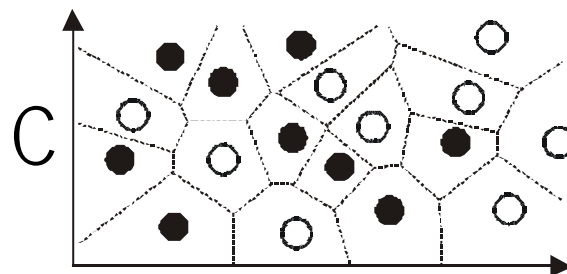
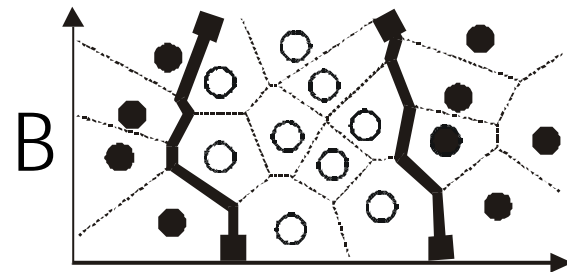
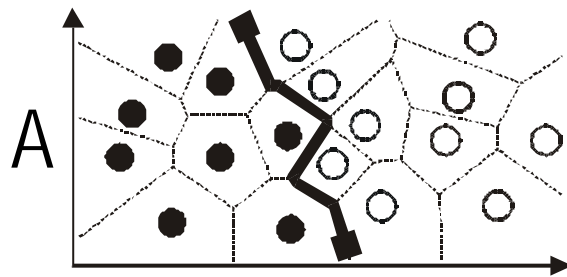
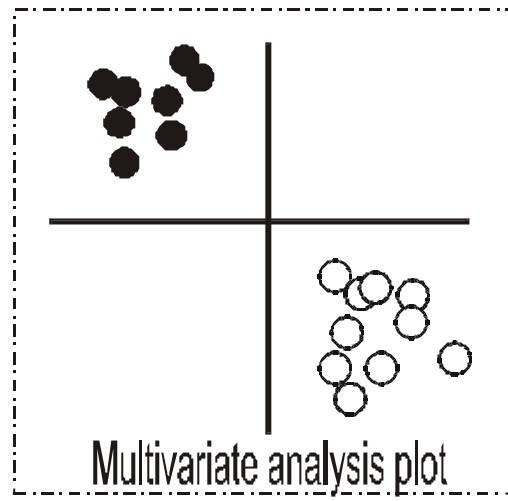


Fig.5

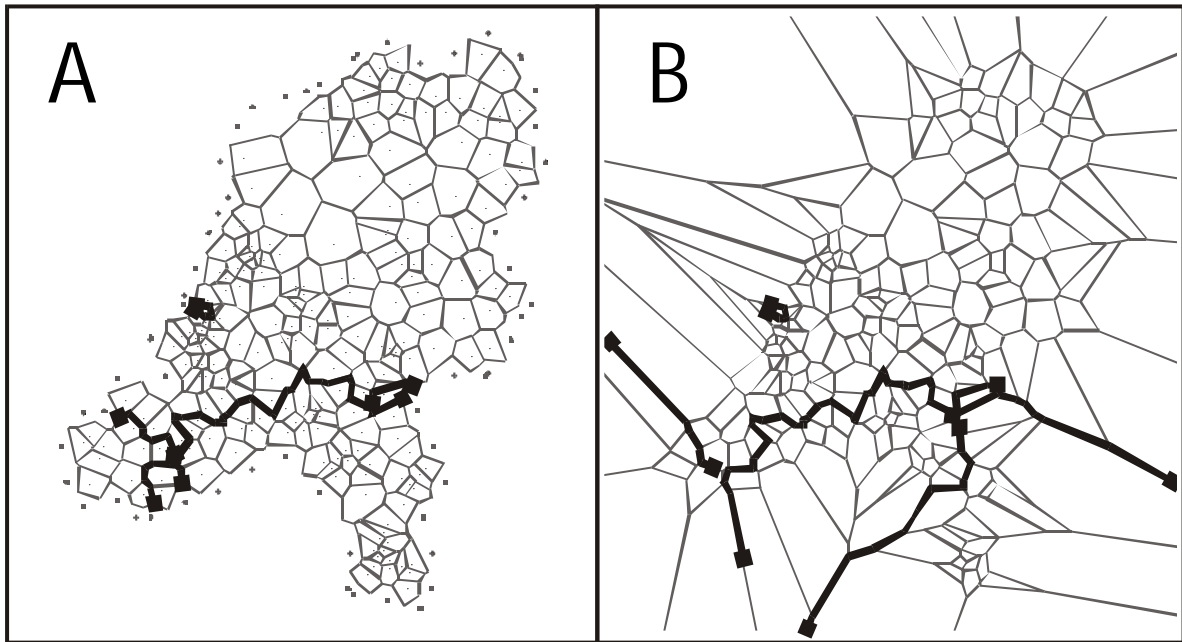


Fig. 6

