COMBINING BIOTECHNOLOGIES AND GISCIENCE TO CONTRIBUTE TO SHEEP AND GOAT GENETIC RESOURCES CONSERVATION

S. Joost* and the ECONOGENE Consortium**

*EPFL-LASIG, GR, Station 2, CH-1015 Lausanne, Switzerland, stephane.joost@epfl.ch
**http://lasig.epfl.ch/projects/econogene/

Summary
Geographic Information Science methods and tools are likely to help to extract useful and so far unknown information from large spatially explicit genetic datasets to understand the distribution of diversity among and within sheep and goat breeds. Considering the vast quantity of data collected within the Econogene project, exploratory data analysis methods were chosen as mean of investigation.

Key words
GIScience – biotechnology – geographic localization – exploratory analysis – visualization

Introduction

After the convention on biological diversity[1] was adopted in Rio in 1992 by a vast majority of the world's governments to conserve biological diversity, to use its components in a sustainable way, and to fairly share the benefits from the use of its genetic resources, FAO initiated a global strategy specifically designed to manage Farm Animal Genetic Resources (AnGR[2]). This action was mainly decided to face the threat constituted by intensive agriculture and its possible deleterious consequences on livestock populations. It consists in identifying and monitoring biodiversity in the different countries, to share the resulting data, and to integrate the conservation and sustainable use of biological resources into national decision-making. General purpose is to propose the introduction of adapted policies aiming at conserving livestock biodiversity. Among other different goals, countries have to identify and understand the genetic resources of each important farm animal species, and to prioritize and conserve unique AnGR.

Methods and techniques have been developed and improved to reach these goals. Since the beginning of the 1990's, the development of biotechnologies led to the elaboration of an array of different molecular techniques able to measure diversity at the DNA level [1] and molecular approaches have been progressively recognized to be appropriate tools to measure, monitor and manage genetic diversity [2]. Biotechnologies are continually evolving, offering new means to investigate the genome and providing additional information favoring the emergence of new hypotheses, and increasing the general intra-genome knowledge. The hypothesis defended here is that extra-organism information has to be taken into account to improve the understanding of genetic variation processes, this additional information being in our case a) the geographic localization of the studied animals, and b) the physical characteristics of their surrounding environment.

As a contribution to a EU program that aims at improving the sustainability of european agriculture, the Econogene project[3] was conceived to promote the sustainable conservation of

[1] http://www.biodiv.org
[2] http://www.fao.org/ag/cgrfa/AnGR.htm
[3] Quality of life V framework program, http://europa.eu.int/comm/research/quality-of-life.html

genetic resources in sheep and goats in marginal rural areas. In this context, it was precisely proposed to boost the contribution of biotechnologies to achieve genetic diversity management goals by combining their results with those of spatial analysis, shedding some light on potential genome-environment interaction processes and migration routes hypothesis. Thus this paper illustrates how Geographic Information Science (GIScience[4]) is likely to contribute in helping to extract useful and so far unknown information from large spatially explicit genetic datasets. It focuses on the choice of methods that have been selected to play a part within the analysis process aiming at improving livestock genetic ressources management.

GISCIENCE AND LANDSCAPE GENETICS

For many years, a GIScience current was directed towards environmental modeling [3], with the constant concern of explaining GIS basic features to show how they could be efficiently applied to natural sciences related fields [4].

Specifically considering genetics, the study of spatial structures exists since Wright (1931) developed adaptation models which were incorporating spatial distribution and distance issues [5]. Distance remains a central topic in spatial genetics as the main reference models directly refer to, or are constrained by it [5][6]. On this basis, GIS were introduced to develop dispersal models to simulate animal population migrations in the landscape [7], or to provide tools for visualization and analysis of geographic population structures [8]. Spatial analysis methods like kriging were used to define diversity zones [8][9]. Hamann *et al.* [10] also exploited kriging to detect areas of genetic differentiation. Directly related to the third method described in this paper, Skøt *et al.* [11] investigated interaction between environmental characteristics and AFLP markers, and so did Pakniyat *et al.* [12] to show association with salt tolerance in wild barley[5].

Using GIS "to place genetic diversity information into a spatial framework" is an approach refered as *landscape genetics* by Dave Galbraith (Royal Botanical Gardens, cited by [13]). The Natural Resources DNA Profiling and Forensic Centre in Peterborough led several studies since the second part of the 1990s' in landscape genetics applied to the canadian's fauna [13]. In 2003, Manel et al. [14] mention landscape genetics as a new approach described as a combination of landscape ecology and population genetics, which is likely "to facilitate our understanding of how geographical and environmental features structure genetic variation at both the population and individual levels, and has implications for ecology, evolution and conservation biology". "Landscape genetics" is becoming a widespread designation to include all research about genetic data and exploiting their geographic dimension, this being confirmed by recent papers [15][16]. This is definitively the domain to which the present research is contributing.

SPATIAL DIMENSION OF GENETIC DATA

The goal of combining spatial analysis with biotechnologies is to increase the power of the latter by exploiting the spatial dimension of the information they provide. GIScience is likely to highlight patterns in the spatial distribution of diversity values, and to discover simultaneousnesses or relationships between genome characteristics and the properties of the environment. This domain offers multiple ways for analyzing spatial information, each of them being closely linked up to a geographic database, the heart of GIS. The set is vast, from

[4] Geographic Information Science is the set of methods and tools (among which Geographic Information Systems) conceived to analyze, manage, use geographic information.

[5] This type of analysis is a scientific field called *Genecology* by the Norsk Institutt for Genøkologi (GenØk[5]) in Tromsø.

elementary functions (consultation via a map interface, database (spatial) requests, representation, etc.) to advanced ones related to spatial processes modeling.
Though many GIScience aspects are involved in the context of the Econogene project, this paper mainly emphasizes exploratory analysis and data representation. Use of geographic information is likely a) to show diversity values on maps to assess potential spatial patterns, b) to compare diversity values provided by the different molecular markers with regard to the geographic location and check for congruence, c) if not congruent, to compare values and expressed diversity patterns, d) to identify particular behaviour of given populations on different scales, and e) to help identifying genes under selection.

EXPLORATORY ANALYSIS

To identify conservation priorities, Econogene teams have collected biological samples from over 3'000 animals spreaded out from Portugal to eastern Turkey. For each animal, 5 to 10 variables per molecular marker were made available. This genetic information has to be compared with more than 100 environmental variables likely to make any interesting relationship emerge. Considering this huge quantity of information, and before trying to formulate initial rough working hypothesis outlines, it is necessary to wander [17] among those large datasets in order to discover information from data. The exploratory analysis (EDA) field was first defined in the John Tukey Exploratory Data Analysis book (cited in [18]) in 1977. It consists of an approach which employs a variety of mostly graphical techniques in order to maximize insight into a data set (uncover underlying structure, extract important variables, detect outliers and anomalies, etc. [19]). Instead of looking for a known model and checking if data is conform, EDA proposes a more direct approach of allowing data itself to reveal its underlying structure. EDA is mainly resorting to graphical techniques for the reason that its main role is to open-mindedly explore data. Visualization of graphics provides matchless power to do so, making it possible to discover structural secrets, and to gain some new unsuspected insight into the data. Scientific visualization was first used as an unformal way to analyze information (Unwin cited in [20]) until it was recognized as a scientific method by the end of the 1980's [21]. While being at it, and on the basis of EDA, a complementary approach emerged to exploit the spatial dimension of data, when available. Exploratory Spatial Data Analysis (ESDA) tools include additional methods elaborated to take into account the specificities of geographic information [17][18][21][22].
The continuation of the paper presents three different ESDA methods which were used to explore genetic diversity data within an explicit spatial context. They are described in the next 3 parts according to a complexity order.

Cartographic visualization

Like EDA often resorts to the visualization of graphics to efficiently investigate important datasets, maps are exploiting human cognition features which are recognized to be essentially sensitive to spatial processes [23]. Thus maps are particularly well suited to stimulate creative thinking by generating mental imagery in the analyst's brain, depending on their personal culture and background.
However, while several sophisticated developments have led to the elaboration of advanced interactive geographic visualization tools, one could believe that the old-fashioned cartography has become totally out of interest. Wood [23] has listed and well argued numerous points to make us still consider static maps as valuable tools, main aspects being that cartography allows to fix and store interesting views, and is appropriate for map design.

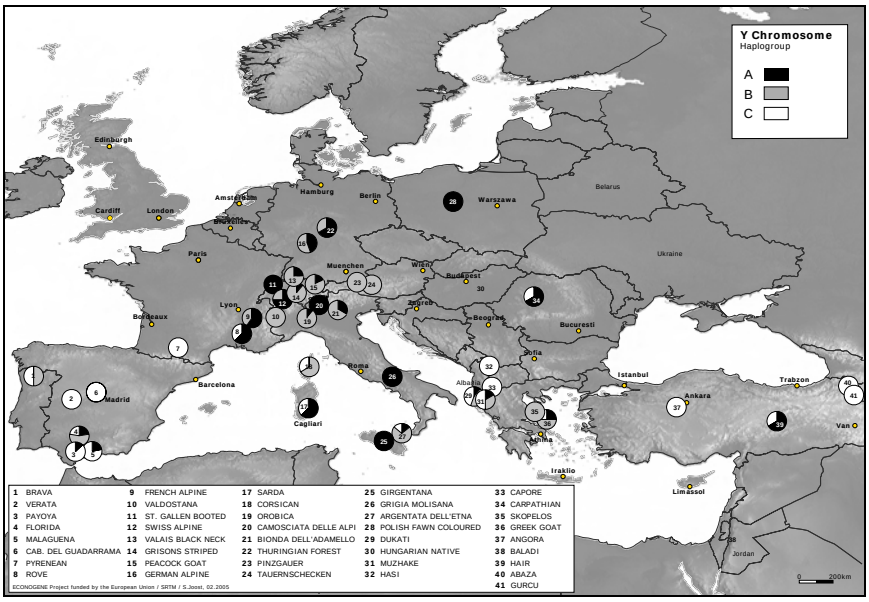

Figure 1 : Goat breeds populations over the Econogene study area and spatial distribution of Y chromosome haplogroup diversity. The position of a centroid is the center of gravity of a breed population sampled in several farms. A digital elevation model provides the main geographic points of reference.

Map design or providing a geographic context

Most of spatial representations realized so far in geographic genetics (for instance [5] p.191, [24] p.306) have recourse to a neutral geographic space, strongly suggesting that only the general location and an approximative distance between symbols is important. Of course, this depends on the goal the map is expected to reach, as one may only wants to focus on a simple function like to compare haplotypes and showing the relationship between diversity and geographic location ([2], p.907). But often are cartographic representations unintentionally limited.

Geographic representations of genetic data are produced because spatial processes are supposed to be explanatory. Then it seems consistent to use available contextual spatial objects at best. Nowadays, as representation technologies henceforth do allow it, giving a concrete expression of land may be invaluable to improve cartographic representation of genetic data, with the constant concern of keeping a high level of readability. This permits to anchor a phenomenon in the landscape and helps analysts to understand a spatial distribution of data and to produce new working hypotheses. Contextual objects may be on the one hand relief (figure 1), forests, rivers, etc., that is to say natural landscape components, and on the other hand anthropic objects like roads, railways, etc. Depending on the working scale, both are likely to play the role of barrier and supply explanatory elements when analyzing their respective position with that of the animals. These points of reference are helpful to locate and inlay observed objects into the geographic space. This has the considerable advantage to reduce the analyst's first intuitive intellectual effort mobilized to locate an object before being able to initiate the visual thinking process to make research hypothesis emerge [23][25].

Representation of genetic information

To make thematic mapping efficient to visualize genetic data, different choices were made to shape the maps. Notably interpolation was used as an artefact to enforce the visual impact of the spatial distribution of diversity measures, to create a continuity that facilitates its rendering. Brodlie [26] mentions this approach as a possible sequence of processes, or "visualization pipeline": a) interpolate scattered data on to a grid, b) generate contour map

from grid, c) render contour map (figure 2). We insist on the fact that this method is not applied in a predictive way, but to help to apprehend a spatial distribution of a phenomenon. Real and trustable information is only contained within breed centroids. Another point is that this reprentation technique allows the superimposition of two different variables to compare them and check for congruence in diversity indicators for instance.
The representation of genetic data was made in order to distinguish at best diversity values variation. Reaching this goal is a subtle balance between the choice of the most adapted discretization method and the most adapted number of classes on one side, and the most adapted choice of colours and hue on the other side, to fit crucial known criteria about visual perception [18]. This exercise is much more restrictive when using greytone maps, and the views produced in this paper (figures 1 and 2) are thus rather basic and less effective to provide a quick apprehension of the analyzed phenomenon.

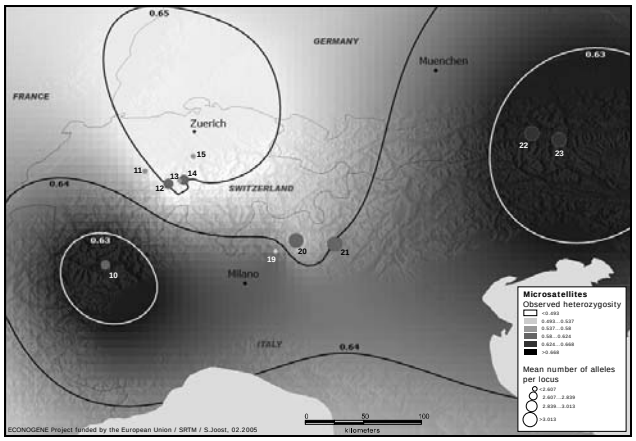

Figure 2 : Goat breeds populations in the Alps. This map illustrates the visualization pipeline mentioned by Brodlie [26]. Centroids are expressing observed heterozygosity (hue) and mean number of alleles per locus (diameter) while the grid provides interpolated values of AFLP frequency of recessive genotype : the darker the cell, the higher the diversity.Contour lines are generated from the interpolated values.

Geographic visualization (GVIS)

For the last 30 years, cartography gradually had to deal with an increasing number of data sources which were becoming larger and larger. Developments in GIS made it possible “to rejoin data storage with display” [25], transforming traditional maps into real interfaces able to support “knowledge construction activities” [25], while keeping their representation function. So emerged a “modern cartography” [25] likely to face the changes occurring in geographic information management and analysis. Geovisualization is an approach stemed from these developments and offering dynamic and interactive access to geodata, fitted to facilitate search for unknowns, information exploration and finally knowledge construction in the absence of pre-determined hypotheses.
In practical terms, GVIS tools are providing interactivity in the sense that they allow users to choose and visualize different variables to assess their simultaneous variation, together with a constant access to the spatial location of the considered objects, in order to facilitate visual thinking. An interactive link is established between the geographic representation of analyzed objects and the genetic information they contain (figure 3). Compared with thematic mapping, GVIS softwares are more powerful to investigate and visualize data, as it is possible to find a value, to see the corresponding location on a map, and to get the values of all other variables describing the breed or the environment at this place. But representation features are not as

extensive as those of cartography and this makes GVIS essential tools to be used upstream, during the first steps of the scientific reasoning.
GVIS tools and visual approaches to data mining are well adapted to the Econogene context as the amount of data collected in the genetics, economics, and environmental fields is sizeable, moreover considering the important number of animals sampled and the large extent of the study area.

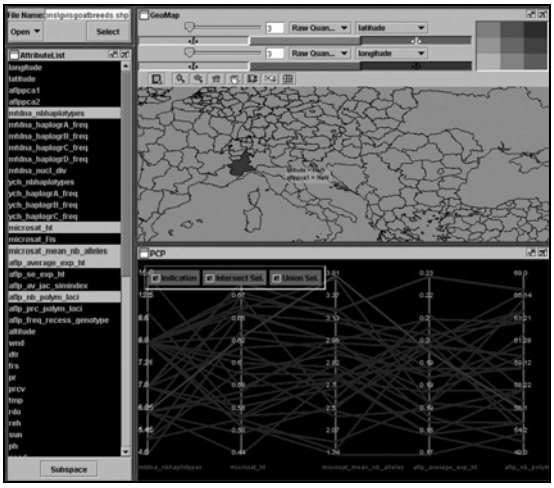

Figure 3 : GVIS softwares are providing powerful tools allowing to display simultaneously selected variables and assess their variation according to the spatial location. Here Geovista Studio [27] displays a list of available variables on the left. In the lower part, a Parallel Coordinate Plots (PCP) shows the values of different variables for a given breed (highlighted area on the map): lines representing breeds in the PCP have a color which is used on the map to point out its corresponding area.

Genome-environment interaction assessment

This approach is different as visualization is not directly involved. GIS is used to extract values of topo-climatic variables constituting the surrounding environment of the sampled animals (altitude, climate). The method to reveal associations consists in using the genotyping of the sampled breeds and to look for markers whose frequency is correlated with selected environmental variables. Univariate logistic regression models [28] are run to assess interaction between environmental characteristics and genetic data. The logit link is of the simple form $g(x) = \alpha + \beta x$ where the significance of the β coefficient is assessed using the Wald test [28]. The statistics of this test may belong to the normal distribution or not, and thus the null hypothesis (H0) being rejected or not. This information is used to make up tables containing H0 test results for all models, allowing first to visualize the global response of molecular markers to environmental stimulus, and then to possibly investigate significant relationships with numerical methods. The high number of processed models (more than 7'000 in the example mentioned hereunder) justifies the use of “rejection” tables (figure 4) where possible structure may be detected.
This methodology was experimented on sheep with AFLP markers, allowing to highlight 10 AFLP markers probably sensible to environmental stimulus [29]. Though AFLP are neutral markers, they can be used to check potential association with environmental parameters and thus to locate genes under selection according to a population genomics approach [30]. This method will soon be applied to other molecular markers available for the same animals (microsatellites and SNPs), and comparing results will most likely provide a reference allowing to evaluate the approach.

ID		1	2	3	4	5	6	48	49	...	50	51	52	53	54	55	56	57	58	59	60	61	62		
		E35T32_25	E35T32_22	E45T36_9	E35T32_32	E35T32_28	E45T38_7	E45T38_19	E35T32_26		E45T38_24	E35T38_16	E45T38_1	E45T38_17	E35T32_27	E35T38_5	E35T32_29	E45T38_30	E45T38_31	E35T32_23	E45T38_34	E45T38_25	E35T32_10	#signif. models	%
1	altitude	0	0	0	0	0	0	0	0	...	0	0	0	1	1	0	0	0	0	0	0	0	0	4	6.5
2	wndjan	0	0	0	1	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0	0	1	0	9	14.5
3	wndfeb	0	0	0	1	0	0	0	0	...	0	1	0	1	0	0	0	0	0	0	0	1	0	8	12.9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	5	8.1
117	sundec	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0	0	0	0	5	8.1
118	sunyear	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0	0	0	0	7	11.3
	#signif. models	3	1	35	57	9	0	0	2	...	11	95	1	10	1	0	1	1	4	1	0	37	0	736	
	%	2.5	0.8	29.7	48.3	7.6	0.0	0.0	1.7	...	9.3	80.5	0.8	8.5	0.8	0.0	0.8	0.8	3.4	0.8	0.0	31.4	0.0		

Synthetic view of over 7'300 models results →

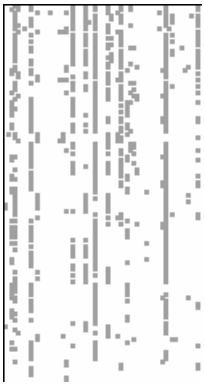

Figure 4 : On the left, detail of a "rejection" table in which each cell contains the result of a H0 test. "1" are indicating significant models. On the right, visual display of a full result table. The vertical structure suggests that AFLP markers' response is often not exclusive to specific environmental variables.

Conclusion

GIScience can offer complementary ways to analyze genetic data by enhancing their spatial dimension. In the absence of pre-determined hypotheses, exploratory analysis approaches are well suited to carry out this task. Cartography allows to inlay spatial patterns of genetic diversity with explicit geographic references; geovisualization makes it possible to interact dynamically with large amounts of different kinds of data while constantly refering to the location of studied populations; spatial analysis permits to study interaction between characteristics of the environment and those of the genome. All of them are supplying means to investigate the unknown, to extract information, and to construct knowledge about how genetic diversity is spatially distributed and why.

Aknowledgments

This work has been supported by the European Commission (Econogene contract QLK5-CT-2001-02461). The content of the publication does not represent the views of the Commission or its services. I would like to thank Elisabetta Milanesi, Marco Pellecchia, Riccardo Negrini, Régis Caloz and Joël Chételat for their valuable comments.

REFERENCE LIST

[1] Karp, A., Edwards, K.J., Bruford, M., Funk, S., Vosman, B., Morgante, M., Seberg, O., Kremer, A., Boursot, P., Arctander, P., Tautz, D., Hewitt, G.M. 1997 Molecular technologies for biodiversity evaluation: opportunities and challenges. Nature Biotechnology, Vol.5, pp 625-628.

[2] Bruford, M.W., Bradley, D.G., Luikart, G. 2003 DNA markers reveal the complexity of livestock domestication. Nature, Vol.4, No.11, pp 900-910.

[3] Goodchild, M.F., Parks, B.O., Steyaert, L.T. 1993 Environmental Modeling with GIS. Oxford University Press.

[4] Caloz, R., Collet, C. 1997 Geographic information systems (GIS) and remote sensing in aquatic botany: methodological aspects. Aquatic botany, Vol.58, No.3, pp 209-228.

[5] Epperson, K.E. 2003 Geographical Genetics. Princeton University Press.

[6] MacArthur, R.H., Wilson, E.O. 2001 The theory of island biogeography. Princeton University Press.

[7] Vuilleumier, S. 2003 Dispersal modelling : integrating landscape features, behaviour and metapopulations. Phd Thesis, SSIE section, EPFL), No.2878

[8] Hoffmann MH, Glass AS, Tomiuk J, Schmuths H, Fritsch RM, Bachmann K 2003 Analysis of molecular data of Arabidopsis thaliana (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS). Molecular Ecology, Vol.12, No.4, pp 1007-1019.

[9] Bucci, G., Vendramin, G.G. 2000 Delineation of genetic zones in the European Norway spruce natural range: preliminary evidence. Molecular Ecology, Vol.9, No.7, 923-934.

[10] Hamann, A., Koshy, M.P., Namkoong, G., Ying, C.C. 2000 Genotype x environment interactions in Alnus rubra: developing seed zones and seed-transfer guidelines with spatial statistics and GIS. Forest Ecology and Management, Vol.136, pp 107-119.
[11] Skøt, L., Hamilton, N.R.S., Mizen, S., *et al.* 2002 Molecular genecology of temperature response in Lolium perenne: 2. association of AFLP markers with ecogeography, Molecular Ecology Vol.11, No.9, pp 1865-1876.
[12] Pakniyat, H., Powell, W., Baird, E., Handley, L.L., Robinson, D., Scrimgeour, C.M., Nevo, E., Hackett, C.A., Caligari, P.D.S., Forster, B.P. 1997 AFLP variation in wild barley (Hordeum spontaneum C. Koch) with reference to salt tolerance and associated ecogeography . Genome, Vol.40, No 3, pp 332-341.
[13] Natural Resources DNA Profiling and Forensic Centre, Biology Department, Trent University, Peterborough, http://www.nrdpfc.ca/landscapegenetics.html
[14] Manel, S., Schwartz, M.K., Luikart, G., Taberlet, P. 2003 Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology and Evolution, Vol.18, No.4.
[15] Hirao, A.S., Kudo, G. 2004 Landscape genetics of alpine-snowbed plants: comparisons along geographic and snowmelt gradients. Heredity, Vol.93, No.3, pp 290-298.
[16] Watts, P.C, Rouquette, J. R., Saccheri, I. J., Kemp, S.J., Thompson, D. J. 2004 Molecular and ecological evidence for small-scale isolation by distance in an endangered damselfly, *Coenagrion mercuriale.* Molecular Ecology, No.13, pp 2931-2945.
[17] Banos, A. 2001 A propos de l'analyse spatiale exploratoire des données. Cybergeo, No. 197, http://193.55.107.45/MODELIS/banos/article.htm
[18] MacEachren, A.M. 1995 How maps work, Representation, visualization and design. The Guilford Press.
[19] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/
[20] Tobon, C. 2001 Visual and interactive exploration of point data. Centre for Advanced Spatial Analysis, UCL, Working Paper SeriesPaper 31.
[21] MacEachren, A.M. 1994 Visualization in modern cartography, setting the agenda. In MacEachren, A.M. and Fraser Taylor, D.R. (ed.). Visualization in modern cartography. Pergamon, pp 1-12.
[22] Haining, R. 2003 Spatial data analyis, theory and practice. Cambridge.
[23] Wood, M., 1994 The traditional map as a visualization technique. In Hearnshaw, H.M., and Unwin, D.J.(ed.). Visualization in Geographical Information Systems. Wiley, pp 9-17.
[24] Petit, R.J., Bialozy, R., Brewer, S., Cheddadi, R., Comps, B. 2001 From spatial patterns of genetic diversity to postglacial migration processes in forest trees. In Silvertown, J. and Antonovics, J. (ed.). Integrating ecology and evolution in a spatial context. British Ecological Society, pp 295-318.
[25] MacEachren, A.M., Kraak, M.-J. 2001 Research challenges in geovisualization, Cartography and Geographic Information Science, Vol.28, No.1.
[26] Brodlie, K. 1994 A typology for scientific visualization. In Hearnshaw, H.M., and Unwin, D.J.(ed.). Visualization in Geographical Information Systems. Wiley, pp 34-41.
[27] Takatsuka, M., Gahegan, M. 2001 GeoVISTA Studio: A Codeless Visual Programming Environment For Geoscientific Data Analysis and Visualization. The Journal of Computers & Geosciences, Vol.5, No.2.
[28] Hosmer, D.W., Lemeshow, S. 2000 Applied logistic regression. John Wiley & Sons, New York.
[29] Joost, S., Econogene Consortium. 2005 Combining biotechnologies and GIScience for livestock genetic resources conservation. Proceedings of the 8th AGILE Conference on GIScience, accepted.
[30] Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P. 2003 The power and promise of population genomics: from genotyping to genome typing. Nature, Vol.4, No.12, pp 981-994.