Characterization, Delineation and Visualization of Agro-Ecozones Using Multivariate Geographical Clustering

Annamaria Castrignanò*, Daniela De Benedetto, Giacoma Girone,
Francesca Guastaferro, Donato Sollitto

CRA – Unità di Ricerca per i Sistemi Colturali degli Ambienti caldo-aridi
Via Celso Ulpiani 5, 70125, Bari, Italy

Received: 6 October 2009. Accepted: 13 November 2009.

Abstract

Agro-ecozoning is a delineation of landscape into relatively homogeneous regions of expected similar crop performance. Past classifications have been subjective, crop specific and did not take into account spatial correlation. A quantitative approach is proposed to unambiguously locate, characterise and visualise agro-ecozones and their boundaries which can be allied to different environmental conditions. In this study the environmental parameters, including climatic and soil characteristics, hypothesized to be generally relevant to many crops in Capitanata-Foggia (South Italy), were used. Cokriged environmental estimates at 500 m scale were used in a clustering algorithm based on non-parametric multivariate density estimation. A 3D map of density estimation and red-green-blue colour triplet were used for visualisation of agro-ecozones as a unique combination of environmental factors.

The proposed approach produced the delineation of the study area in five compact classes in the space of environmental attributes that were also contiguous in geographic space. The resulting agro-ecozones may provide a framework for useful application in land use decision making.

Key-words: agro-ecozone (AEZ), management zone, geostatistics, fuzzy c-means classification.

1. Introduction

The delineation of landscape in regions where crop performance is expected relatively homogeneous may have potential benefits for improved agricultural production and natural resource conservation. Agro-ecozones (AEZs) are geographic units containing similar land resource potentials and limitations relevant to agriculture (Williams et al., 2008). In the past AEZ delineation has been crop-specific (FAO, 1996; Caldiz et al., 2001; Swinton et al., 2001), using detailed information on crop requirements and human expertise in a qualitative, weight-of-evidence approach (McMahon et al., 2001). However, when regionalisation mostly depends on observer interpretations and on his personal experience, it is not suitable for statistical extrapolation (Metzger et al., 2005), because science requires transferability and repeatability of the results and then the use of more objective, quantitative models. However, quantitative regionalisation techniques will not be perfectly objective, because they require expertise in the choice of data layers to include and in the interpretation of the resulting agro-ecozones. Nevertheless, the definition of a quantitative approach for regionalisation is a desirable goal, because it allows, among the other things, to custom the input variables to specific uses and has broad application to numerous crops, including potential alternative crops. Moreover, the variable nature of the boundaries between agro-ecozones may add further ambiguity to boundary location and meaning. Actually, boundaries between agro-ecozones can be sharp or more commonly gradual, which causes edges to be indistinct and makes difficult to locate a line of demarcation between distinct regions, or can change their characteristics along their length. Approaches to distinguish these different types of border have been used in the

* Corresponding Author: Tel. +39 080 5475024; Fax: +39 080 5475023. E-mail address: annamaria.castrignano@entecra.it

past, including fuzzy set theory (Leung, 1987; Lark, 1998), wavelet analysis (Csillag and Kabos, 2002) and multivariate geographic clustering (Hargrove and Hoffmann, 1999), but no single method has been widely adopted. Locating agro-ecozone boundaries is a multivariate process, which requires to analyse large geographical data sets for the different environmental conditions. Growth in computing power and increased availability of spatial environmental data in geographic information systems (GIS) have made agro-ecozoning feasible.

However, though GISs allow the use of spatial data in a digital environment and integration of data from different sources, and separate scales and have several application in agricultural research, their use is no guarantee of objectivity if a quantitative analytical approach is not defined to delineate AEZs.

Statistical clustering is a well known technique which groups similar individuals into distinct classes in attribute space, called clusters (Jensen, 1996; Irvin et al., 1997). Several investigators have used multivariate clustering for delineating homogeneous climatic and physiographic regions (Host et al., 1996), uniform regions of geology (Harff et al., 1990), regions of uniform crop (Lark, 1998) and regions of constant fertility (Carter et al., 1997). At present several algorithm options exist but no unified theory is widely accepted.

Most current clustering methods, based on least-squares criterion (Sarle, 1982), are biased, because they tend to fit globular clusters of equal size in data space, characterised by a similar upper limit on within-group variance and a similar maximum radius around each centroid. Therefore, the uniform heterogeneity across clusters prevents the creation of regions with highly elongated or irregular shapes and vastly changeable within-unit variance. Conversely, the methods based on nonparametric density estimation are the ones with the least bias (Silverman, 1986; Scott, 1992) and according to such an approach a cluster is defined as a region surrounding a local maximum of probability density function or a connected set of local maxima.

Furthermore, existing traditional clustering techniques do not account for the spatial correlation between observations and take little account of gradual change, either from one class to another or within any one class, because it is assumed that the variability of most properties is less within clusters than between clusters. However, where high variable levels of management are applied, within-unit variation may exceed that between units. In such conditions it is very difficult to unambiguously associate the boundaries of the delineated clusters with important changes of landscape. Geostatistics uses a completely different paradigm, because it treats multivariate indices of spatial variation as continua in a joint attribute and geographical space. In the univariate form each attribute is considered as a random regionalized variable, varying continuously and its gradual geographical variation is described by a covariance function. In order to obtain geographically continuous regions, proximal information of the cells then has to be directly used in the geographical classification of agro-ecozones.

Finally, density estimation is now recognized as a powerful graphical tool for detecting and summarizing the multivariate structure of complex data. As each cluster's centroid, obtained by averaging on all the point in the cluster, is assumed as representative of the cluster region, the Euclidean distance from each cell to its centroid measures its deviation from the cluster norm. Therefore, cells close to their centroids are more representative of the cluster than cells for from their centroids in environmental space (Belbin, 1993). In a 3D representation of probability function, hypothetical clusters appear as a series of peaks with border regions tracing along the lowest geographic locations. Visualization then is a key aspect of effective multivariate analysis but requires a proper array of both statistical and graphical tools.

The objective of this paper is to propose a multivariate statistical clustering approach to delineate and visualise agro-ecozones. The proposed approach is a combination of geostatistical techniques with a non-parametric density algorithm and a red-green-blue colour triplet.

2. Materials and methods

2.1 Study area

The study area extends for about 1.979 km^2 and is located in "Capitanata" plain (Apulia region, Southern Italy). "Capitanata" plain is the northern sector of the Apennines foredeep, a geological structure delimited by the Apennines Chain West and by Gargano Promontory East. The plain is mainly constituted by continental and fluvial sediments and some terraced marine deposits of Pliocene and Pleistocene ages. The area is characterized by a flat morphology, with topographic elevation values ranging between 50 m and 200 m a.s.l. On the west side of the plain, the elevation increases gently toward the Apennine piedmont region, where the morphology becomes hilly and the region is crossed by not very deep river valleys. Rivers are characterized by a torrential regime and cross the study area forming meanders and braided landscapes.

According to the Atlas of the Soil Regions of Italy (Centro Nazionale di Cartografia Pedologica, 2002) the main pedological region in "Capitanata" is the 62.1 type, with a typical Mediterranean subtropical climate, characterized by an average annual air temperature between 12 and 17 °C and an average annual precipitation between 400 and 800 mm distributed mainly during the autumn period (October-November).

The geological and geomorphological settings of the region cause the variability of the pedological and micro-climate conditions, determining the presence of different soil sub-regions. Different soil sub-regions are delineated in the study area related to the local morphological and hydraulic conditions (Apulia Region, 2001): higher Capitanata, lower Capitanata, southern Capitanata, Fortore Valley and middle-west Gargano.

The study area is characterized by different soil types such as Vertisols, Calcisols, Kastanozems-Calcisols and Vertisols-Cambisols (Apulia Region, 2001).

2.2 Environmental dataset

We defined a quite flexible approach in regionalisation of land resources by using the environmental variables hypothesized to be potentially limiting growth and production of a wide variety of crops. The choice of variables (climatic, topographic and edaphic) was based on Loomis and Connor (1992) and FAO (1996). Elevation was not included because it varies very little within the flat study area. Three climatic properties were considered: total monthly precipitation, monthly minimum temperature and monthly maximum temperature. They were derived from daily weather observations for the period 2001-2007 at 14 weather stations distributed across the study area and provided by the "Consorzio per la Bonifica della Capitanata" (Consortium for Capitanata Reclamation) (Fig. 1).

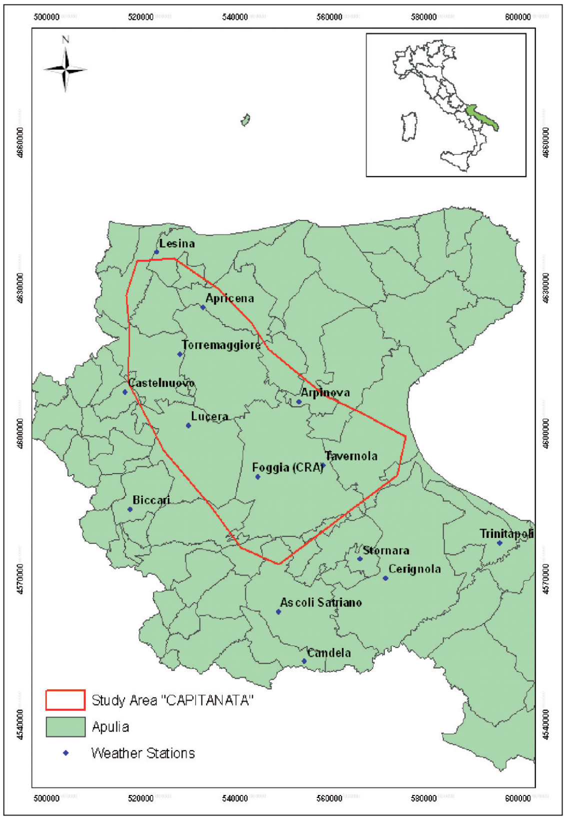

Figure 1. Study area and locations of the weather stations.

Because of the small number of weather stations inside the study area, some external weather stations were included in the dataset. Monthly means and standard deviations of climatic properties were calculated for the period of observation. Monthly means of climatic properties at each observation station were then used to interpolate a raster surface of 500 m x 500 m resolution using inverse squared distance weighting method (ArcGIS 9.1, ESRI, 2005) because the few observation stations precluded the application of geostatistical techniques for spatial interpolation. To reduce the number of climatic variables, the normalised variable values for each raster cell were used in a Principal

Component Analysis (PCA) using the FACTOR procedure in SAS (SAS Institute Inc. release 9.1.3, 2009), where the 36 components were retained to create 36-dimensional data space. Using VARIMAX rotation, the 36 components were transformed into 36 orthogonal axes in the climate data space. Only the PCs explaining most of total variance were retained for successive analysis of clustering.

Seven soil properties were used: clay (%), silt (%), sand (%), field capacity (FC) (%), permanent wilting point (PA) (%), pH (-) and organic matter (OM) (%). Soil data (749) were obtained from the soil properties database of Apulia Region (Regione Puglia, 2001) and other datasets produced in various public research projects (Fig. 2).

As the datasets contained values for the selected soil characteristics from different profile depths, the values were averaged over the two depths: 0-0.4 m and 0.4-2 m.

2.3 Geostatistical and clustering analyses

The geostatistical technique used for spatial interpolation was cokriging (Goovaerts, 1997). The application of cokriging requires modelling the coregionalization of the set of variables using the Linear Model of Coregionalization (LMC) developed by Journel and Huijbregts (1978). The LMC considers all the studied variables as the result of the same independent physical processes, acting on different spatial scales. The $p(p+1)/2$ simple and cross variograms of the p variables are modelled by a linear combination of N_s standardized variograms to unit sill. Fitting[1] of LMC is performed by weighted least-squares approximation under the constraint of positive semi-definiteness of the matrix of sills (coregionalization matrix), using an iterative procedure (Lajaunie and Béhaxétéguy, 1989). Finally, the soil variables were then interpolated on a 500 by 500 m-grid. As silt is statistically linked to clay and sand, the coregionalization matrix was not positive semi – definite and silt was then excluded from the geostatistical analysis; and its estimates were calculated as complementary to 100 of the sum clay + sand.

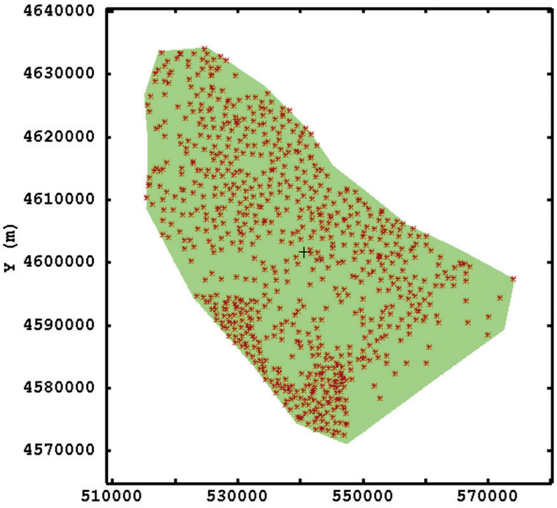

Figure 2. Location of soil properties data.

All the geostatistical analyses were carried out using ISATIS® software package (Geovariances, 2009, release 9.1).

To divide the study area into a number of AEZ without any previous information about the existence and the number of the groups, an algorithm, based on nonparametric density estimate, was used (Silverman, 1986; Scott, 1992).

2.4 Density estimation

The approach utilises hyperspherical uniform kernels of fixed radius to estimate density. The density estimation at any point of the attribute hyperspace is computed by dividing the number of observations, within a sphere centred at the point, by the product of the sample size by the volume of the sphere. The size of the sphere is determined by a smoothing parameter (R) to be pre-specified, which represents the kernel radius and is expressed as a Euclidean distance. The sphere of support of the kernel at observation x_i is referred as the neighbourhood of x_i and the observations within the neighbourhood are referred as the neighbours of x_i. Therefore, the estimated density at x_i is given by:

$$\hat{f}_i = \frac{n_i}{nv_i}$$

where $\hat{f}$ is the estimated density, n_i the number of neighbours of x_i, nv_i the sample size and the volume of the neighbourhood. There is no

[1] To fit the linear model of coregionalization, it needs to decompose each coregionalization matrix into an orthogonal matrix by using the diagonal matrix of eigenvalues. A brief determination of eigenvalues is given in Davis (1986, pp. 107-148) and Webster and Oliver (1990, pp. 291-298).

simple answer to the question of which smoothing parameter to use even though the problem of choosing how much must be smoothed is of crucial importance in density estimation. After Silverman (1986), the appropriate choice of smoothing parameter must always be influenced by the purpose for which the density estimate is to be used. We chose the smoothing parameter subjectively, by trying several values and retaining the one corresponding to the classification deemed the best one capturing the environmental differences described by the previous geostatistical analysis. The number of clusters is a function of the smoothing parameter and generally tends to decrease as the smoothing parameter increases. However, the relationship is not strictly monotonic and several different values of the smoothing parameter generally have to be specified before seeing as the number of cluster varies.

The method is not inherently hierarchical, however, it can do approximate nonparametric significance tests for the number of clusters. An approximate p-value for each cluster is computed by comparing the estimated maximum density in the cluster with the estimated maximum density on the cluster boundary. The least significant cluster is then repeatedly joined with a neighbouring cluster until all remaining clusters are significant.

Finally, it is necessary to consider questions of scaling of variables, so that the attribute variances do not affect the resulting clusters. As the variables are not measured in comparable units, some sort of standardization or scaling is required if the variables are wanted to have equal importance in the analysis. The standardization used in this work scales all the variables to the same mean 0 and to the same variance 1.

In order to obtain spatially contiguous clusters, the clustering algorithm was applied to the data set of the interpolated soil variables and retained PCs of the estimated climatic variables and also the geographic coordinates of grid-cells were included in the attribute dataset.

Table 1. Relative contribution of the first three components of climatic variables.

Component	Eigenvalues	Proportion	Cumulative
PC1	7.9519	0.57	0.57
PC2	2.5252	0.18	0.75
PC3	1.3894	0.10	0.85

The clustering approach was implemented by using the MODECLUS procedure of the SAS/STAT software package (SAS, 2009, release 9.1.3).

2.5 Visualising agro-ecozones similarity using RGB colour triplet

A statistical colouring scheme (Hargrove and Hoffman, 2005) was used to visualise environmental similarities within different agroecoregions. PCA was applied either before or after clustering to condense a larger number of "raw" environmental variables into orthogonal principal component axes and the top three PCA components were mapped with a red-green-blue (RGB) colour triplet. The unique colour of each AEZ was derived by a mixture of RGB weighted by relative contribution of each of the three components to each AEZ, so reflecting the degree of similarity among AEZs.

3. Results and discussion

To condense the number of climate variables, a PCA was applied to all 36 variables followed by a VARIMAX orthogonal rotation to make easier the interpretation of the components. The relative contribution of the first three components is provided in Table 1, which shows that the three top PCs cumulatively explain about 85% of the total variance.

Factor loadings (not reported) were used to interpret the first, second and third principal components, which were used to create the RGB colour triplets. Maximum temperatures had the highest scores for the first factor; minimum temperatures for the second factor, whereas the winter-spring rainfall had the highest scores for the third factor. Based on these loadings, red colours represent dominance of high T_{max}, green colours dominance of high T_{min} and blue colours dominance of winter-spring rainfall within a zone. The RGB image (Fig. 3) shows as the warmer zones are located at south, whereas the rainier zones during the winter and spring months north-east at the border with Gargano promontory and along the west side at the foot of the Apennines.

To account for spatial dependence, an

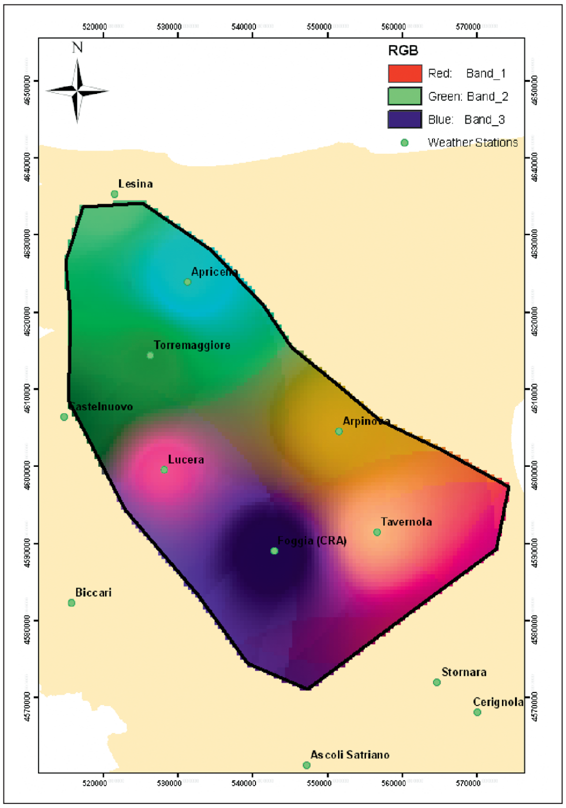

Figure 3. RGB map of climate attributes.

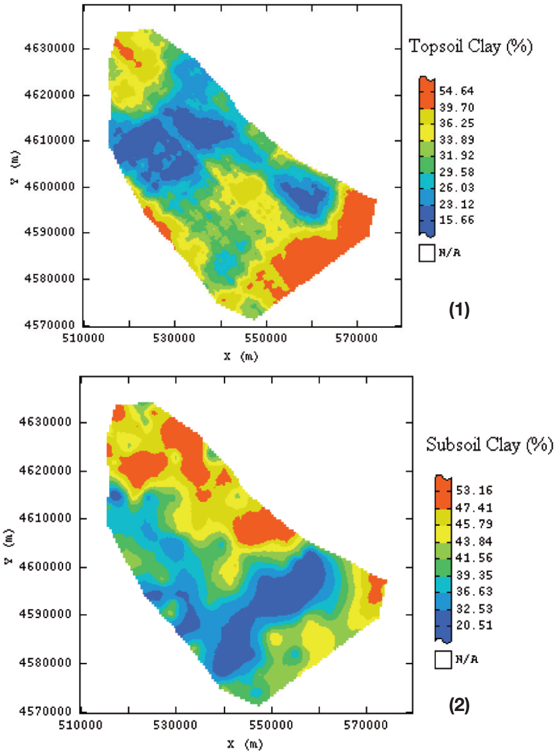

Figure 4a. Maps of clay percentage in topsoil (1) and subsoil (2).

isotropic LMC was fitted to all variograms of the soil attributes. No significant spatial anisotropy was disclosed, and the fitted LMC included (1) a nugget effect, (2) a spherical model with range = 8500 m and (3) an exponential model with range of 35000 m. The fitted multivariate model of spatial dependence is reported in Table 2 where, for each one of the three basic structures, the coregionalization matrix, composed by the sills of the direct and cross-variograms, is shown together with the eigenvalues and the percentage of explained variance (%). From the interpretation of eigenvalues, firstly it results the variance of the soil attributes to be mostly dominated by the erratic component (nugget effect) and secondly by short-range variation. However, the selected soil attributes were so strongly correlated between them at short range that the first eigenvalue explained more than 89% of the variance associated with the corresponding basic structure (Tab. 2).

In Figures 4a and 4b there are reported the cokriged maps of the soil variables using an

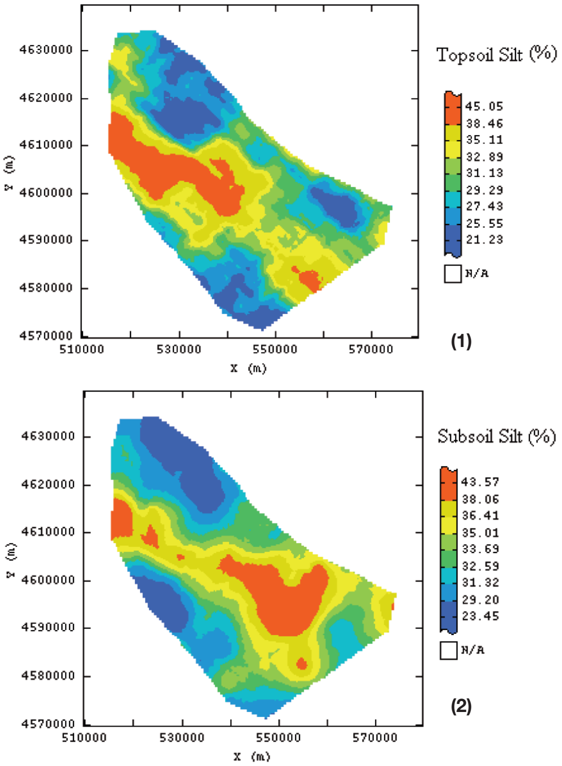

Figure 4b. Maps of silt percentage in topsoil (1) and subsoil (2).

Table 2. Coregionalization matrices of soil attributes corresponding to three basic spatial structures. The eigenvalues (EV) and the explained variance (P, %) corresponding percentage to each spatial scale are also reported. T is for topsoil and S for subsoil.

	pH$_T$	pH$_S$	Sand$_T$	Sand$_S$	OM$_T$	OM$_S$	PA$_T$	PA$_S$	CIC$_T$	CIC$_S$	Clay$_T$	Clay$_S$	EV	P
Nugget effect														
pH$_T$	0.05												268.33	52.39
pH$_S$	0.01	0.03											139.60	27.25
Sand$_T$	0.17	-0.20	98.22										61.52	12.01
Sand$_S$	0.50	0.43	23.58	108.64									20.34	3.97
OM$_T$	0.00	0.07	1.88	0.80	0.32								11.89	2.32
OM$_S$	-0.01	-0.02	0.15	-0.64	0.06	0.42							7.06	1.38
PA$_T$	0.00	0.06	-16.14	-3.81	-0.17	-0.05	8.60						1.96	0.38
PA$_S$	-0.01	0.07	-12.97	-28.76	-0.13	-0.46	3.52	23.87					1.10	0.21
CIC$_T$	0.04	0.10	-24.69	-11.43	-0.17	0.06	7.92	4.15	16.56				0.39	0.08
CIC$_S$	0.10	0.15	-14.58	-34.67	-0.01	-0.53	3.51	22.69	3.37	26.02			0.03	0.00
Clay$_T$	-0.19	0.05	-52.79	0.97	-1.36	-0.50	17.48	7.58	14.62	9.30	61.28		0.00	0.00
Clay$_S$	0.17	-0.61	-10.53	-77.28	-2.37	-0.84	5.06	57.16	7.41	52.79	10.05	168.20	0.00	0.00
Spherical - Range = 8500 m														
pH$_T$	0.00												223.08	89.32
pH$_S$	0.00	0.01											16.37	6.55
Sand$_T$	-0.11	0.12	44.09										7.87	3.15
Sand$_S$	0.35	0.13	-61.97	101.65									2.42	0.97
OM$_T$	-0.01	-0.03	-0.73	0.02	0.09								0.00	0.00
OM$_S$	0.01	0.01	-0.29	1.11	-0.05	0.04							0.00	0.00
PA$_T$	0.00	0.00	-9.94	12.58	0.22	0.00	3.61						0.00	0.00
PA$_S$	-0.04	0.05	23.23	-33.25	-0.41	-0.15	-5.65	14.01					0.00	0.00
CIC$_T$	0.04	0.02	-14.78	22.00	0.17	0.14	4.04	-8.94	6.01				0.00	0.00
CIC$_S$	-0.01	0.12	18.99	-24.12	-0.50	0.00	-4.23	9.72	-5.91	8.71			0.00	0.00
Clay$_T$	0.09	-0.08	-26.00	39.14	0.37	0.27	4.53	-14.68	8.86	-10.96	17.80		0.00	0.00
Clay$_S$	-0.15	-0.24	37.98	-61.20	-0.02	-0.58	-12.09	23.69	-17.14	13.86	-21.45	53.72	0.00	0.00
Exponential - Scale = 35000 m														
pH$_T$	0.04												107.78	56.04
pH$_S$	0.03	0.08											43.02	22.37
Sand$_T$	-0.49	-0.56	47.10										25.51	13.26
Sand$_S$	0.60	1.40	0.89	30.71									12.72	6.61
OM$_T$	0.08	-0.07	-1.61	-0.57	0.70								1.78	0.93
OM$_S$	-0.04	-0.12	0.55	-2.60	0.02	0.27							1.52	0.79
PA$_T$	0.01	-0.06	-7.25	-1.40	0.44	0.01	2.48						0.00	0.00
PA$_S$	-0.09	-0.12	-2.17	-2.93	-0.21	0.25	1.62	2.05					0.00	0.00
CIC$_T$	-0.13	-0.20	-9.60	-1.83	1.04	-0.40	3.93	1.78	10.73				0.00	0.00
CIC$_S$	-0.27	-0.30	1.60	-5.47	-0.15	0.28	0.18	0.37	2.46	2.44			0.00	0.00
Clay$_T$	1.02	1.44	-37.74	20.49	0.11	-1.13	4.06	-0.53	-4.00	-9.02	76.10		0.00	0.00
Clay$_S$	-0.04	-0.69	-1.83	-8.39	1.16	1.26	3.62	2.88	2.22	-0.31	10.11	19.63	0.00	0.00

isofrequency classes representation so to enhance the differences among the spatial patterns. The maps do not reveal a well-defined gradient and a general discontinuity occurs between topsoil and subsoil. Clay soils are predominant in the south, whereas north-eastern soils are coarser textured on the topsoil but finer textured in depth. Field capacity and wilting point vary concordantly with clay content. A wide median diagonal strip is characterised by higher content of silt on both top and subsoil.

To synthesise the complex multivariate variation above described in a restricted number of agro-ecozones, the clustering approach was applied to the all raster variables, i.e. the three retained PCs of the climatic variables and the 14 soil variables including the geographic coordinates. After several trials, the smoothing parameter was chosen equal to 1.8, because it produced the subdivision of the study area into 5 distinct clusters, deemed in good accordance with the prior description of spatial variation.

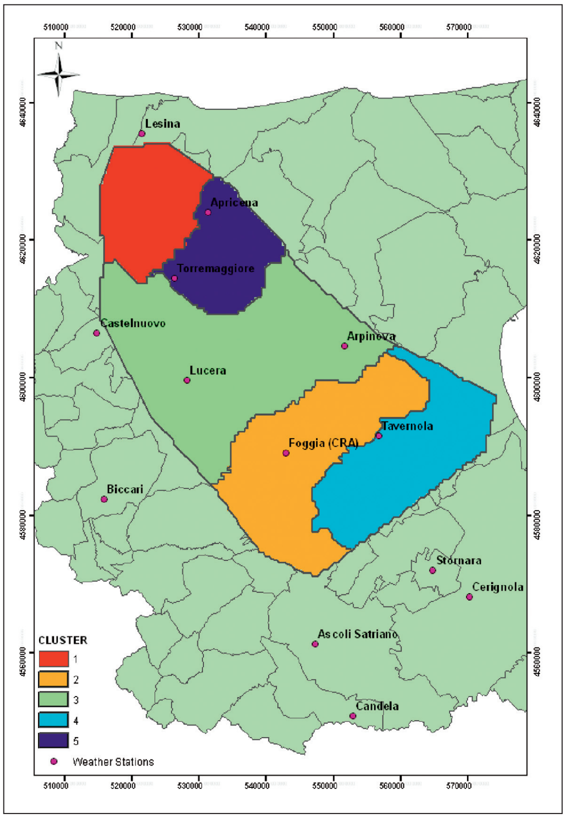

Figure 5. Agro-ecozones delineation.

This coarse delineation (Fig. 5) captures the maximum dissimilarity of environmental conditions across the area and the AEZs occur as globular, continuous patches.

In Table 3 means and standard deviations of the attributes for each cluster are reported.

The former represents the coordinates in the attribute space of the centroids, which provide a description of the average ecological conditions in each AEZ. The cluster 1 is the rainiest and coldest and is characterised by clay soils in depth. The cluster 2 is the most rainy in summer and its soils are characterised by the highest proportion of sand on the top. In the median clusters 3 and 4 the silt component increases on the top but the soil of the cluster 4 is mostly clay + silt along the whole profile. In this cluster the highest T_{max} values are recorded. The cluster 5 does not show clear distinctive properties and here loamy soil is predominant.

The randomly coloured map of the AEZs (Fig. 5) emphasizes the location of the borders between the agro-ecozones, but does not give any indication of how different environmental conditions mix across the borders or vary within the AEZs. One of the main advantages of the proposed approach, compared with the other traditional methods of clustering, is that it also gives information on the intrinsic spatial structure of the cluster and on the distribution of the residual variation within each class.

Table 4 reported the main statistics of each cluster, i.e. the boundary frequency, the maximum estimated density, the estimated saddle density (saddle point will be defined below), the number of observations within the neighbourhood of the modal observation, the number of observations within the neighbourhood of the saddle observation, the ratio between these two last counts, the number of observations within the overlap of the two previous neighbourhoods and the approximate p-value for the cluster.

To interpret the results shown in Table 4 and then to make inferences regarding cluster populations, it is necessary to remind what is meant by a cluster according to the method using non-parametric density estimation. A cluster is defined as a region (modal region) surrounding a local maximum of the multivariate probability density function or a connected set of local maxima, therefore, if a population has two clusters, there must be two modal regions and then a ‘valley’ between them. According to Hartigan and Hartigan (1985) there must be a ‘dip’ between the two modes and the maximum value of the neighbourhood distribution function along the boundary occurs at a ‘saddle’ point. A useful parameter to characterise the nature of the boundary between two clusters is then the ratio, which compares modal density with saddle density.

From the inspection of the first 4 statistics of the Table 4, it results that the cluster number 3 is the widest but the cluster 1 is the best defined, realising the highest value of maximum estimated density. However, this cluster also realises the highest value of estimated saddle density, which means that it might not be well distinct from the other neighbouring clusters. On the contrary, cluster 2 has the minimum value for estimated saddle density and the lowest estimated density. The nature of the cluster boundaries and then their “sharpness” is better defined by the last five statistics in Table 4, including also the results of the approximate sig-

Table 3. Means and standard deviations of the environmental attributes relative to each cluster (only statistics for $T_{min,}$ T_{max} and precipitations of the months: January (1), April (4), July (7) and October (10) are reported).

Variable	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev	Mean	St dev
$Clay_s$	46.69	1.96	46.57	2.86	40.75	4.85	42.66	3.13	33.32	4.40
$Clay_T$	34.82	3.00	24.55	3.12	28.41	5.75	40.59	5.82	31.76	4.47
CIC_T	43.62	1.38	43.54	1.43	42.45	2.91	43.11	1.56	38.50	2.40
CIC_S	31.61	1.82	27.92	1.78	33.60	1.73	34.11	2.39	29.06	2.66
$Silt_T$	30.13	3.16	29.55	2.81	34.45	3.50	34.80	1.81	35.25	3.56
$Silt_S$	29.36	3.28	25.85	1.80	35.41	4.47	32.28	3.72	28.72	3.62
PA_S	28.00	1.21	27.97	1.60	25.15	2.55	25.71	1.75	20.69	2.79
PA_T	17.81	1.10	15.57	1.23	18.66	1.11	19.50	1.37	16.82	1.19
pH_S	7.84	0.27	8.00	0.00	8.00	0.03	8.30	0.46	8.00	0.03
pH_T	7.88	0.16	8.00	0.00	7.97	0.09	8.00	0.00	8.00	0.00
$Sand_S$	23.18	3.21	23.76	1.42	24.87	6.32	22.37	3.44	31.39	3.18
$Sand_T$	35.82	5.11	49.56	3.25	36.19	5.47	27.16	7.82	39.54	6.22
OM_S	1.27	0.33	1.80	0.40	1.84	0.34	1.48	0.50	1.44	0.50
OM_T	1.33	0.38	1.20	0.40	2.37	0.52	2.57	0.61	2.84	0.46
P_1	79.51	6.82	71.07	1.71	71.67	6.45	62.23	1.27	65.92	2.86
P_4	78.11	3.28	64.38	5.17	64.93	10.63	49.53	1.04	54.00	3.81
P_7	20.09	0.89	20.93	0.36	20.38	0.68	19.14	0.34	20.03	0.61
P_{10}	50.45	1.92	46.08	2.10	45.79	3.94	40.74	1.00	41.67	1.59
T_{min1}	3.26	0.16	3.49	0.04	3.42	0.14	3.36	0.05	3.47	0.07
T_{min4}	7.15	0.16	7.44	0.05	7.39	0.18	7.70	0.08	7.55	0.08
T_{min7}	18.99	0.20	19.41	0.10	19.35	0.27	19.55	0.06	19.58	0.05
T_{min10}	11.95	0.27	12.27	0.04	12.14	0.19	11.91	0.09	12.14	0.14
T_{max1}	12.12	0.58	12.47	0.07	12.48	0.38	13.20	0.21	12.76	0.24
T_{max4}	20.04	0.88	20.44	0.10	20.45	0.53	21.30	0.24	20.75	0.31
T_{max7}	33.93	1.09	34.43	0.12	34.37	0.62	35.16	0.25	34.62	0.31
T_{max10}	23.40	0.81	23.86	0.11	23.89	0.54	24.78	0.21	24.27	0.31

nificance test. The clusters 3 and 2 look as the most distinguishable from the neighbouring clusters, realising the highest values of the ratio, on the contrary the cluster 4 looks as the most muddled.

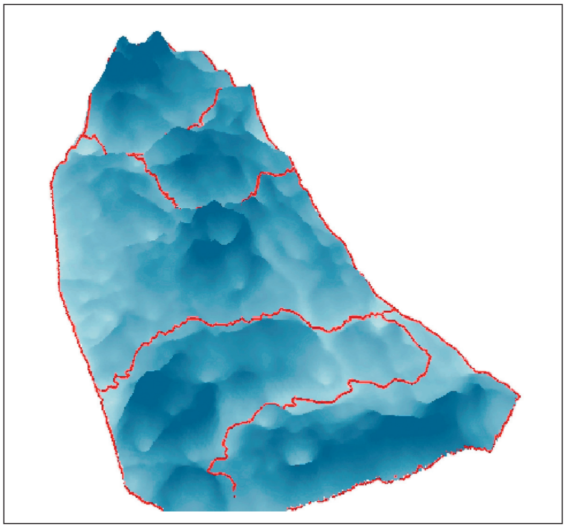

Figure 6. Density estimation topography for the study area. The boundaries between agro-ecozones are drawn as bold lines.

However, these statistics describe the overall behaviour of the clusters, whereas the visualization of the density function in a 3D space can aid to understand the differences in the environmental properties of the AEZs. The map of the density function into geographic space (Fig. 6) depicts these values as elevations and creates a surface whose peaks correspond to the location of the cluster's centroid. Because we can calculate such a value for all cells, this probability surface is complete and continue across the map.

The elevation surface also reflects the cell-representativeness and the idealised hypothetical cluster agro-ecozones appear as a series of peaks with borders tracing along the valleys or depressions between the peaks. Therefore, the highest geographic locations represent the cells at or near the cluster's centroid. The clusters 1 and 4 are the most distinct, with the highest peaks and each one essentially with one single mode. The cluster 3 shows one single central

Table 4. Statistics for the 5 clusters.

Cluster	Boundary frequency	Maximum estimated density	Estimated saddle density	Mode count	Saddle count	Ratio	Overlap count	Approximate P-value
1	967	1.32 x 10^{-5}	4.12 x 10^{-6}	317	98	3.23	0	1 x 10^{-6}
2	849	9.57 x 10^{-6}	2.04 x 10^{-6}	229	48	4.77	0	1 x 10^{-6}
3	2854	1.09 x 10^{-5}	2.20 x 10^{-6}	260	52	5.00	0	1 x 10^{-6}
4	1785	1.14 x 10^{-5}	3.74 x 10^{-6}	273	89	3.07	0	1 x 10^{-6}
5	1441	1.08 x 10^{-5}	2.83 x 10^{-6}	259	67	3.87	0	1 x 10^{-6}

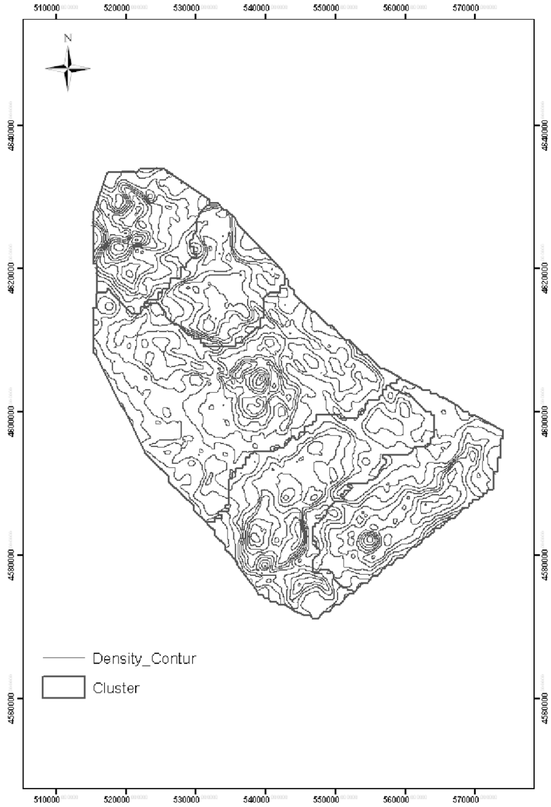

Figure 7. Agro-ecozones represented with equal-elevation contours of the density estimation.

mode but most of the outlying cells are at low elevation. The cluster 5 looks more compact than the others, showing a set of connected modes, whereas for the cluster 2 no clear high spot is evident corresponding with the cluster's centroid. A great advantage of this clustering method and the 3D visualization of the probability density is the possibility of viewing the morphology of the surface and then of making an assessment of the residual local variation within each cluster. The traditional methods produce average properties of each cluster and a global estimation of within-cluster variation, supposed homogeneously distributed. The 3D graph clearly shows how the clusters have different degrees of compactness in attribute space and how some clusters, such as the clusters 2 and 5, actually show several local modes, making not clearly defined what a cluster is meant.

Since a cluster is defined by the properties of its modal point, the 3D graph shows how the departures from these properties may occur abruptly or gradually on the slope of a peak indicating a fuzzy, gradual area of transition. Moreover, the borders may also change from fuzzy to sharp or vice versa and, because edge properties are dependent on each adjacent cluster, each side may have distinct and different properties. This "sideness" property may appear initially as counterintuitive, but it is completely logical because we are characterising the transition from the border to the centroid independently on each side (Hargrove and Hoffman, 1999). Figure 7 shows equal-elevation probability contours visualising the sharpness of the AEZ borders. The contour's random orientation and meandering character at the borders between the clusters 1-2 and 3-4 clearly indicate that these edges are transition areas. On the contrary, closely-spaced, parallel contour lines separating the clusters 2 and 3 and the cluster 5 from the outside the study area indicate abrupt changes. Moreover, the border between the cluster 4 and 5 looks sharp on the western part and on the side of the cluster 4, but more fuzzy on the other side. Contour lines then have the flexibility to represent mixed gradual / sharp borders, as well as borders whose characteristics change along their length. In the particular environmental conditions under study the borders generally change sharpness characteristics along their length, since deep valleys alternate with high saddle points.

Figure 8 shows the same AEZs but using the RGB-colour triplet, created applying PCA to the whole set of the nineteen environmental

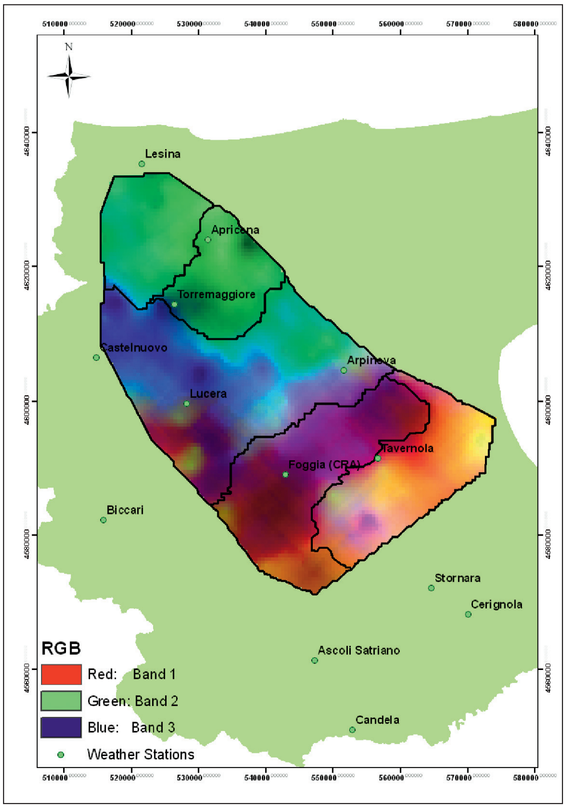

Figure 8. Visualization of study area using RGB colours based on nineteen environmental characteristics.

variables, including the geographic coordinates, the three PCs of the previous meteorological analysis and the fourteen soil variables. Varimax rotation factor loadings were used to interpret the first three PCs, which cumulatively explained about 76% of the total variance. Based on these loadings (not reported), red colours represent dominance of maximum temperature, clay content, pH and organic matter on the top soil; green colours represent dominance of clay content, FC and WP in depth, and blue colours of silt content along the whole profile, FC and WP on the top soil and low rainfall during the winter and spring months. In this representation individual cluster borders disappear and the colours reveal similarities among the environments in each AEZ. The red south is dominated by high temperature and clay content on the top soil; the blue central part is dominated by higher content of silt and low precipitation; the green north-east is characterised by higher content of sand on the top soil and clay content in depth. Abrupt colour changes are generally related to parallel contours (Fig. 7), whereas subtle colour changes are generally accompanied by the meandering contours of a transition area. However, locations with equal elevations may have different colours within the same cluster due to differences in environmental conditions.

Visualisation, such as that produced, can also provide a way to assess the appropriateness of geographic clustering. For example, the appearance of multiple peaks within a single cluster, such as in the cluster 5, might suggest that we need more divisions, whereas border passing through high saddle areas might suggest we need fewer, as the cluster 4.

4. Conclusions

The achievement of compact classes in the space of environmental attributes, which are also contiguous in geographic space, is highly desirable in agro-ecozoning and mapping. This study has demonstrated that combining continuous classification with geostatistical interpolation can provide useful means for automatically locate boundaries between AEZs. By limiting input variables to those more specifically crop relevant, AEZ delineation is more likely to reflect agriculturally relevant differentiation of the environment than the traditional regionalisation. Therefore, a primary application of this method is the evaluation of crop suitability. If the crop growth needs are known, the approach can be used as a screening tool, through an indicative transform, in suitability analysis.

Further, the method overcomes the limitations associated with a discrete delineation between AEZs by creating complex borders and regions of transition between the zones. A great advantage consists in giving a visual and quantitative assessment of environmental both between - and within AEZ variability in crop perspective, which is quite useful in agricultural land use decision making.

An animation of 3D graph of the density function is available at: http://www.siagr.org/riproducivideo.asp?file=3D_scene_1_minute.avi

Acknowledgment

This work is supported by Italian Ministry of Agriculture, Food and Forestry Policies under contract n. 209/7303/05 (AQUATER Project).

References

Belbin L. 1993. Environmental representativeness: regional partitioning and reserve selection. Biological Conservation, 66:223-230.

Caldiz D.O., Gaspari F.J., Haverkort A.J., Struik P.C. 2001. Agroecological zoning and potential yield of single or double cropping of potato in Argentina. Agric. For. Meteorol., 109:311-320.

Carter M.R., Gregorich E.G., Anderson D.W., Doran J.W., Janzen H.H., Pierce F.J. 1997. Concepts of soil quality and their significance. In: Gregorich E.G., Carter M. (eds.): Soil Quality For Crop Production And Ecosystem Health. Elsevier Science Publishers, Amsterdam, Netherlands, 1-17.

Castrignanò A., Giugliarini L., Risaliti R., Martinelli N. 2000. Study of spatial relationships among soil physical-chemical properties using Multivariate Geostatistics. Geoderma, 97:39-60.

Csillag F., Kabos S. 2002. Wavelets, boundaries and the spatial analysis of landscape pattern. Ecoscience, 9, 2:177-190.

Davis J.C. 1986. Statistics and data analysis in Geology, 2° ed. John Wiley & Sons, New York, 107-148.

ESRI, Environmental Systems Research Institute, 2005. Grid commands. ERSI, Redlands, CA.

FAO 1996. Agro-Ecological Zoning: Guidelines. Food and Agricultural Organization of the United Nations, Rome.

Geovariances 2009. ISATIS, Software manual, release 9. Geovariances ècole des Mines de Paris, France.

Goovaerts P. 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, 483 pp.

Harff D., Eiserbeck H.J. and Davis J.C. 1990. Regionalization in Geology by Multivariate Classification. Math. Geol., 22, 5:573-588.

Hargrove W.W. and Hoffman F.M. 1999. Using Multivariate Clustering to Characterize Ecoregion Borders. Computing in Science & Engineering, 1, 4:18-25.

Hargrove W.W., Hoffman F.M. 2005. Potential of multivariate quantitative method for delineation and visualization of ecoregions. Environmental Management, 34, suppl. 1:S39-S60.

Hartigan J.A., Hartigan M. 1985. The Dip Test of Unimodality. Annals of Statistics, 13:70-84.

Host G.E., Polzer P.L., Mladenoff D.J., White M.A., Crow S.J. 1996. A quantitative approach to developing regional ecosystem classifications. Ecological Applications, 6:608-618.

http://inrm.cip.cgiar.org/home/publicat/99oth83.pdf

Irvin B.J., Ventura S.J. and Slater B.K. 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma, 77:137-154.

Jensen J.R. 1996. Introductory Digital Image Processing: A Remote Sensing Perspective. Prentice-Hall, Inc, New Jersey, 197-256.

Journel A.G., Huijbregts C.J. 1978. Mining Geostatistics. Academic Press, New York.

Lajaunie C., Béhaxétéguy J.P. 1989. Elaboration d'un programme d'ajustement semi-automatique d'un modèle de corégionalisation - Théorie. Technical report N21/89/G. Paris: ENSMP, 6 p.

Lark R.M. 1998. Forming spatially coherent regions by classification of multivariate data: An example from the analysis of maps of crop yield. Int. J. Geogr. Inf. Sci., 12:83-98.

Leung Y. 1987. On the imprecision of boundaries. Geographical Analysis, 19:125-151.

Loomis R.S. and Connor D.J. 1992. Crop ecology: Productivity and management in agricultural systems. Cambridge Univ. Press, Cambridge, UK.

McMahon G., Gregonis S.M., Waltman S.W., Omernik J.M., Thorson T.D., Freeouf J.A., Rorick A.H., Keys J.E. 2001. Developing a spatial framework of common ecological regions for the conterminous United States. Environ. Manage., 28:293-316.

Metzger M.J., Bunce R.G.H., Jongman R.H.G., Mucher C.A., Watkins J.W. 2005. A climatic stratification of the environment of Europe. Global Ecol. Biogeogr., 14:549-563.

Regione Puglia 2001. Progetto ACLA2: "Caratterizzazione agro ecologica della Regione Puglia in funzione della potenzialità produttiva: Sottoprogetto carta pedologica in scala 1:100.000".

Sarle W.S. 1982. Cluster Analysis by Least Squares. SAS Users Group International Conference Proceedings: SUGI 7, 651-653.

SAS Institute Inc. 2009. SAS/STAT Software Release 9.1.3, Cary, NC, USA.

Scott D.W. 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons Inc., New York, 33-46.

Silverman B.W. 1986. Density Estimation. Chapman and Hall, New York, 75-94.

Swinton S.M., Quiroz R.A., Paredes S., Reinoso J.R., Valdiva R. 2001. Using farm data to validate agroecological zones in the Lake Titicaca basin, Peru. In: Proceedings of the 3rd International Symposium on Systems Approaches for Agricultural Development.

Webster R., Oliver M.A. 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, Oxford.

Williams C.L., Hargrove W.W., Liebman M. 2008. Agroecoregionalization of Iowa using multivariate geographical clustering. Agriculture, Ecosystems and Environment, 123:161-174.