MSEL
Modelling in Science Education and Learning

Modelling in Science Education and Learning
Volume 4, No. 7, 2011.
Instituto Universitario de Matemática Pura y Aplicada

Semivariogram calculation optimization for object-oriented image classification

A. Balaguer-Beser, T. Hermosilla, J. Recio, L.A. Ruiz
Universidad Politécnica de Valencia
abalague@mat.upv.es, txohergo@topo.upv.es, jrecio@cgf.upv.es, laruiz@cgf.upv.es

Abstract

En este trabajo se proponen y evalúan diferentes parámetros matemáticos extraídos del semivariograma experimental para la clasificación de los usos del suelo mediante imágenes de alta resolución, usando los límites catastrales para la definición de los objetos de análisis. En primer lugar, se describe el proceso de cálculo del semivariograma a partir de los valores de niveles de gris del objeto imagen. Con el fin de optimizar el tiempo de cálculo se presentan dos técnicas de selección de píxeles que conservan la forma original del semivariograma. A continuación se definen varios parámetros del semivariograma. Finalmente, se usan diferentes técnicas estadísticas para la selección de los parámetros más discriminantes. La última sección muestra los resultados obtenidos con las imágenes digitales aéreas de una zona agrícola en la costa mediterránea de España. El estudio de la aplicación práctica que se presenta facilita la comprensión de la relación entre el comportamiento del semivariograma experimental y la variabilidad de los valores de intensidad en una imagen digital. Con el fin de seguir el desarrollo de este trabajo, el lector debe conocer algunos métodos estadísticos de clasificación y algunas técnicas de procesamiento digital de imágenes.

In this paper we propose and evaluate different mathematical parameters extracted from the experimental semivariogram for land use/land cover classification using high-resolution images and cadastral mapping limits for the definition of the objects of analysis. First, we describe the process of calculating the semivariogram from the gray level values in an image object. In order to optimize the computation time we present two pixel selection techniques that preserve the original shape of the semivariogram. Several parameters are then extracted from the semivariogram. Finally, we use various statistical techniques to select the most discriminant parameters. Last section shows the results obtained using aerial digital images of an agricultural area on the Mediterranean coast of Spain. The study of the practical application presented in this paper facilitates the understanding of the relationship between the behaviour of the experimental semivariogram and the variability of the intensity values in a digital image. In order to follow the development of this work, the reader should know some basis of classification methods and digital image processing techniques.

Keywords: Semivariogram features, object-oriented classification, digital image processing.

1 Introduction

The semivariogram is a key mathematical tool for geo-statistical studies. The semivariogram describes the spatial variability of the values of a variable. It relates the semivariance with the spatial separation, providing a concise and unbiased description of the scale and the pattern of spatial variability (Curran [5]). The semivariogram curve quantifies the spatial associations of the values of a variable, and measures the degree of spatial correlation between different pixels in an image (Chilés and Delfinder [6]).

Known a spatial dependent dataset, the semivariogram is used to estimate the analyzed variable value in different locations considering the spatial correlation of the sample data. This process is known as kriging estimation and it has been studied in various earth sciences disciplines (see Goovaerts [8] and Portalés [11]). The semivariogram is a mathematical function that depends on the distance between points. Various parameters obtained from the sample data are required to perform its definition. The correct semivariogram modeling largely depends on the study of the experimental semivariogram behaviour, especially at close to zero distances and over long distances. It is important to know if the semivariogram is stabilized around a sill and to compute the distance from which this occurs, known as range. Besides to define the theoretical semivariogram model and thereby to obtain a kriging estimate, experimental semivariogram parameters can also be applied for texture features definition for digital image classification (see Carr and Miranda [4]).

The semivariogram has been frequently used in remote sensing studies focused on the extraction of texture features to perform image classification, using different types of imagery and in several different applications. The semivariogram has demonstrated to have a superior performance than traditional methods such as the grey level co-occurrence matrix (Balaguer et al. [2]). Besides, the information obtained from different features describing the semivariogram graph complements the features extracted with other methodologies and mathematical tools for land use/land cover classification using high resolution imagery and object-oriented approach (Ruiz et al. [15]).

In the object-oriented approach images are divided in smaller segments or image-objects by employing automatic segmentation techniques or cartographic limits. Then, these image-objects are described with descriptive features extracted from images or ancillary data. In previous works, semivariogram derived features were developed and tested (Recio [14]; Balaguer et al. [2]) describing effectively the spatial patterns in the objects. The computation of the experimental semivariogram in high spatial resolution digital images for each object is a time-consuming process, due to the large amount of pixels composing the image-objects, being this the main drawback of this technique. Finally each object is classified in one of the different land use/land cover classes.

In this study, we analyze the effect that the use of different proportion of pixels in the calculation of the semivariogram of each object has in the image classification accuracy. We present a methodology to optimize the semivariogram computing time by using a reduced number of pixels per image-object.Two different pixel selection strategies are used: random and stratified. Random pixel selection strategy arbitrarily selects a defined percentage of pixels from the image-object, whereas in stratified pixel selection strategy the image-objects are divided in regular subgroups from where the defined percentage of pixels is randomly selected.

The use of the semivariogram has been extended to a large range of applications. In our case, a detailed study of the computation of the experimental semivariogram is required, as

well as the relationship between the different parameters extracted from this function, and the variability of the spatial data. Thus, the methodology shown in this paper may be used as research and teaching material for the higher levels of engineering, showing the importance of modeling in science education and its use to improve learning. This methodology can be used in standard subjects devoted to geostatistical techniques, in order to teach the parameters to be considered when a theoretical semivariogram model is adapted to a sample data, or in subjects related to Image Processing techniques, as a tool to analyse the structure of the elements of an image. It may have especial relevance for undergraduate students in Geomatics and Surveying Engineering, as well as for graduate level students in Geodetic Engineering and Cartography, since it combines concepts from Geostatistics, Remote Sensing and Digital Image Processing.

This paper is organized as follows. First, the data and study area are presented. Later, the methodology used to calculate the semivariogram and image classification is explored. Finally, the results obtained are discussed, leading to final conclusions.

2 Data and study area

The study was performed in a rural area in the municipality of Benicarló in the north of the province of Castellón, on the Mediterranean coast of Spain. The study area is composed of a variety of land cover types, presenting large surfaces of citrus trees coexisting with vegetable crop fields in areas near the coast, as well as carob-tree orchards, pine forests, and shrubland in mountainous areas. Eight different classes were considered: citrus orchards, young citrus orchards, buildings, forest, carob-trees, irrigated crops, shrubland, and arable land.

The remotely sensed data used were aerial ortoimages with 0.5 m/pixel resolution, acquired in August 2005 with a digital mapping camera (DMC). The final spatial resolution is achieved through a fusion between panchromatic and multispectral bands. The system is composed of three bands in the visible part of the electromagnetic spectrum (0.4-0.58 μm, 0.50-0.65 μm, and 0.59-0.675 μm), one band in the near infrared (0.675-0.85 μm), and a panchromatic band.

The limits of the plots with homogeneous land uses were provided by vectorial cadastral cartography at a scale of 1:2000. The actual land use of each plot on the image acquisition date was obtained by photo-interpretation techniques. This information was employed as ground truth data for training the classifier and for evaluation processes, being composed of 150 plots per class, making a total of 1200 plots. Fifty plots per class were used as training samples, reserving the remaining 100 as evaluation samples. An equal number of samples per class was taken in order to avoid the influence of under and over-represented classes in the results, since decision-trees are sensitive to large discrepancies in the number of training samples between individual classes (Borak and Strahler [3]; McIver and Friedl [10]).

3 Methodology

Experiments have been performed using an object-oriented approach. The scheme carried out begins by creating image-objects based on the cartographic limits of the plots. The process followed consists of three stages.

1. First, the experimental semivariogram is computed. Semivariograms representing each image-object are computed by considering different percentage of pixels for their calcula-

tion by using two sampling methods: random and stratified.

2. Second, several mathematical parameters characterizing the shape and properties of the semivariogram are defined.

3. To conclude, image-objects are classified in one of the defined classes using the selected parameters and decision trees, ending with an evaluation of the classification.

Considering one sampling percentage, the use of different pixel values produces different semivariograms, and consequently, different descriptive parameters. Therefore, iterations are required to obtain robust and comparable results between sampling strategies and pixel percentages.

3.1 Experimental semivariogram computation

The semivariogram is a particularly suitable tool in the characterization of regular patterns. It provides information on plant pattern distribution in agricultural plots, indicating the type of spatial arrangement. For continuous variables, such as reflectance in a given spectral band, the experimental semivariogram is defined as half of the average squared difference between values separated by a given lag, where this lag is a vector in both distance and direction (Atkinson and Lewis [1]). The experimental semivariogram is computed as:

$$\gamma\left(\vec{h}\right) = \frac{1}{2N(\vec{h})} \sum_{i=1}^{N(\vec{h})} \left[z(x_i) - z(x_i + \vec{h})\right]^2 \tag{7.1}$$

where $z(x_i)$ represents the value of the variable at the location x_i, $\vec{h}$ the separation between elements in a given direction and $N(\vec{h})$ the number of data pairs occurring at locations x_i and $x_i + \vec{h}$. Since pixels in image data are regularly spaced, no tolerance on the separation vector $\vec{h}$ is considered. Subsequently, in expression (7.1) we consider that x_i represents the central position of pixel i, $z(x_i + \vec{h})$ is equal to $z(x_k)$, and x_k is located at the central position of the nearest pixel to the location $x_i + \vec{h}$. Considering an object-oriented approach, each image-object is characterized with one object-specific experimental semivariogram. Thus, only those pixels inside an object are considered for the computation of the semivariogram, completely eliminating the classification *border effect* (Ruiz et al. [15]).

The omnidirectional semivariogram is obtained by averaging the semivariograms of all possible directions, requiring a long processing time. In order to reduce this time, the multidirectional semivariogram is obtained by computing the mean of the semivariograms calculated in six directions, ranging from 0^o to 150^o with a step of 30^o. Moreover, each semivariogram curve is filtered using a Gaussian filter with a stencil of 3 positions, in order to smooth its shape and to eliminate experimental fluctuations. The experimental semivariograms derived from the infrared band show different behaviours for each class (see Figures 1 and 2).

In figures 1 and 2 we observe the following:

1. Arable land plots, which have low internal variability, lead to a monotone increasing semivariogram with very low semivariance values.

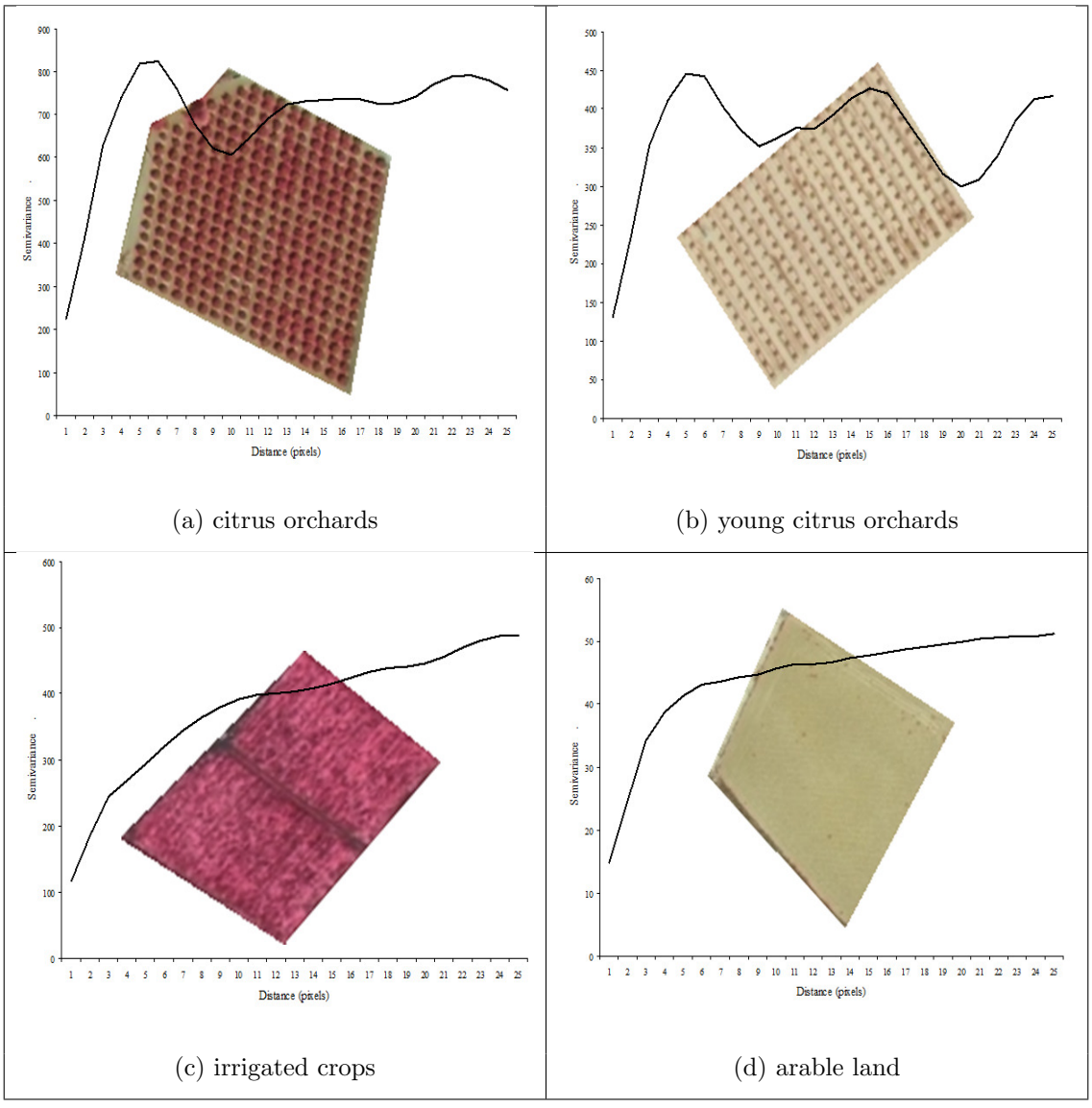

Figure 1: Semivariogram graphs for image-objects belonging to the classes: (a) citrus orchards, (b) young citrus orchards, (c) irrigated crops and (d) arable land.

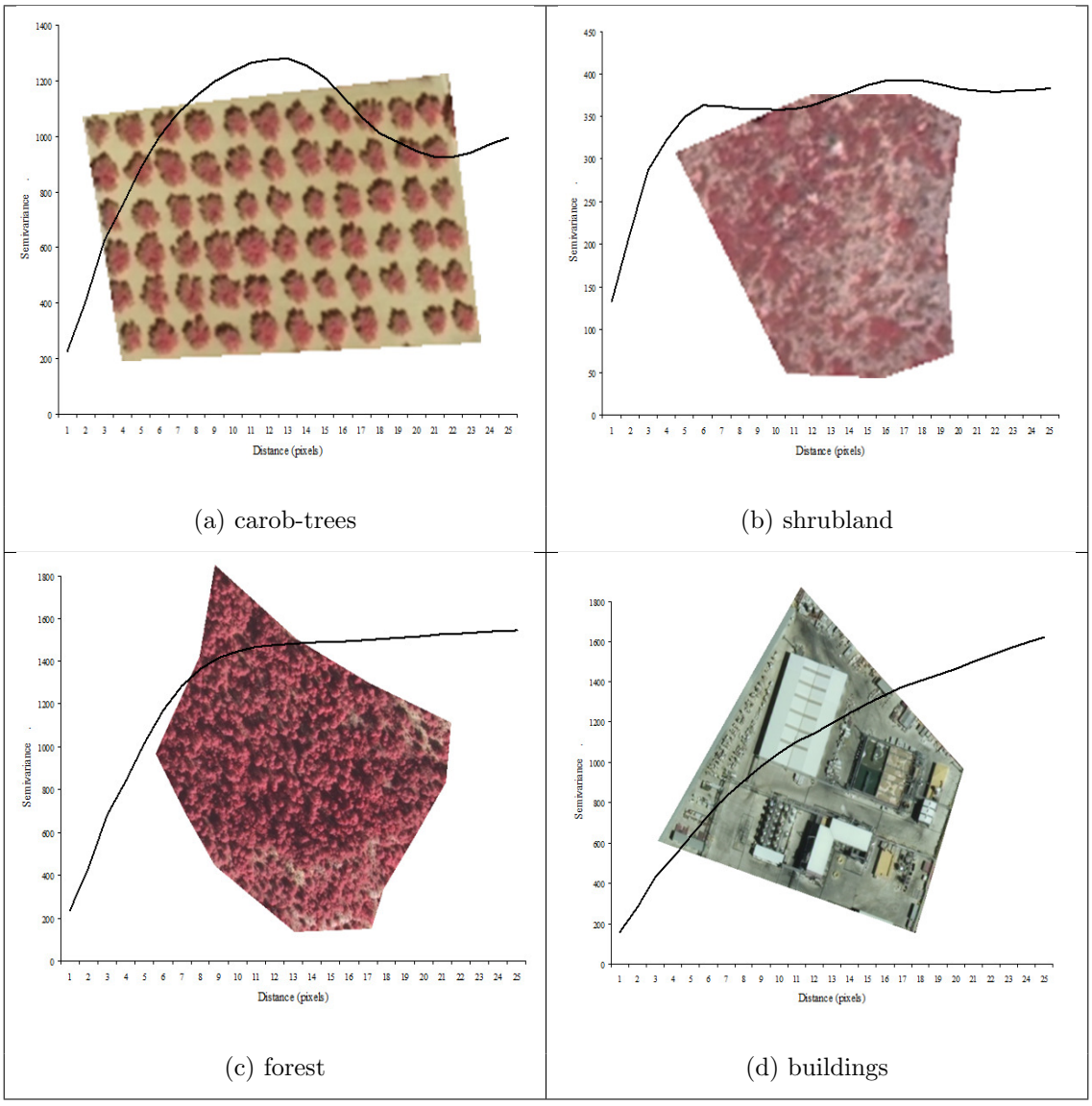

Figure 2: Semivariogram graphs for image-objects belonging to the classes: (a) carob-trees, (b) shrubland, (c) forest and (d) buildings.

2. The semivariogram curves of forest and buildings classes show a shape comparable to arable land, but having higher semivariance values.

3. The semivariograms corresponding to the class irrigated crops and shrubland present a rising trend with some irregularities due to the presence of micro-patterns, with medium semivariance values.

4. While experimental semivariograms often increase with lag distance, the classes citrus orchards and young citrus orchards present cyclic semivariogram curves, known as *hole effect* semivariogram (see Pyrcz and Deutsch [12]), that is typically produced when the studied variable has a periodic spatial behaviour.

5. Finally, the class carob-trees also presents a *hole effect* semivariogram. However, this is not completely registered by the corresponding graph due to the fact that the analysis distance used to compute the semivariogram is insufficient to capture the complete periodic spatial pattern defined by the larger planting distance in this crop.

3.2 Semivariogram descriptive features

In this work, eight semivariogram features were employed to describe the image-objects. They are fully described in Balaguer et al. [2]. These features characterize the semivariogram behaviour according to the position of the lags used in their definition: near the origin, up to the first maximum, and between first and second maxima.

1. The ratio between the values of the total variance and the semivariance at first lag: $RVF = \frac{Variance}{\gamma_1}$.

2. The ratio between semivariance values at second and first lag: $RSF = \frac{\gamma_2}{\gamma_1}$.

3. The first derivative near the origin: $FDO = \frac{\gamma_2 - \gamma_1}{h_2 - h_1}$.

4. The lag value where the curve $\gamma(h)$ reaches the first local maximum: $FML = h_{\max_1}$.

5. The features mean of the semivariogram values up to the first maximum:

$$MFM = \gamma^{mean}_{\max_1} = \frac{1}{\max_1} \sum_{i=1}^{\max_1} \gamma_i.$$

6. The variance of the semivariogram values up to the first maximum:

$$VFM = \frac{1}{\max_1} \sum_{i=1}^{\max_1} \left(\gamma_i - \gamma^{mean}_{\max_1}\right)^2.$$

7. The ratio between the semivariance at first local maximum and the mean semivariogram values up to this maximum: $RMM = \frac{\gamma_{\max_1}}{\gamma^{mean}_{\max_1}}$

8. The distance between the first maximum and the first minimum: $DMM = h_{\min_1} - h_{\max_1}$

RVF and RSF features are related to the homogeneity values of the grey levels at long and short distances respectively. FDO feature shows the variability changes of the data at short distances. FML, MFM, VFM and RMM features are related with the overall variability of the grey level values. DMM feature characterizes periodic patterns within an image-object and quantifies the hole effect, which is directly related to the variability or contrast of the regularity patterns.

3.3 Sampling techniques for pixel selection for semivariogram computing

The calculation of the semivariogram requires a lot of time when working with high-resolution images. To avoid high computation time, this paper shows the result of performing a selection of a percentage of data from the total of $N(\vec{h})$ pairs points, preserving the shape of the semivariogram. Thus in the calculation of $\gamma(\vec{h})$ (see formula (7.1)) only a percentage of the total points $N(\vec{h})$ is considered. There are different strategies to make the selection, seeing the distribution of points displayed on a h-scatterplot. An h-scatterplot shows all possible pairs of data values whose locations are separated by a certain distance in a particular direction. In a h-scatterplot the first coordinate coincides with the values of $z(x_i)$ and the second coordinate is $z(x_i + \vec{h})$ (see Isaaks and Srivastava [9]).

In the case of a rectangular grid, this selection process can be simplified. The first selection technique considered here -denoted as the **random pixel selection method**- is precisely the random selection of a percentage of values of $z(x_i)$ on the x-axis of the h-scatterplot. Then, all the points are selected in the h-scatterplot with first coordinate coinciding with the points selected in the previous step, and only this selection of points will be taken into account for calculating the semivariogram. To calculate $\gamma(\vec{h})$, for every distance and direction, the same pixel selection in the x-axis of the h-scatterplot is chosen. The results after performing a different selection of pixels in every distance and direction were analyzed. However, the overall accuracies of the classification do not improve the classification results with respect to the method mentioned above.

Random pixel selection strategy arbitrarily selects a defined percentage of pixels from the image-object. This strategy is compared to a stratified pixel selection strategy at which the image-objects are divided in regular subgroups from where the defined percentage of pixels is randomly selected, using the same process described above for the random pixel selection method. This method will be denoted by **stratified pixel selection method**.

3.4 Classification

Classification is performed using decision trees built using C5.0 classification algorithm, which is the latest version of the algorithms ID3 and C4.5 developed by Quinlan [13]. Decision trees are created following the boosting multi-classifier method (Freund [7]). The algorithm searches the features that best separate one class form the others by dividing data using mutually exclusive conditions until the newly generated subgroups are homogeneous, i.e. all the elements in a subgroup belong to the same class or a stopping condition is fulfilled. For each classification, the assessment is based on the analysis of the confusion matrix (Story and Congalton [16]), by comparing the class assigned to each evaluation sample with the reference information. The overall accuracies of the classifications were computed.

4 Results and discussion

Highest classification accuracies have been obtained in those classes presenting heterogeneity in their grey level values: citrus orchards, young citrus orchards, buildings, irrigated crops and carob-trees. On the other hand, the lowest accuracy performances have been given in shrubland and arable land classes. The overall classification accuracy considering all the pixels

of the image-objects for semivariogram computation was 81.25%.

Figure 3 shows a comparison between the accuracies obtained using random and stratified sampling strategies and 10 iterations per considered sampling percentage. Comparable mean, minimum and maximum values have been obtained using both sampling strategies. In both cases, mean overall accuracy values trend to be more stable with sampling percentages higher than 30%. These results show the poor relevance, in this case, of the sampling strategies.

Using random sampling strategy a new simulation has been performed by applying 100 iterations per pixel sampling percentage (figure 4).

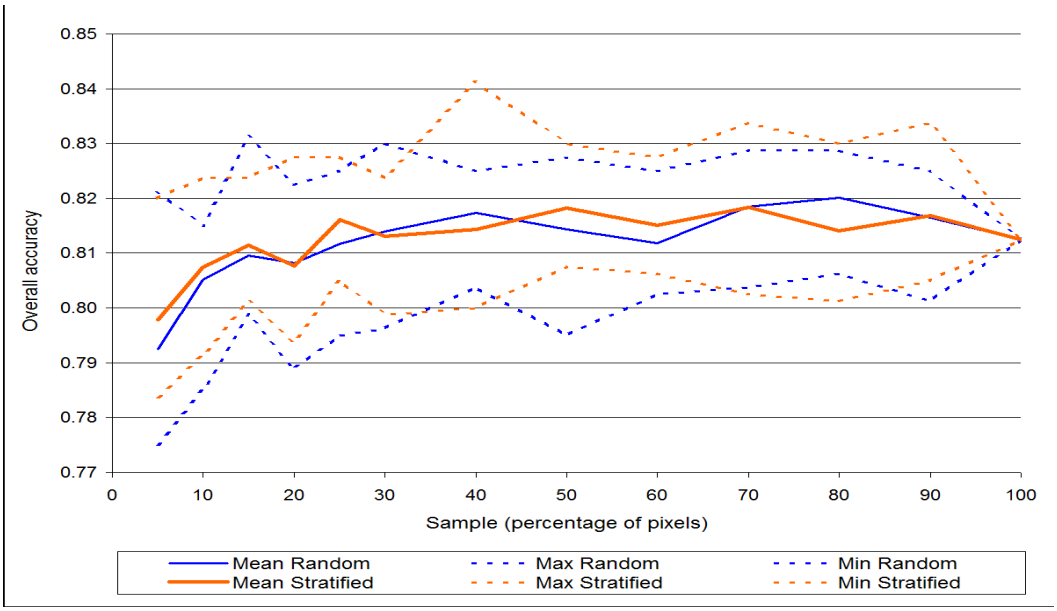

Figure 3: Comparison of mean, maxima and minima overall accuracies values applying 10 iterations for each pixel sampling percentage considered using random and stratified pixel selection methods.

In figure 4, mean and median values increased slightly when the sampling percentage is incremented. This variation increases from a 79 % up to 81 %, respectively, considering percentages ranging from 5 % to 100% of the pixels for semivariogram computation. Minima values indicate the presence of some outliers in all the considered percentages. The percentiles inform about the high degree of stability of the results, even when low sampling percentages are employed.

Figure 5 shows the average processing time (seconds) per plot for several percentage of pixels in the selection process, being the computing time lower when a random selection of pixels is done. The processing time grows linearly in the two selection methods considered in this work.

5 Conclusions

A real application of the semivariogram in image classification has been presented in this paper. This may complete the teaching examples about kriging estimation and simulation. The methodology shows how the properties of a semivariogram curve (slope, curvature, maximum, minimum, concavity, convexity, etc.) can be parameterized in order to characterize the spatial distribution of the input data. The relationship between the behaviour of the semivariogram and the spatial distribution patterns of the elements of a plot has been analyzed. In order to reduce computation time, the results obtained by two pixel selection techniques have been

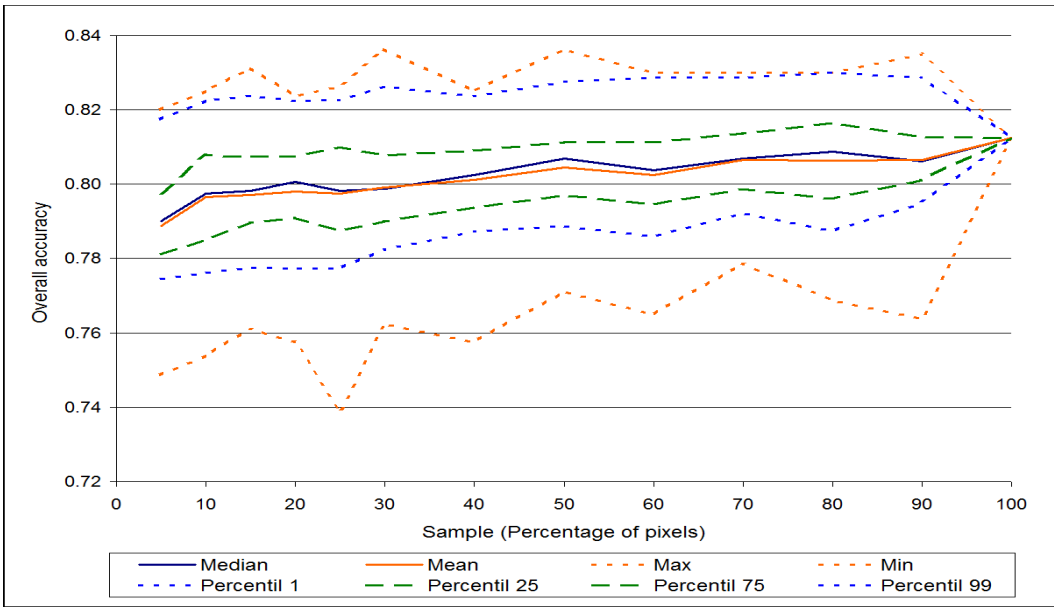

Figure 4: Overall accuracy values applying 100 iterations for each pixel sampling percentage considered using the random pixel selection methods.

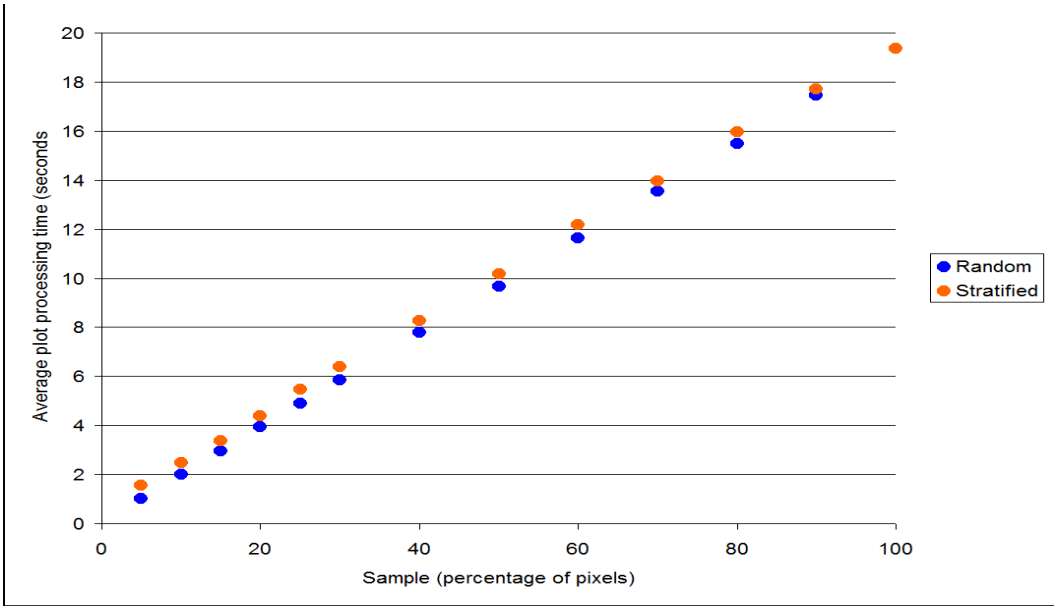

Figure 5: Average plot processing time (seconds) for different percentages of pixels in the selection process.

compared, in order to optimize the required time for calculating the semivariogram.

An analysis of the effect of the pixel sampling percentage required for computing a semivariogram using object-oriented image classification techniques has been made. The results show that the use of low percentages of pixels produces slightly lower classification overall accuracy results than considering all image-object pixels. This indicates that the descriptive features extracted from the histogram have a robust behaviour. The methodology of pixel selection, random or stratified, does not have significant influence on the classification results. Besides, the employment of only a proportion of the pixels within image-objects significantly reduces the computation time required for extracting the semivariogram.

Acknowledges

The authors appreciate the financial support provided by the Spanish Ministry of Science and Innovation and the FEDER in the framework of the Projects CGL2009-14220-C02-01 and CGL2010-19591/BTE.

References

[1] P.M. Atkinson, P. Lewis, 2000, Geostatistical classification for remote sensing: an introduction. Computers and Geosciences 26, 361-371.

[2] A. Balaguer, L.A. Ruiz, T. Hermosilla, J.A. Recio, 2010, Definition of a comprehensive set of texture semivariogram features and their evaluation for object-oriented image classification. Computers and Geosciences 36, 231-240.

[3] J.S. Borak, and A.H. Strahler, 1999, Feature selection and land cover classification of a MODIS-like data set for a semiarid environment. International Journal of Remote Sensing 20, 919-938.

[4] J.R. Carr, F.P. Miranda, 1998, The semivariogram in comparison to the cooccurrence matrix for classification of image texture. IEEE Transactions on Geoscience and Remote Sensing 36, 1945-1952.

[5] P. Curran, 1988, The semivariogram in remote sensing: an introduction: Remote Sensing of Environment 24, 493-507.

[6] J.P. Chilés, P. Delfinder, 1999, Geostatistics. Modeling Spatial Uncertainty, John Wiley and Sons, New York.

[7] Y. Freund, 1995, Boosting a weak learning algorithm for majority. Information and Computation 121(2), 256-285.

[8] P. Goovaerts, 1997, Geostatistics for Natural Resources Evaluation. Oxford University Press: New York.

[9] E.H. Isaaks, R.M. Srivastava, 1989, An introduction to applied geostatistics. Oxford.

[10] D.K. McIver, M.A. Friedl, 2002, Using prior probabilities in decision tree classification of remotely sensed data. Remote Sensing of Environment 81, 253-261.

[11] C. Portalés, N. Boronat-Zarceño, J.E. Pardo-Pascual, A. Balaguer-Beser, 2009, Seasonal precipitation interpolation at the valencia region with multivariate methods using geographic and topographic information. Int. J. Climatol. DOI: 10.1002/joc.1988.

[12] M.J. Pyrcz, C.V. Deutsch, 2003, The Whole Story on the Hole Effect. In: Searston, S. (Eds.) Geostatistical Association of Australasia, Newsletter 18.

[13] J.R. Quinlan, 1993, C4.5: Programs For Machine Learning. Morgan Kaufmann, Los Altos.

[14] J. Recio, 2009, Técnicas de extracción de características y clasificación de imágenes orientada a objetos aplicadas a la actualización de bases de datos de ocupación del suelo, PhD Thesis. Universidad Politécnica de Valencia, Valencia, Spain, 289 p.

[15] L.A. Ruiz, J.A. Recio, T. Hermosilla, 2007, Methods for automatic extraction of regularity patterns and its application to object-oriented image classification. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI, Munich, Germany, 117-121.

[16] M. Story, R. G. Congalton, 1986, Accuracy assessment: a user's perspective, Photogrammetric Engineering and Remote Sensing, 52(3), 397-399.