Volume-4, Issue-4, April-2015 • ISSN No 2277 - 8160

Research Paper Engineering

Correlation between digital Landsat levels With Uranium content in soil using Support Vector Machines Regression Algorithms

Luis Hernán Ochoa Associated Professor - Universidad Nacional de Colombia - Colombia

Luis Hernán Alvarez Support Engineer - Universidad Jorge Tadeo Lozano - Colombia

ABSTRACT

Geophysical and Geochemical Field surveying is a hardly and costly activity where remote sensing tools can be a very useful tool to increase coverage and resolution in that kind of studies. Here we present the results obtained using a Support Vector Machine Regression model that relates contents of Uranium measured in laboratory to field soil samples withlocation coordinates and digital Landsat levels at the sample site. Soil samples were taken during a geochemical field survey made by the "InstitutoColombiano de Geología y Minería - INGEOMINAS"in the Vichada and Guainía Departments - Colombia in 2006. The maximum correlation factor obtained was 0.67 which is quite low but indicates the presence of an important relation between Uranium content and digital Landsat levels. This research can be considered as an important contribution to the classical geostatistical models because let's to increase model resolution in areas where sampling density is low getting valuable information from inaccessible areas using remote sensing higher spatial variability

KEYWORDS : Uranium,Landsat, Regression, Support Vector Machines, Mineral Prospecting, Machine Learning, Unsupervised Classification

Introduction

The study of our planet is based on multiple disciplines interacting in a holistic manner, providing a complete set of tools tosolve multiple problems related to a wide range of problems, especiallyUranium prospecting, which is essential for the economic and technological development.

Mineral depositprospecting is an activity that requires field surveying to collect soil and rock samples to be analyzed in laboratory,where concentrations of a particular mineral can be measured directly. Spatial distribution of these geo-referenced valuescan be modeledusing geostatistical tools to obtain distribution maps, which are the key for the discovery of economically viable prospects.Geostatisticaltools were developed to calculate models with a good similarity with real worldbehavior, based on solid statistical principles;however, the accuracy of these models is affected by the amount and distribution of data taken directly in field, mainly because field survey is costly and some sites are inaccessible due to logistical and/or security issues, makingthat most of thedata sets don't have enough data points toreach statistical requirements for development of reliable models.

Due to rapid development of technology over the past years, actually, it is possible to have large amounts of digital information, which is not used properly despite its high availability and low cost. Multiple Data Mining algorithms have been developed to extract valid mathematical relations between multiple variables using large volumes of information. Most of these algorithms are based on strong statistical principles and use concepts and strategies of artificial intelligence and bio-inspired algorithms, which have become a very useful analysis tool.

The presence of some minerals in the soil affect vegetation health in different ways and can be correlated with bands 3 and 4 of Landsat images, as well as other minor effects on other bands of the same remote sensing data set.This is the reason why we proposethe use of Data Mining algorithms to find a correlation between combination of digital levels from different bands of Landsat images in the field sample site, which is known as spectral signature, with the Uranium content values obtained in laboratory to improve the determination of the spatial variability of the prospected mineralusing the large amount of information present in satellite images. This correlation will lead to construct a model based on the information obtained from the eight bands of the Landsat image in those places where samples could not be collected.

The model was developedusingUraniumcontent data, obtained by INGEOMINAS (2006), during an exploration campaign in the departments of Vichada and Guainía; these datawere correlated with the satellite image 0458-100314 GEOCOVER, taken on February 12, 2004. Several regression models were generated using Support Vector Machinesto select the most acceptable correlation, establishing that it is possible to find a mathematical relation for future applications.

Background

Research work associated with spatial distribution of geophysical and geochemical variables, obtained on field surveys, has been developed over time, giving origin to disciplines such as Geostatistics, which is based on solid and reliable statistical principlesand has become the cornerstone in spatial behavior analysis ofEarth features. However,despite the successful of geostatistical models, these depend heavily on rigorous data statistical restrictions, which are not always possible to achieve because the high cost of field work associated with logistical difficulties and laborious and costly laboratory work.Therefore, is not always possible to get enough number and quality of samples limited by project budget and available time.

Another important field for mineral exploration are remote sensors, that count at present with an important number of satellite images, which are composed bymultiple electromagnetic wavelengths reflectance, exceeding human narrow visible range, which combination of values generates a unique particular spectral signatureto achieve differentiation of manytypes of vegetation, soil, etc. With these images, is possible to achieve spectral values of the areas of study, to obtain spatial variations in these places, leading to have a valuable surface detail of spectral features.

There are many research works related with determination of the effect of methane gas presence in the soil, emerged from deep reservoirs leaks, affecting vegetationhealth, which can be represented by a variation in its spectral signature, in particular in the change of the vegetation index, which is a relation between bands 3 and 4 from Landsat. Authors correlated values of gas methane present in field samples, associated to this index, finding a high correlation and therefore allowing to extend that relationship to sampling areas, generating a model that usesthe high spatial variability in satellite images, obtaining the contents of methane from the reflectance values on sites where sampling was takenimproving the model of distribution of that value in the study area, improving results obtained with classical geostatistics.

With the emergence of intelligent systems and the availability of large amounts of data, now is possible to perform data mining to find rela-

tionships between different variables automatically. Neural Networksalgorithms were used to find relations of spectral signatures with geophysical variables modeling, Krasnopolsky (2003).

Researchassociated with the determination of the subsoil chemical elements effect in the spectral signatures of different types of vegetation has basically focused on hydrocarbonsprospection, through the surface manifestation. Schumacher (2006) carried out a review of soil hydrocarbons content effect inspectral signature of soils, vegetation and sediments andconclude that such effects are caused by chemical and mineralogical changes and are detectable using those modifications of spectral signatures. Petrovic (2008), performed an anomaly detection caused by hydrocarbons presenceworking with ASTER images analysis.Similar models have been documented by Hong Yang, et al (2000), Almeida Filho (2001, 2002), Van der Meer (2002) and Noomen et al (2007).

Van Der Meer (2012), made a compilation of research works related to the influence of elements and compounds present in the soil in the spectral signatures, finding some relationsbetween bands 5 and 7 of Landsat,that can be used to separate materials of argillic hydrothermal alteration, reflecting the presence or absence of hidroxilica absorption bands, or the evidence of the relation of the bands 3 and 1 of Landsat which allows to differentiate materials with presence or absence of iron oxides (FeO), andmany othercombinations of bands that are used for geochemical soil characterizations.

Methodology

Considering the existence of multiple mathematical relations between the properties of Earth materials and spectral features recorded by the remote sensors, it is possible to extend the strategy used formethanegas, to other geochemical and geophysical variables, under the assumption that geochemical components, geophysical field values and other Earth surface properties produce an effect on vegetation and soil, modifying and therefore causing its spectral signature modification in a characteristic pattern.Thanks that,it is possible to use the big amount of available satellite imagery, with high spectral content making necessary the use of new and powerful analysis toolsthat allow to find patterns hidden in a hugenumber of data of many variables. This is possible usingData Mining strategies.

This is the reason why we used intelligent systems algorithms to find a model to relate a large number of remote sensing spectral bands with direct field geophysical and geochemical surveys and laboratory measurements. This model is used to find measured variables in places where no field samples were taken and with all this new information makea more detaileddistribution model. Initially, we used techniques of Artificial Neural Networks and Support Vector Machines, without ruling out other complementary techniques like clustering and some Swarms and Ant Colonies algorithms.

Information used for this research was taken from a study ofGeochemistry and Geophysics in the departments of Vichada and Guainía, Colombia, Ingeominas, 2006.We extracted, at surveyed points, the values of digital levels for different bands from a Landsat image, not only in pixels corresponding to the sampling site but also around it, to compare values at neighborhood involving digital levels around the sampling area. Then, we implemented correlation models based on Support Vector Machine algorithms which made possible to find basic correlations that we presented in this work.

Development

Developments of the model and correlation parameters founded are presented below:

Data Collection

Collected data belong to a geochemical prospectioncampaign,madeby INGEOMINAS (2006), to find optimal areas for mineral exploration in the East Colombian through geochemical modeling. In Figure 1we present the general study area. Soil samples were taken during the survey for goldand Uraniumlaboratory contentdetermination, aswell as some direct radiometry measurements, in different places. For this research, Uranium content was selected, as study variable because this element can present much greater impact for vegetation and soil reflectance values than the other measured variables.

Sampling sites belongto an area covered by at least 6 satellite Landsat images. Only the area corresponding to the image Landsat Row 58 Path 04, will be used to create the model because it's an area covered by 706 Uranium samples, considering a large enough number to make an exploration of mathematical relations through data mining. Sampling sites can be observed in figure 2.

Figure 1. General location

An application for automatic extraction of digital levelswas developed. This application reads the pixel value of the image that corresponds to the sample site, and also the digital levels of neighbors. figure 3 presents the main GUI of the program.

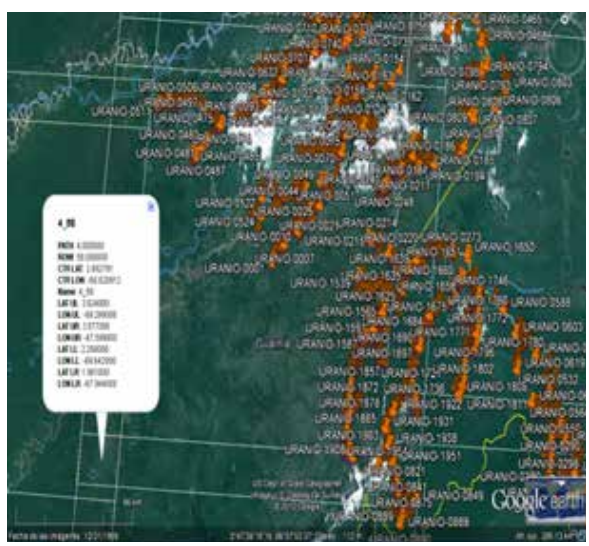


Figure 2. Samples Location and Landsat images.

Unsupervised classification

Sampling sites arelocated in different types of ground. That's the reason why a previous classification was made, to ensure homogeneity i*n the coverage of the ground in order to have a more stable model*, preferably associated withsimilar vegetation. The first clustering was made with the k-Means algorithm, generating 4 *homogenous groups. F*igure 4 shows different groups in which were divided the spectral *signatures. The* left upper groupcorresponds to cloudiness areas, the right upper and also lower left groupsare related with ground and finally the lower right groupcorresponds to mixed vegetation.This last gather of spectral values,corresponding to vegetation

cluster,were taken to perform the model.

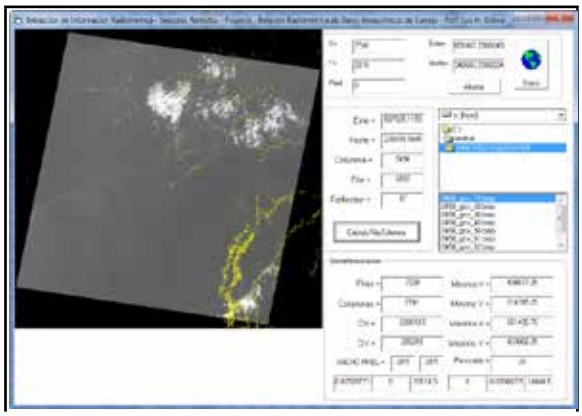


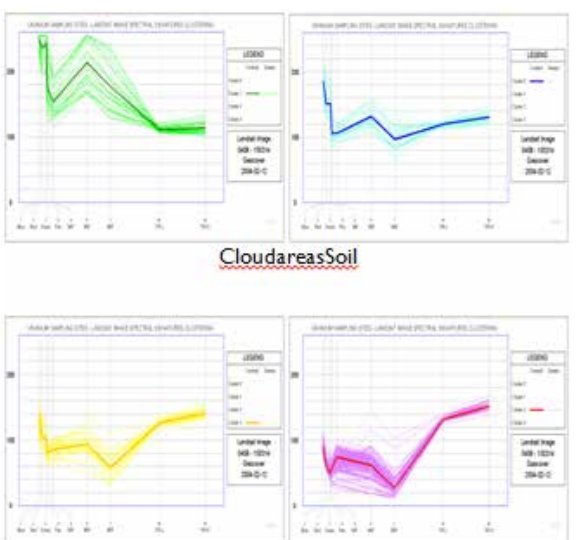
Figure 3.Application - Digital Levels Extractor

Initial cluster Reclassification

In the cluster associated with vegetation, showed in figure 4 (lower-*right),exist* some spectral signatures which are not associated with this feature. For this reason, it was necessary to make a new clustering in order to purge the data before to make the finalmodel.Figure 5 shows the new clustering results for the initial vegetation group. *There* are three new classes, two of them have *a* central spectral *signature (*thicker line) very similar to vegetation and the other corresponding to cluster 0 which is related to abnormal values that must beignored for the final model. The values presented in figure 6 were*chosen to* develop the final regression model based on Support Vector Machines, all of them belong to vegetation coverage with a high grade of certainty.

Experimentation

Using data selected in the preceding chapter, many experiments were made to determine the values of the parameters of interpolation with which best fit is obtained and was executed with free software(WEKA)in which data and predictive modeling were analyzed.



Soil Vegetation

Figure 4. Unsupervised classification

Tests and Results

Determination of optimum model parameters for the relations between values of the Uraniumcontent logarithm with reflectance of Landsat bands and the position of the sample was developed by the execution of regression models based on Support Vector Machines for each of the desired set of parameters evaluating each of them with the value of the statistical correlation obtained from the comparison of real and predicted values of the Uranium content. A Normalized Polynomial Kernel was used for different combinations ofKernel Exponentand complexity factor values calculating correlations with a 10 fold cross-validation scheme.

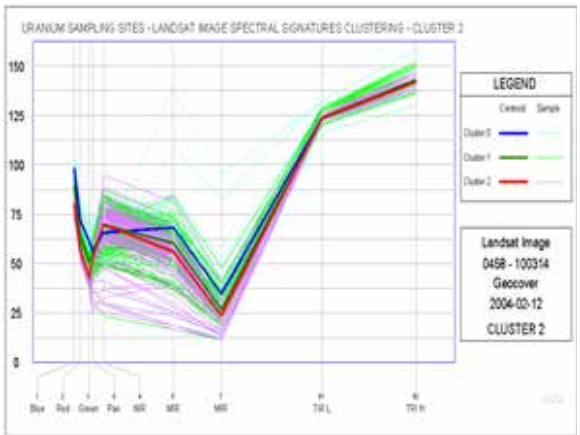


Figure 5. Spectral signatures - first grouping

Three different tests were done for values of complexity factor 1, 0.1 and 0.01. Data subsets were selected usingthen Euclidean distance of each spectral signature to the average spectral signature or centroid as selection rule,to evaluatethe dispersion effect of selectedspectral signatures. The numbers of data points for each subset, based on euclidean distance to centroid, were 170,433, 513, 537, 551, 568, 573 and 583. For each complexityfactor, different kernel exponents (2, 5, 10, 20, 30, 35, 40, 50, 60, 80 and 100) and various distances to central spectral signature of total group (10, 15, 20, 25, 30, 35, 50 and 72), were used. Finally, for each combination of parameters the model correlation was used as evaluation parameter.

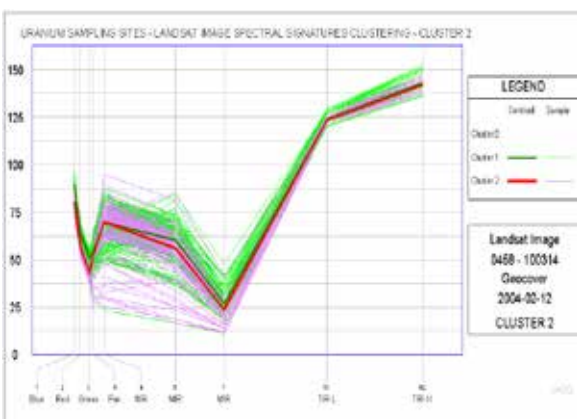


Figure 6. Selected spectral signatures

Figure 7 shows the variation of correlation values for complexity factors equal to 1, 0.1 and 0.01 respectively. Comparing the exponent of the kernel and the Euclidean distance to the spectral signature centroid which controls the number of data points used in each subset of data.

Maximum correlation factor values for each case are summarized in table 1. Note that for the factor of complexity C=1, the distribution of correlation factors presents two local maxima, and for the complexity factor of C=0.1 and C=0.001 there is a single global maximum.

The maximum value of correlation wasobtained for an exponent of 30 and a distance to the centroid of 25, working in this case with a subset of data from 537, which means a 92% of the total number of available information

It can be seen in Figure 7, which corresponds to the maximum values of correlation, that the increase in the number of data from 92% does not generate improvements in the correlation factor whereas the decrease of the subset of training data, generates an accelerated diminutionof the correlation factor.

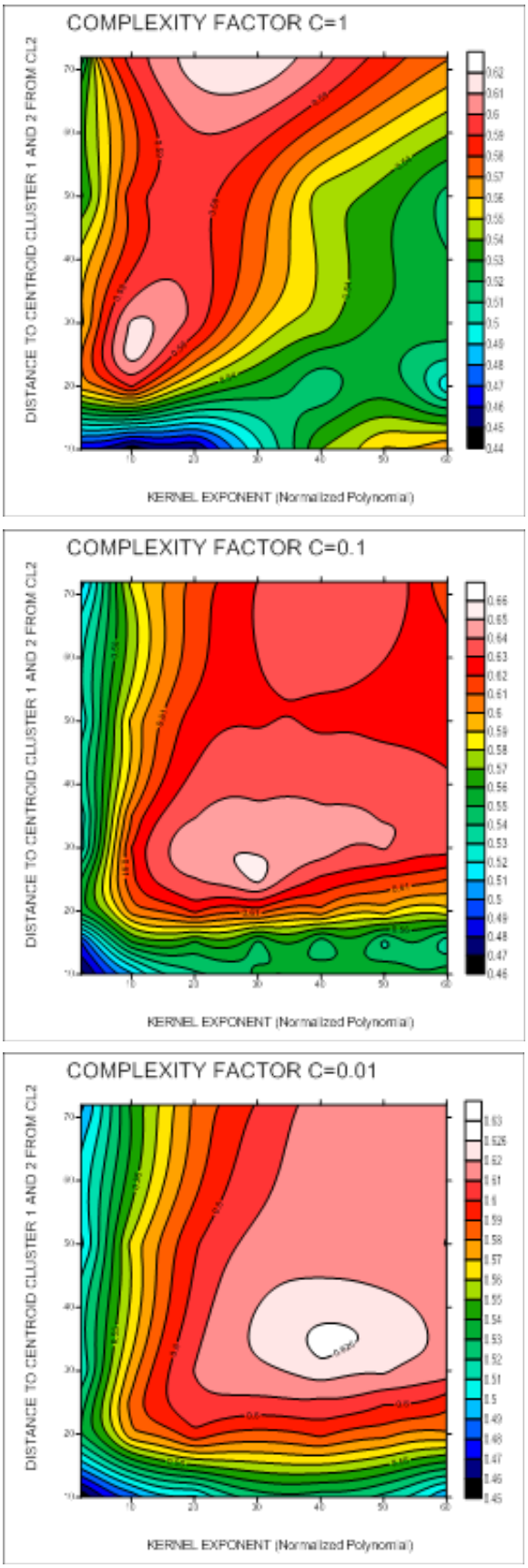


Figure 7. Factor of correlation for various Kernel parameters

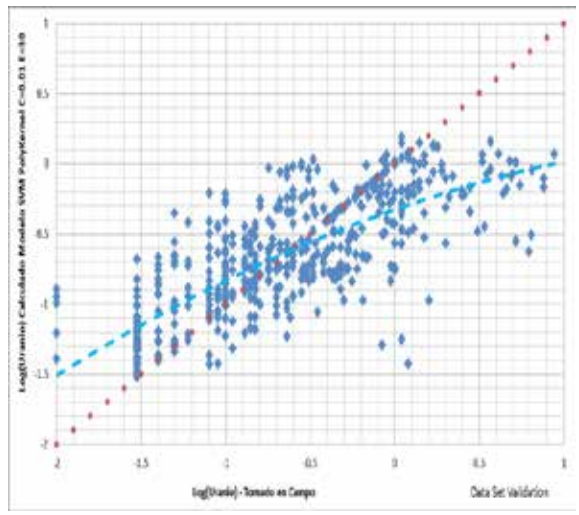


Figure 8.Model Vs Laboratory Uranium Content

Once obtained the optimal values for the parameters of regression model the values of Uraniumof each sample were obtained and were compared with the values measured in the laboratory. Figure 10 shows a comparative plot between those values.For a perfect modelplot, it is expected that pointsbe over a line at 45 degrees, where the values of the model beequal to the values measured in the laboratory. Although, our model is not perfect but shows a tendency to follow this ideal behavior and is a model that allows to get values of Uranium from digital Landsatlevels and the geographical location of the point.In general, our model predicted values tend to be higheratlow Uranium contents and vice versa.

Table 1. Overview of factors of correlation

Complexity factor	Exponent Kernel NPK	Distance to the centroid	Number of data	Correlation factor
1	30	72	583	0.6193
1	10	30	551	0.6117
0.1	30	25	537	0.6532
0.01	40	30	551	0.6246

Conclusions

The regression model based on Support Vector Machine, between radiometric values of Landsat Imageswithlaboratory measured Uraniumcontent of field samples, allowed to obtain a correlation factor of 0.65, a relatively low result, but that represents the effect of the Uranium content in spectral signatures, specially vegetation areas, allowing its correlation and modeling to show a trend tomake more detailed geochemical models using a relatively low number of samples to be correlated with Landsat spectral signatures in a bigger number of points where model can be applied .

The correlation parametersmaybe better forareas with higher Uranium values whichcan affect the vegetation in a more intense way, and thus allow amuch higher correlation, because the spectral signatures of vegetation could be much more distinguishable than those employed in this research

The effects of the above statement, allow us to deduce that a decrease in field samples required for the modeling, which in turn affects directly and implicit in a significant reduction in costs and time in large areas, also if consider unique conditions and characteristics of territory (problems of public order) as it is in our case will be able to achieve a remarkable profit in the improvement of the quality and resolution of the distribution of the variables of prospecting models.

The methodology can be applied to other minerals and compounds of interest.

We recommend that some corrections can be applied to images to obtain values of reflectance that can have a better correlation because some adverse effects can be eliminated.

It is recommended to extend the scheme proposed to the other available Landsat images with the rest of the samples.

REFERENCES

Almeida Filho, Almeida-R, Miranda P. Etal., RADARSAT-1 Images In Support Of Petroleum Exploration: The Offshore Amazon River Mouth Example, Canadian Journal of Remote Sensing, Vol. 31, no. 4, 2001, pp. 289-303. | Krasnopolsky V.M., H. Schiller, Some Neural Network Applications In Environmental Sciences. Part I: Forward And Inverse Problems In Geophysical Remote Measurements, Neural Networks, not 16., 2003, pp 321-334. | Noomen, M.F., Skidmore, A.K., van der Meer, F.D., K.L., Steven, Smith, M.D. and Colls, J.J., The influence of gas pipeline leakage on plant development and reflectance. In: ACRS 2004: Proceedings of the 25th Asian conference on remote sensing, ACRS 2004 Silver jubilee: Chiang Mai, Thailand, November 22-26, 2004, Vol. 1. and 2, pp. 637-642 | Noomen, MF, Van Der Werff H.M.A., Van Der Meer F.D. To Be Submitted To Terra Nova, Circular Detecting Patterns In Vegetation Canopies Caused By Underground Natural Gas Seepage. | Petrovic, a., Khan S.D., Chafetz, H.S., Remote Detection And Geochemical Studies For Finding hydrocarbon-Induced Alterations In Lisbon Valley, Utah, Marine And Petroleum Geology, 2008, Vol 25, pp 696-750. | Corner, E.C., L.G., (2006). Determination of optimal zones for exploration in eastern Colombia through geochemical modeling. Tech-nical report, Ingeominas. | D. Schumacher, Hydrocarbon Microseepage - A Significant But Underutilized Geologic Principle With Broad Applications For Oil/Gas Exploration And Production, 1996. | Van Der Meer F.D. Etal., Multi And Hyperspectral Geologic Remote Sensing: A Review. International Journal Of Applied Earth Observation And Geoinformation, Vol. 14, pp 112-128, 2012. | Vicente-Serrano S.M., Cuadrat-Prats, Romo a., Aridity Influence On Vegetation Patterns In The Middle Ebro Valey (Spain): Evaluation By Meenas Of AVHRR Images And Climate Interpolation Techniques. Journal Of Arid Environments, Vol. 66, pp 353-375, 2006. | Yang, Hong Imaging Spectrometry For Hydrocarbon Microseepage, ISBN 90611641721, ITC Publication, not 76. 1999 |