Fuzzy Classification of Surface Geochemistry Data Applied to the Determination of HC Anomalies

Alexandre G. Evsukoff
NTT - COPPE/UFRJ,
Rio de Janeiro, Brazil
evsukoff@coc.ufrj.br

Felix T. T. Gonçalves
GIMAB/LAB2M -
COPPE/UFRJ,
Rio de Janeiro, Brazil
felix.goncalves@
lab2m.coppe.ufrj.br

Ricardo P. Bedregal
GIMAB/LAB2M -
COPPE/UFRJ,
Rio de Janeiro, Brazil
ricardo.bedregal@
lab2m.coppe.ufrj.br

Nelson F. F. Ebecken
NTT - COPPE/UFRJ,
Rio de Janeiro, Brazil
nelson@ntt.ufrj.br

Abstract: **In this study, fuzzy reasoning numerical techniques were applied to integrate surface geochemical (headspace C1 to C6+ concentrations from soil samples) and geologic data in a Sub-Andean sedimentary basin. A methodology is proposed to compute anomalous regions combining Fuzzy c-Means clustering and fuzzy classifiers. The results of the proposed approach have allowed a good definition of areas anomalous areas, taking into account all the geochemical parameters in an integrated way.**

Index Terms: Fuzzy cluster analysis, Fuzzy classification, Fuzzy geo processing, Surface geochemistry.

I. Introduction

Surface geochemical methods use surface or near-surface occurrences of hydrocarbons (micro seepage) as clues to the location of oil and gas accumulations. The rationale of such methods is that hydrocarbons are generated and/or trapped at depth and leak in varying quantities to the surface. Surface geochemical surveys provide direct evidence of the existence of an active petroleum system, helping in the identification of most prospective areas and in the evaluation and ranking of exploration leads and prospects.

During the last decades, a remarkable advance in analytical techniques has allowed the detection of minute traces of hydrocarbons. Conversely, interpretative methods have been mostly limited to straightforward statistical approaches that define background and anomalous hydrocarbon concentrations assuming a lognormal or normal distribution.

In this study, fuzzy reasoning techniques were applied to locate surface geochemical (headspace C1 to C6+ concentrations from soil samples) data in a sub Andean sedimentary basin.

Fuzzy reasoning techniques are a key for human-friendly computerized devices, allowing symbolic generalization of high amount of data by fuzzy sets and allowing its interpretation by domain experts [1].

A fuzzy geo-processing methodology is proposed to compute anomalous regions (see Fig. 1). Firstly, clusters of similar geochemical values are computed by the fuzzy c-means algorithm disregarding the location of the samples. In a second phase, a fuzzy classifier is trained to recognize the anomaly region generated by cluster analysis. The inputs to the classifier are sample's coordinates and the outputs are the classes identified in the cluster analysis. Finally, a grid of geographic coordinates is generated to cover the whole domain and the fuzzy classifier is used to map the clusters into the grid.

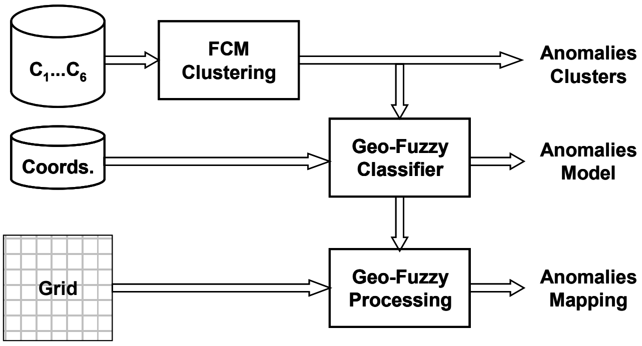

Fig. 1: Fuzzy Geo-processing methodology.

The paper is organised as follows: next section discussed the surface geochemistry for oil exploration. Section III presents the fuzzy clustering analysis with the Fuzzy c-Means algorithm. Section IV presents the fuzzy classifier, referred as geo-fuzzy classifier, which is used to locate clusters of similar headspace samples in the geographic domain. Section V presents the results computed with a data set from a sub Andean basin. Finally, the conclusions and future extensions of this work are highlighted.

II. Surface Geochemistry

Surface geochemical for petroleum exploration is the search for surface or near-surface occurrences of hydrocarbons. It extends through a range of observations from clearly visible oil and gas seepage at one extreme to identification of minute traces of hydrocarbons (micro seepage) or hydrocarbon-induced changes at the other.

The principal objective of a geochemical exploration survey is to establish the presence and distribution of hydrocarbons in the area, in order to help determining the location of petroleum accumulation in subsurface. If the objective is to evaluate individual exploration leads and prospects, the results of geochemical surveys can lead to better risk assessment by identifying those associated with strong hydrocarbon anomalies, thereby improving prospects on the basis of their probable hydrocarbon charge.

The underlying assumption of all near-surface geochemical exploration techniques is that hydrocarbons are generated and/or trapped at depth and leak in varying but detectable quantities to the surface. This has long been an established fact, and the close association of surface geochemical anomalies with faults, productive fairways, and specific leads and prospects is well known. It is further assumed, or at least implied, that the anomaly at the surface can be reliably related to a petroleum accumulation at depth (Fig. 2). Hydrocarbon gases diffuse and flow from deep-seated petroleum accumulations through the sedimentary rock and reach the surface, being absorbed by the soil particles.

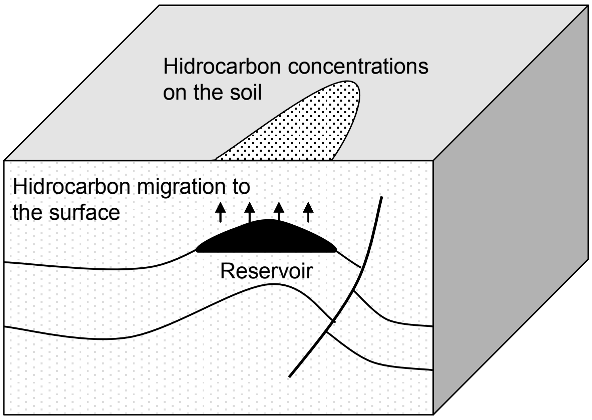
Hidrocarbon concentrations on the soil
Hidrocarbon migration to the surface
Reservoir

Fig. 2: Surface geochemical prospecting.

III. Fuzzy c-Means Cluster Analysis

In the first step of the proposed methodology, a cluster analysis is performed to find similar patterns in the geochemical data. The geochemical data, regardless of their location are used as input to the Fuzzy c-means algorithm as described next.

A. *Fuzzy c-Means Algorithm*

The Fuzzy c-means (FCM) algorithm proposed by Bezdek [2], is the well known fuzzy version of the classical ISODATA clustering algorithm.

Consider the data set $T = \{(\mathbf{x}(t)), t = 1..N\}$, where each sample contain the hydrocarbons' concentration data, represented by the vector $\mathbf{x}(t) \in R^p$. The algorithm aims to find a fuzzy partition of the domain into a set of K clusters $\{C_1 \ldots C_K\}$, where each cluster C_i is represented by its center's coordinates vector $\mathbf{w}_i \in R^p$.

In the fuzzy cluster analysis, each sample in the training set can be assigned to more that one cluster, according to a value $v_i(t) = \mu_{C_i}(\mathbf{x}(t))$, that defines the membership of the sample $\mathbf{x}(t)$ to the cluster C_i.

The FCM algorithm computes the centers' coordinates by minimizing the objective function J defined as:

$$J(m, \mathbf{W}) = \sum_{t=1..N} \sum_{i=1..K} v_i(t)^m \, d(\mathbf{x}(t), \mathbf{w}_i)^2 \tag{1}$$

where $m > 1$, generally referred as the "fuzziness parameter", is a parameters to adjust the effect of membership values and $d(\mathbf{x}(t), \mathbf{w}_i)$ is the Euclidean distance from the sample $\mathbf{x}(t)$ to the cluster center $\mathbf{w}_i$.

The membership of all samples to all clusters defines a *partition matrix* as:

$$\mathbf{V} = \begin{bmatrix} v_1(1) & \cdots & v_K(1) \\ \vdots & \ddots & \vdots \\ v_1(N) & \cdots & v_K(N) \end{bmatrix}. \tag{2}$$

The partition matrix is computed by the algorithm such that:

$$\forall \mathbf{x}(t) \in T, \; \sum_{i=1..K} v_i(t) = 1. \tag{3}$$

The FCM algorithm computes interactively the clusters centers coordinates from a previous estimate of the partition matrix as:

$$\mathbf{w}_i = \frac{\sum_{t=1..N} v_i(t)^m . \mathbf{x}(t)}{\sum_{t=1..N} v_i(t)^m}. \tag{4}$$

The partition matrix is updated as:

$$v_i(t) = \frac{1}{\sum_{j=1..K} \left(\frac{d(\mathbf{x}(t), \mathbf{w}_i)}{d(\mathbf{x}(t), \mathbf{w}_j)} \right)^{\frac{2}{(m-1)}}}. \tag{5}$$

The FCM algorithm is described as follows:

0. Set $m > 1$, $K \geq 2$ and initialize the cluster centers' coordinates randomly, initialize the partition matrix as (5).
1. For all clusters $(2 \leq i \leq K)$, update cluster centers coordinates as (4).
2. For all samples $(1 \leq t \leq N)$ and all clusters $(2 \leq i \leq K)$, update the partition matrix as (5).
3. Stop when the norm of the overall difference in the partition matrix between the current and the previous iteration is smaller than a given threshold ε; otherwise go to step 1.

The FCM algorithm computes clusters centers' coordinates and the partition matrix from the specification of the number of clusters K, that must be given in advance. In practice, the FCM algorithm is executed to various values of K, and the results are evaluated by a cluster validity function, as described next.

B. Cluster validity

Many cluster validity criteria have been proposed in the literature in the last years ([4], [5] and [6]). Validity indexes aim to answer two important questions in cluster analysis: (i) how many clusters are actually present and (ii) how good the partition is.

The main idea present in many of the validity indexes is based on the geometric structure of the partition, such that samples within the same cluster should be compact and different clusters should be separate. When the cluster analysis assigns fuzzy membership functions to the clusters, "fuzziness" must be taken into account in such a way that the less fuzzy the partition is the better it is.

In this work, the recently proposed PBM index [6] is used to evaluate the number of clusters in the data set. The PBM index is defined as a product of three factors, of which the maximization ensures that the partition has a small number of compact clusters with large separation between at least two of them. Mathematically the PBM index is defined as follows:

$$PBM(K)=\left(\frac{1}{K}\cdot\frac{E_1}{E_K}.D_K\right)^2 \tag{6}$$

where K is the number of clusters; E_1 is the sum of the distances of each sample to the geometric center of all samples $\mathbf{w}_0$ as:

$$E_1=\sum_{t=1..N} d(\mathbf{x}(t),\mathbf{w}_0). \tag{7}$$

E_K is the sum of within cluster distances of K clusters, weighted by the corresponding membership value:

$$E_K=\sum_{t=1..N}\sum_{i=1..K} v_i(t)d(\mathbf{x}(t),\mathbf{w}_i)^2 \tag{8}$$

and D_K that represents the maximum separation of each pair of clusters:

$$D_K=\max_{i,j=1..K}\left(d(\mathbf{w}_i,\mathbf{w}_j)\right). \tag{9}$$

The greater the PBM index is, the better is the cluster fuzzy partition. As other indexes, the PBM index is an optimizing index, such that it can be used to search the best number of clusters. The PBM index has achieved a good performance in several data sets [6] when compared with the Xie-Beni index [4]. This index is thus used as a validity index of the methodology presented in this work.

IV. RULE-BASED FUZZY CLASSIFIER

In the second step of the proposed methodology, the clusters generated by the FCM algorithm are used as classes to train a rule-based fuzzy model of the anomalies. The resulting fuzzy classifier is referred as geo-fuzzy classifier since it computes the membership to classes $\mathbf{C}=\{C_1\ldots C_K\}$, form the geographic coordinates vector $\mathbf{y}(t)=\left(y_1(t),y_2(t)\right)$ given as input.

A. Geo-Fuzzy Classifier

Each coordinate $y_i(t)$ is described by a fuzzy partition $\mathbf{A}_i=\{A_{i1},\ldots,A_{in}\}$ where $A_{ij}\in\mathbf{A}_i$ is a fuzzy set. The number of fuzzy set for each coordinate is set the same to simplify the computations.

Strong normalized and triangular fuzzy partitions are used to represent each input variable. Trapezoidal membership functions are used for the two fuzzy sets at each end of the domain, as shown in Fig. 3, to deal with off-limit points.

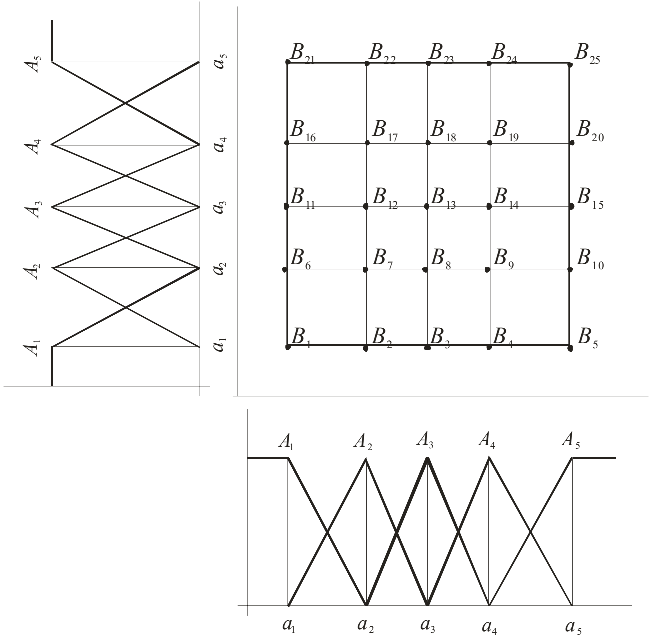
A_5
A_4
A_3
A_2
A_1
a_5
a_4
a_3
a_2
a_1
B_{21} B_{22} B_{23} B_{24} B_{25}
B_{16} B_{17} B_{18} B_{19} B_{20}
B_{11} B_{12} B_{13} B_{14} B_{15}
B_6 B_7 B_8 B_9 B_{10}
B_1 B_2 B_3 B_4 B_5
A_1 A_2 A_3 A_4 A_5
a_1 a_2 a_3 a_4 a_5

Fig. 3: Construction of the geographic fuzzy set.

The fuzzy rule base relates input fuzzy sets to the classes, in rules like:

$$\textit{if } \mathbf{y}(t) \textit{ is } B_k \textit{ then class is } C_j \textit{ with } cf=\varphi_{kj} \tag{10}$$

The fuzzy set B_k in rule (10) represents the combination of the fuzzy sets in the partition of each coordinates (see Fig. 3) and defines a geographic region. For a given input $\mathbf{y}(t)=\left(y_1(t),y_2(t)\right)$, all the combinations of fuzzy sets in each fuzzy partition must be considered in such a way that the model is complete, *i.e.* it produces an output for whatever input values. In Fig. 3, the combination of two fuzzy partitions of 5 fuzzy sets each is shown.

Each component $u_k(t)=\mu_{B_k}(t)$ of the fuzzification vector $\mathbf{u}(t)$ is computed as:

$$u_k(t)=\mu_{A_{1i}}(y_1(t))\mu_{A_{2j}}(y_2(t)),\ i,j=1...n. \tag{11}$$

The confidence factor $\varphi_{kj} \in [0,1]$ in rule (10) represents the rule certainty. The confidence factor weight all rules in the fuzzy rule base. The value φ_{kj} represents how much the term B_k is related to the class C_j in the model described by the rule base.

The rule base can be represented by the matrix $\Phi = \left[\varphi_{kj}\right]$, of which each line is related to a geographic fuzzy sets B_k and each column is related to a class C_j.

The rule base weights are the kernel of the model described by the fuzzy rules (10) and its determination is computed as described next.

B. Rule Base Identification

The rule base weight are computed from a data set T', where each sample $t = 1..N$ is a pair $\left(\mathbf{y}(t), \mathbf{v}(t)\right)$, of which $\mathbf{y}(t)$ is the coordinates vector and $\mathbf{v}(t) = \left(v_1(t) \ldots v_K(t)\right)$ is a row in the partition matrix (2) computed by the FCM algorithm. Each sample t in the data set T' is related to the hydrocarbons' concentration data in the sample t of the data set T.

Each rule in the rule base is a sub-model that assigns a class (computed in fuzzy cluster analysis) to the corresponding region of the domain. Each rule base weight φ_{kj} can be seen as a measure of how frequent the class C_j occurs in the region B_k. Under this interpretation [3], the rule base weights are computed as:

$$\varphi_{kj} = \frac{\sum_{t=1..N} u_k(t) v_j(t)}{\sum_{t=1..N} u_k(t)} \tag{12}$$

where $u_k(t)$ is the membership of the register t to the geographic fuzzy set B_k, computed as (11) and $v_j(t) = \mu_{C_j}(\mathbf{x}(t))$, i.e. the membership of the hydrocarbons' concentration data in the register t to the cluster C_j.

C. Geo-Fuzzy Processing

In the third and last step of the methodology, a grid of testing points $\mathbf{z}(t) = \left(z_1(t), z_2(t)\right)$ is generated to create a map of the clusters of concentration data into the geographic domain. The grid is represented by a testing set $T'' = \left\{\left(\mathbf{z}(t)\right), t = 1..M\right\}$, where M is the number of registers in the grid.

The geo-fuzzy processing aims to compute the output of the geo-fuzzy classifier to each point of the grid, i.e. the class membership vector $\hat{\mathbf{v}}(t) = \left(\mu_{C_1}(\mathbf{z}(t)), \ldots, \mu_{C_K}(\mathbf{z}(t))\right)$, where $\mu_{C_j}(\mathbf{z}(t))$ is the output membership value of the grid coordinates $\mathbf{z}(t)$ to the class C_j.

The fuzzification vector $\hat{\mathbf{u}}(t)$ is computed for every point in the grid. Each component of the fuzzification vector is computed as product of the membership of each coordinate value to the respective fuzzy partition:

$$\hat{u}_k(t) = \mu_{A_{1i}}(z_1(t)) \mu_{A_{2j}}(z_2(t)),\ i, j = 1 \ldots n. \tag{13}$$

The class membership vector is computed from the input membership vector $\hat{\mathbf{u}}(t)$ and the rule base weights matrix Φ. Using the sum-product composition operator for the fuzzy inference, the class membership vector $\hat{\mathbf{v}}(t)$ can be computed as a standard vector matrix product as:

$$\hat{\mathbf{v}}(t) = \hat{\mathbf{u}}(t).\Phi \tag{14}$$

The number fuzzy sets as so as the number of points in the grid controls the accuracy of the map generated.

V. Results and Discussion

This section presents the results of the application of the proposed methodology to a real data set composed of $N = 350$ samples containing headspace concentration of $p = 8$ hydrocarbons (C1 to C6+) and their respective UTM coordinates. The data set was collected in a sub Andean basin.

A. FCM Cluster Analysis

The FCM cluster analysis was performed with the PBM validation index to determine optimal the number of clusters in the data. As it can be seen in Table 1, for many values of the fuzzy parameter m the validity index always indicates 4 clusters in the data set. Thus, the classification was performed considering four classes.

Table 1: Determination of the number of clusters.

$m = 1.2$	$m = 1.4$	$m = 1.6$	$m = 1.8$	$m = 2.0$
0.3508	0.3415	0.3234	0.2958	0.2606
0.3078	0.2912	0.2633	0.2299	0.1957
0.6428	**0.5850**	**0.5013**	**0.3983**	**0.2778**
0.5301	0.4787	0.4044	0.3239	0.2406
0.4245	0.3948	0.3893	0.2650	0.2298
0.4627	0.3227	0.3468	0.2782	0.2111
0.3870	0.3409	0.3074	0.2542	0.2219
0.3588	0.3229	0.2643	0.2306	0.1867
0.3191	0.2850	0.2485	0.2024	0.1605

The results shown in Table 1 show that for greater values of the parameter m, the PBM index is lower, since the partition is more "fuzzy". The fuzzy partitions allow that a register be assigned to more than one class allowing gradual transitions between clusters. The geo-fuzzy classifier was thus adapted to classify the clusters generated with $m = 1.2$.

B. Classification results

The geo-fuzzy classifier was computed from the results of cluster analysis to map each cluster into the geographic domain. For this application, the rule base was generated using $n = 10$ fuzzy sets to each coordinate, resulting in 100 rules like (10) and a grid of 62500 (250×250) points regularly spaced within the domain.

The results are shown in Fig. 4 to Fig. 7, where class membership is represented by a degree of colors: the green (light gray in the B&W printing) represents null (0.0) membership and the red (dark gray) represents full (1.0) membership.

The figures also show the location of the registers. As it can be seen, for the regions that are not covered by any point, the membership to all classes is always zero.

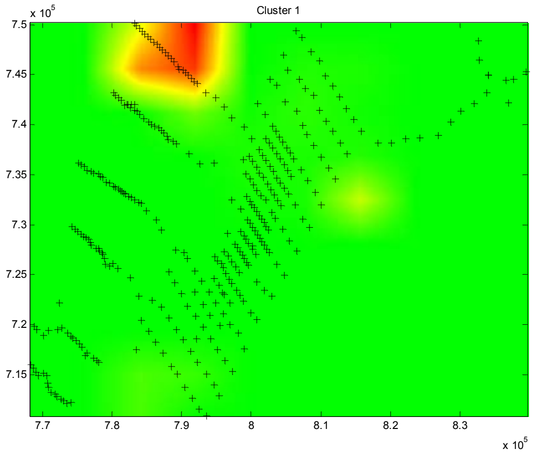

Fig. 4: Map of the cluster 1.

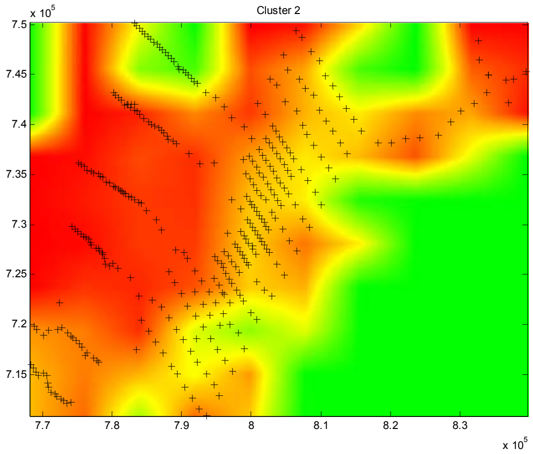

Fig. 5: Map of the cluster 2.

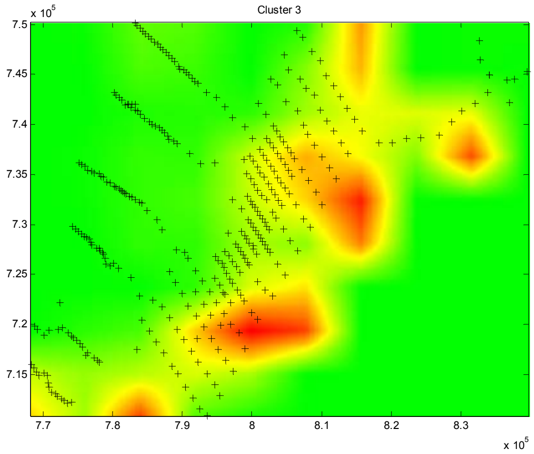

Fig. 6: Map of the cluster 3.

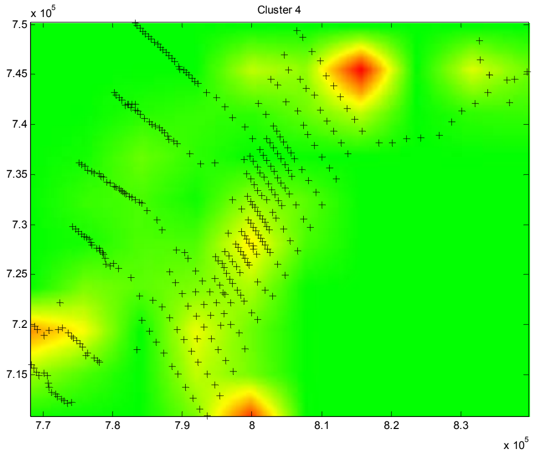

Fig. 7: Map of the cluster 4.

The interpretation of each cluster can be done from their centers' coordinates, shown in Table 2. The columns of Table 2 indicate the hydrocarbons' concentrations (C1 to C6+) that were used in the cluster analysis expressed in ppm (parts per million) units.

Cluster 1, of which the map is shown in Fig. 4, presents the higher concentrations for all gases (from C1 to C6+), supporting the interpretation of a subsurface source for these hydrocarbons (e.g. a oil and gas field).

Table 2: Hydrocarbon concentrations on clusters' centers, expressed in ppm units.

	Methane (C1)	Ethane (C2)	Propane (C3)	i-Butane (iC4)	n-Butane (nC4)	i-Pentane (iC5)	n-Pentane (nC5)	Hexane + (C6+)
Cluster 1	651.3375	91.9572	34.7065	2.7693	11.9485	3.4243	5.3249	2.9407
Cluster 2	97.6568	9.9916	3.5075	0.0156	1.2170	0.0784	0.7412	0.3487
Cluster 3	159.4038	25.1123	10.6860	1.0798	3.7594	1.1578	1.7204	0.9486
Cluster 4	1077.8756	9.5655	3.7722	0.1427	1.2930	0.3128	0.7485	0.3062

Cluster 2, shown in Fig. 5, is the most common concentration of the samples with lower concentrations of hydrocarbons, indicating the absence of significant sources in subsurface or the existence of permeability barriers between the sources and the surface.

Cluster 3 represents intermediary concentrations between cluster 1 and 2. Finally, cluster 4 concentrations are almost similar to cluster 2, presenting only outliers values for variable C1, indicating a possible contribution of biogenic gases generated by the degradation of organic matter in the soil.

The samples can be grouped according to their higher membership values. The absolute and relative number of samples grouped in each cluster is shown in Table 3.

Table 3: Number of samples of each cluster.

	Number	Relative
Cluster 1	21	6.00%
Cluster 2	242	69.14%
Cluster 3	59	16.86%
Cluster 4	28	8.00%

Results in Table 3 are similar to the area of higher membership regions in Fig. 4 to Fig. 7. Cluster 2 is the most common cluster, grouping 69.14% of data, while clusters 1 and 4 are the ones which regroup the least number of samples.

These results may be an indication of an anomaly region in cluster 1, but they must further confirmed by a detailed geologic studies.

VI. Conclusions

This work presented a fuzzy geo-processing methodology to map surface geochemistry similar data. In the first step, the groups of similar geochemistry data are computed by a fuzzy cluster analysis with the FCM algorithm. Geochemistry data are presented by headspace hydrocarbon concentration, without their respective coordinates.

In the second step, a fuzzy classifier is then used to map the clusters into the geographic coordinates. The inputs to the fuzzy classifier are the UTM coordinates of the geochemical registers; the outputs are the clusters recognized by the cluster analysis. The rule base computed by the learning algorithms represents a model of the anomalies location.

The results of the methodology have allowed a better definition of anomalous areas, taking into account all the geochemical parameters in an integrated way (instead of considering concentrations of each gas independently), providing an encouraging alternative to standard geo statistic techniques.

The extension of this work is in the direction of the integration of other fuzzy cluster analysis algorithms as so as the comparison of the results with standard geo statistics techniques.

Acknowledgment

The authors are grateful to the Brazilian Research Council, CNPq for the financial support for this research.

References

[1] L. Zadeh, “Fuzzy logic = computing with words,” *IEEE Trans. on Fuzzy Systems*, vol. 4, no. 2, pp. 103-111, 1996.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, Plenum, 1981.

[3] A. G. Evsukoff, A. C. S. Branco and S. Gentil, “A knowledge aquisition method for fuzzy expert systems in diagnosis problems,” *Proc. 6th IEEE International Conference on Fuzzy Systems – FUZZIEEE’97*, Barcelona, 1997.

[4] X. L. Xie, G. A. Beni, “Validity measure for fuzzy clustering,” *IEEE Trans. Pattern Anal. Machine Intelligence* vol. 3 no. 8, pp. 841–846, 1991.

[5] J. Bezdek, N.R. Pal, “Some new indexes of cluster validity,” *IEEE Trans. Systems Man and Cybernetics B*, vol. 28, pp. 301–315, 1998.

[6] M. K. Pakhira, S. Bandyopadhyay, U. Maulik, “Validity index for crisp and fuzzy clusters,” *Pattern Recognition*, vol. 37, pp. 487-501, 2004.