COMO SELECCIONAR LA VARIANTE OPTIMA DE AGRUPAMIENTO Y CLASIFICACION DE LOS DATOS

RICARDO ALBERTO VALLS ALVAREZ

- Expedición Geológica "Habana - Matanzas"
- Especialista en Geoquímica y modelaje Geomatemático

RESUMEN

El tratamiento previo de los datos antes de su elaboración y procesamiento, es un paso primordial para poder garantizar resultados confiables e interpretaciones exitosas. Al conjunto de estos trabajos pertenece el proceso de clasificación (agrupamiento o estratificación) de los datos, el cual -si no es correcto- puede malograr la investigación desde su inicio. En el presente artículo se explica un sencillo, pero muy seguro modelo geomatemático para seleccionar entre varias variantes de estratificación de los datos la óptima para obtener los mejores resultados del procesamiento de los mismos. También se abordan algunas implicaciones genéticas de este análisis. El método V.O.C. (variante óptima de clasificación), dado su sencillez, efectividad e importancia, debe aplicarse previamente a cualquier procesamiento estadístico de la población.

1. INTRODUCCION

Con las tareas de clasificación previa de los datos, el geólogo tropieza constantemente y si no realiza una correcta estratificación de sus datos (o no realiza clasificación alguna en general), llegar a resultados alejados de la realidad.

Por ejemplo, un geoquímico que procesó los resultados del bario de un muestreo metalométrico en un sector sin haber diferenciado las muestras al menos por tipos litológicos, obtuvo el "prometedor" cuadro de anomalias que se muestra en la Fig. 1 y planificó, en base a estas, trabajos más detallados (y más costosos) de verificación.

En realidad, tal como se aprecia en la Fig. 2, lo que dicho geoquímico obtuvo fue un mapa de anomalías "litofílicas". O sea que mapeó los principales grupos litológicos presentes en el sector de los trabajos que se diferencian por sus clarks (granitoides- 0.083%; tobas andesíticas-0,065 %; calizas- 0.08% y serpentinitas- 0,0001%, según datos de Voitkievich et al. (1970).

También tenemos que recurrir a la clasificación de los datos cuando enfrentamos la tarea de establecer lla especialización geoquímica de los grupos petrológicos del área de estudios.

La clasificación "por tipo petro o litológico" es la más usualmente empleada, aunque puede presentar algunas desventajas, tales como:

- Mala documentación del punto.
- No representatividad estadística de alguna clase, cuando la cantidad de datos correspondientes a la misma es inferior a 20 - 30 unidades.
- La existencia de grupos petrológicos con características similares, lo cual dificulta enormemente su clasificación.

Algunos de estos problemas son casi imposibles de solucionar y pertenecen al así llamado 5% de inseguridad estadística de los trabajos. No obstante, el presente modelo geomatemático le será de gran ayuda, sobre todo en lo correspondiente a la última dificultad enumerada, permitiéndole seleccionar la mejor variante de estratificación, con lo cual se garantiza una alta calidad y objetividad de las futuras investigaciones.

2. MODELO V.O.C.

A continuación se detallan los pasos que componen el modeloV.O.C. (variante óptima de clasificacón) para la correcta estratificación de sus datos.

2.1 Selección de los elementos más informativos

Esta tarea puede ser solucionada de dos formas diferentes. bien, confeccionando varios perfiles geoquímicos para establecer visualmente el comportamiento de los distintos elementos o parámetros analizados (Valls, 1989) y de esa forma seleccionar los que mejor varíen su comportamiento en correspondencia con los cambios geológicos del corte (Fig. 3); o bien determinando mediante las ecuaciones 1, 2 y 3, el coeficiente de variación de cada elemento, seleccionando como "más informativos" aquellos con variaciones que oscilen entre el 50 y el 150% (Valls, 1987). No es recomendable incluir en estos cálculos aquellos elementos donde más del 50% de sus valores sean nulos o trazas.

$$(1)\quad \bar{X} = \frac{1}{2} \sum_{i=1}^{n} X_i$$

$$(2)\quad S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n - 1}}$$

$$(3)\quad V = S.100/\bar{X}$$

Donde,

X_i Es el valor del parámetro analizado en cada muestra.

$\bar{X}$ Es la media aritmética.

n Es la cantidad total de muestras analizadas.

S Es la desviación típica o varianza

V Es el coeficiente de variación del parámetro analizado.

Como puede apreciarse en la Fig. 3, la Plata a diferencia del Bario, mantiene un comportamiento completamente independiente a las variaciones geológicas del corte, razón por la cual no se toma la Plata en calidad de elemento informativo.

Este análisis se hace bien para el elemento más importante -en el caso de tratarse del estudio de un yacimientoo manifestación mineral- o bien de un conjunto de elementos representativos.

En este segundo caso, debemos llegar a obtener un parámetro único representativo del conjunto de elementos seleccionados como los más informativos. En este caso propongo emplear el Coeficiente Correlacional (Valls, 1989) pues es más sencillo que otros modelos geomatemáticos como el de Componente Principal, por sólo citar uno de ellos.

Para determinar el Coeficiente Correlacional (C.C.), se efectúa con los elementos seleccionados un análisis correlacional binario. Aquellos elementos que presenten correlaciones positivas y sensibles se ubicarán multiplicados en el numerador del Coeficiente Correlacional. Cuando la correlación sea inversa y sensible con alguno de los elementos ubicados en el numerador, dicho elemento se ubicará en el denominador del C.C.

Al terminar de distribuir todos los elementos en el C.C., verifique que no exista ninguna contradicción en el mismo, como es el caso de un elemento que pueda ubicarse tanto en el numerador como en el denominador, producto de sus correlaciones positivas o negativas con diferentes elementos del C.C. En esos casos, ambos elementos se excluyen del C.C.

Para ejemplificar lo anterior, en el Anexo I se muestra un caso real, con toda la secuencia de operaciones a efectuar. En dicho ejemplo fueron seleccionados gráficamente tres elementos como los más informativos. La correlación en cada caso se determina (Kazhdan et al., 1979) mediante la ecuación (4), comprobándose su sensibilidad por medio del criterio de Student según la ecuación (5). Si los cálculos se realizan "a mano", la ecuación (6) puede ser más fácil de utilizar para determinar la correlación.

$$(4)\quad r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(n-1)\ S_x \times S_y}$$

$$(5)\quad T_c = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}\ \ [tp, n-2]$$

$$(6)\quad r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]\left[\sum_{i=1}^{n} Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

Donde,

r Es la correlación.

t_c Es el valor crítico de la correlación

$t_{p,n-2}$ Es el valor del criterio de Student para =0,05% y n-2 g1.

2.2 Estratificación de la información

La "estratificación" o clasificación y agrupamiento de los datos iniciales, es un proceso largo y a menudo tedioso, pero sin el cual no se pueden garantizar resultados correctos del proceso investigativo. Es por ello preferible dedicarle al mismo todo el tiempo que sea necesario.

La estratificación exige un gran profesionalismo al investigador quien deberá escoger entre varias opciones de clasificación, desde las más generales a las más detalladas. En el ejemplo representado en la Fig. 2, se podrían proponer las siguientes variantes:

A. Serpentinitas, granitoides, tobas y calizas.

B. Serpentinitas masivas, otras serpentinitas, granitoides, rocas vulcanógeno-sedimentarias.

C. Complejo ofiolítico, tobas y calizas.

Yvarias otras más, en dependencia a los intereses y objetivos del investigador.

2.3 Tratamiento de los valores "huracanados"

La determinación de la existencia de valores huracanados se realiza independientemente en cada variante de estratificación adoptada. Este paso es muy importante pues la presencia de tan sólo uno de estos valores, puede llevar a resultados equívocos con posterioridad.

Existen múltiples formas para determinar la existencia de valores huracanados (Bondarienko et al., 1985), pero para las necesidades de este modelo puede satisfacernos el uso de la ecuación (7).

$$(7)\quad X_{hur} \geq \bar{X} \pm 5\,S$$

Donde,

X_{hur} Son los valores huracanados

$\bar{X}$ Es la media de los valores en la variante analizada (Vid. ecuación 1)

S Es la varianza de los valores en la variante analizada (Vid. ecuación 2).

Una vez determinado el valor límite "X hur", comprobamos si existen en la variante valores iguales o superiores al mismo. Si así fuera, elimine dicho dato de la variante. Si la cantidad de datos en la misma es menor de 30, entonces en vez de eliminar el dato, sustitúyalo por el valor de media ($\bar{X}$) antes determinada.

Las variantes así preparadas ya están listas para ser investigadas en busca de la variante óptima de clasificación.

2.4 Verificación de la calidad de la estratificación

Como criterio estadístico para seleccionar la variante óptima, se tomó el que plantea que una estratificación correcta, la dispersión interna de cada grupo, clase o variante ha de ser la mínima, en tanto que la dispersión entre grupos ha de ser la máxima (Yamane, 1980).

Expresándolo en un lenguaje "menos matemático", pudiera decirse que cada clase ha de ser lo más homogénea posible internamente y lo más diferente posible en comparación con las demás. En otras palabras, se escogerá como variante óptima aquella que presente la mayor diferencia entre sus varianzas internas y externas.

Para efectuar estos cálculos, se emplearán las ecuaciones (8), (9) y (10).

$$(8)\quad W_h = N_h / N$$

$$(9)\quad S^2_w = \sum_{i=1}^{n} W_h \cdot S^2_h$$

$$(10)\quad S^2_b = \sum_{i=1}^{n} W_h \cdot \left(\bar{x}_h - \bar{x}\right)$$

Donde,

N Es la cantidad total de datos

N_h Es el total de datos en el estrato analizado

W_h Es la ponderación del estrato

S^2h Es la varianza del estrato analizado

S^2w Es la varianza interna total

$\bar{X}$ Es la media total de los datos

$\bar{X}_h$ Es la media del estrato analizado y

S^2 b Es la varianza externa total.

En la tabla 1 se muestran los resultados reales de la aplicación del modelo V.O.C. en una mina aurífera en serpentinitas. Como elemento informativo se tomó el oro no sólo por ser el principal elemento de la mineralización, sino también por ser el único elemento determinado cuantitativamente mediante análisis dosimástico, lo cual garantiza una mejor calidad de la información.

Teniendo en cuenta las labores de muestreo realizadas (Leal et al., 1982), fundamenté tres variantes de estratificación:

A. Por tipos petro-estructurales, sin tener en cuenta el nivel de donde fue tomada la muestra (cinco estratos).

B. Por niveles, sin tener en cuenta el tipo petro-estructural (cinco estratos).

C. Y por tipos petro-estructurales, teniendo en cuenta el nivel de donde se tomó la muestra (21 estratos).

Tabla 1. Verificación de la calidad de la estratificación de los datos en una mena aurífera en serpentinitas.

VARIANTES	S^2_w	S^2_b	DIFERENCIA
A	1. 321,21	2. 912,72	1. 591,51
B	2. 100,92	3. 909,59	1. 808,67
C	5. 057,97	5. 766,97	707.00

Los resultados presentados en la tabla 1, permiten seleccionar como la variante óptima de clasificación a la estratificación por niveles, sin tener en cuenta el tipo petro-estructural, por lo cual el resto del análisis lo efectué con los datos agrupados en esa forma. Lo anterior me permitió además suponer (dado el comportamiento indendiente del oro con respecto a los tipos petro-estructurales muestreados) que la génesis de la mineralización aurífera era eminentemente hidrotermalmetasomática ya que afectaba casi por igual a todas las rocas del área y sólo existían variaciones en la medida que nos alejábamos del foco hidrotermal (de ahí la influencia de los niveles de profundidad muestreados). Esta suposición fue colaborada posteriormente, gracias al modelaje geomatemático detallado que efectué en dicha zona mineral.

3. CONCLUSIONES Y RECOMENDACIONES

En los trabajos de investigación geológica, principalmente aquellos que trabajan con grandes volúmenes de datos (geoquímica, geofísica, hidrogeología, etc.), la estratificación o agrupamiento de los datos mejora la calidad de los resultados y permite lograr interpretaciones más seguras y objetivas. Es por ello que debe realizarse siempre este proceso y dedicarle al mismo todo el tiempo que le sea menester.

En los casos en que disponga de más de un elemento o parámetro informativo, recurra al cálculo del Coeficiente Correlacional (C.C.) en base al análisis de correlacional binario de dichos elementos informativos, tal como se explica en detalle en el Anexo I.

En ciertos casos, la aplicación del modelo V.O.C. a una estratificación bien planeada permite llegar a conclusiones importantes, incluso acerca de la génesis de un yacimiento. Es por ello que le recomiendo no proceder de forma mecánica ni a la selección de las variantes, ni a la aplicación de este modelo. Recuerde -en todo caso- que el modelo, por muy eficiente y sencillo que sea, es sólo una herramienta y depende de la maestría del que lo utiliza.

ANEXO I. SECUENCIA DE OPERACIONES PARA DETERMINAR EL COEFICIENTE CORRELACIONAL (C.C.) EN UN SECTOR DE ESTUDIOS

NOTA 1.- Teniendo en cuenta que el objetivo de este anexo es explicar cómo realizar los cálculos y no el resultado en sí, es que presentamos las tablas reducidas. De esta forma ahorramos también un poco de espacio.

NOTA 2 - Dado que el Bario presenta en el caso analizado una distribución log-normal, el análisis del mismo se efectuó a partir del logaritmo de sus datos iniciales.

CORRELACION Log Ba: Pb

NN	log Ba	$(\log Ba)^2$	Pb	Pb^2	log Ba x Pb
1	2,0	4,0	0,3	0,09	0,6
2	2,3	5,29	0,6	0,36	1,38
3	2,6	6,76	1,0	1,0	2,6
• • •					
60	2,48	6,15	0,4	0,16	0,992
Total	147,36	367,32	75,6	211,32	206,06

Sustituyendo en (6)

$$r = \frac{206,06 - \frac{1}{60}(147,36)(75,6)}{\sqrt{\left[367,32 - \frac{1}{60}(147,36)^2\right]\left[211,36 - \frac{1}{60}(75,6)^2\right]}}$$

$r = 20,39 / 25,04$

$r = 0,81$

Sustituyendo en (5)

$$t(c) = \frac{0,81}{\sqrt{1 - 0,81^2}} \sqrt{60 - 2}$$

$t(c) = 10,62$

En tanto que t (95,58) = 2,00

Conclusión 1 - Dado que t (c) > t (p, n-2), se acepta que existe una correlación positiva y sensible entre ambos elementos. Por lo tanto el C.C. toma la siguiente forma parcial:

C.C. = Log Ba x Pb

NOTA 3 - De no disponer a mano de una tabla de Student (vid Anexo II) para determinar el valor de t (p, n-2), y siempre que n > 30, usted puede asumir que

Para P = 95,0 % t = 2
Para P = 99,9 % t = 3

CORRELACION Log Ba: Cu

NN	log Ba	$(\log Ba)^2$	Cu	Cu^2	log Ba x Cu
1	2,0	4,0	10	100	20
2	2,3	5,29	8	64	18,4
3	2,6	6,76	4	16	10,4
• • •					
60	2,48	6,15	2	4	9,92
Total	147,36	367,32	306	2.243,6	699,2

Sustituyendo en (6)

$$r = \frac{699,2 - \frac{1}{60}(147,36)(306)}{\sqrt{\left[367,32 - \frac{1}{60}(147,36)^2\right]\left[2243,6 - \frac{1}{60}(306)^2\right]}}$$

$$r = -52,34 / 60,75$$

$$r = -0,86 + ab^{1-4}$$

Sustituyendo en (5)

$$t(c) = \frac{-0,86}{\sqrt{1-(-0,86)}}\sqrt{60-2}$$

$$t(c) = 12,9$$

En tanto que t (95,58) = 2,00

Conclusión 2 - Dato que t (c) > t 9p, n-20, se acepta que existe una correlación positiva y sensible entre ambos elementos. Por lo tanto el C.C. toma la siguiente forma parcial:
C.C. - (Log Ba x Pb) / Cu

NOTA 4. - Queda ahora demostrar que entre el Pb y el Cu también existe una correlación negativa y sensilbe. En caso contrario, ambos elementos quedarían excluidos del C.C., quedando como único elemento informativo el Log Ba.l

CORRELACION Pb : Cu

NN	Pb	Pb^2	Cu	Cu^2	Pb x Cu
1	0,3	0,09	10,0	100,0	3,0
2	0,6	0,36	8,0	64,0	4,8
3	1,0	1,0	4,0	16,0	4,0
•					
•					
•					
60	0,4	0,16	2,0	4,0	0,8
Total	75,6	211,32	306,0	2.243,6	150,3

Sustituyendo en (6)

$$r = \frac{150,3 - \frac{1}{60}(75,6)(306)}{\sqrt{\left[211,32 - \frac{1}{60}(75,6)^2\right]\left[2.243,6 - \frac{1}{60}(306)^2\right]}}$$

$$r = -235,26 / 281,55$$

$$r = -0,84$$

Sustituyendo en (5)

$$t(c) = \frac{-0,84}{\sqrt{1-(-0,84)^2}}\sqrt{60-2}$$

$$t(c) = 11,79$$

En tanto que t (95,58) = 2,00

Conclusión 3.- Dado que t (c) > t (p, n-2), se acepta que existe una correlación negativa y sensible entre ambos elementos. Por lo tanto el C.C. toma definitivamente la siguiente forma:

C.C. (Log Ba x Pb) / Cu

Para las determinaciones aplicando el modelo V.O.C. se emplearán los resultados de dicho C.C.

Nota 5 - Todos estos cálculos se pueden hacer en muy poco tiempo empleando algún sistema de programas estadísticos. Personalmente recomiendo el MICROSTAT, por la potencia y variedad de sus cálculos y el efectivo sistema para el manejo de datos que posee.

ANEXO II. EXTRACTO DE LA TABLA DE DISTRIBUCION DE STUDENT, TOMADA DE KAZHADAN, A.B. et al. (1979)

N	P α	95 0,05	99 0,01	99,9 0,001
5		2,57	4,03	6,86
10		2,23	3,17	4,59
20		2,09	2,85	3,85
30		2,04	2,75	3,65
40		2,02	2,70	3,55
60		2,00	2,66	3,46
120		1,98	2,58	3,37
00		1,96	2,58	3,29

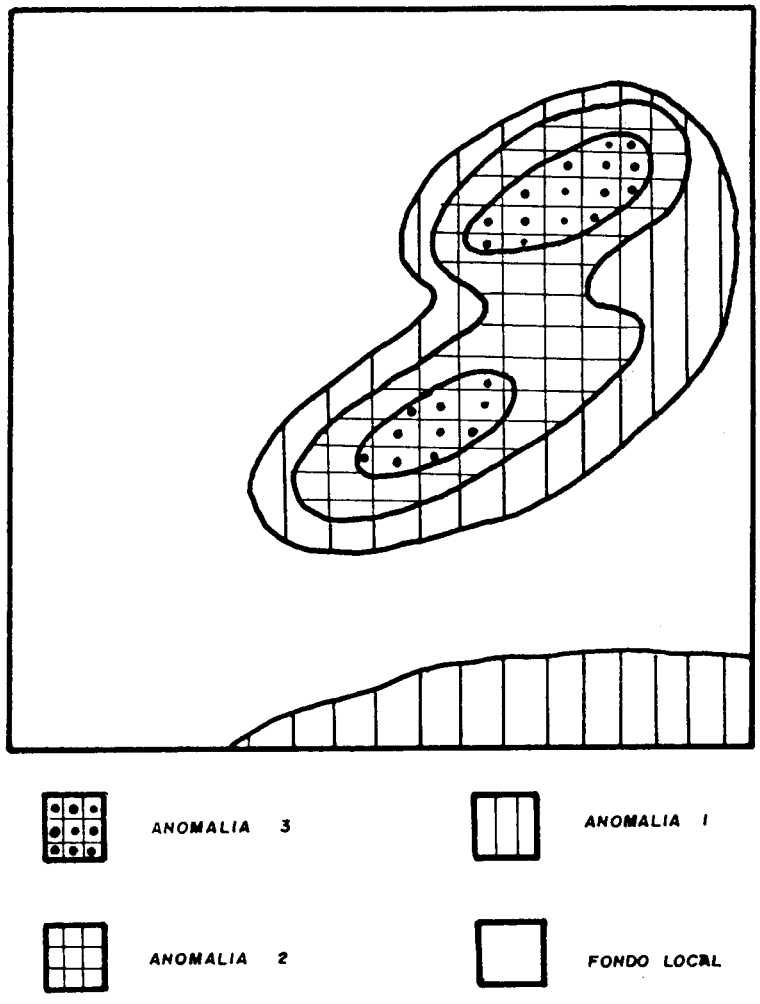

FIG. 1 MAPA DE LOS NIVELES DE ANOMALIAS DEL BARIO EN UN SECTOR DE ESTUDIOS

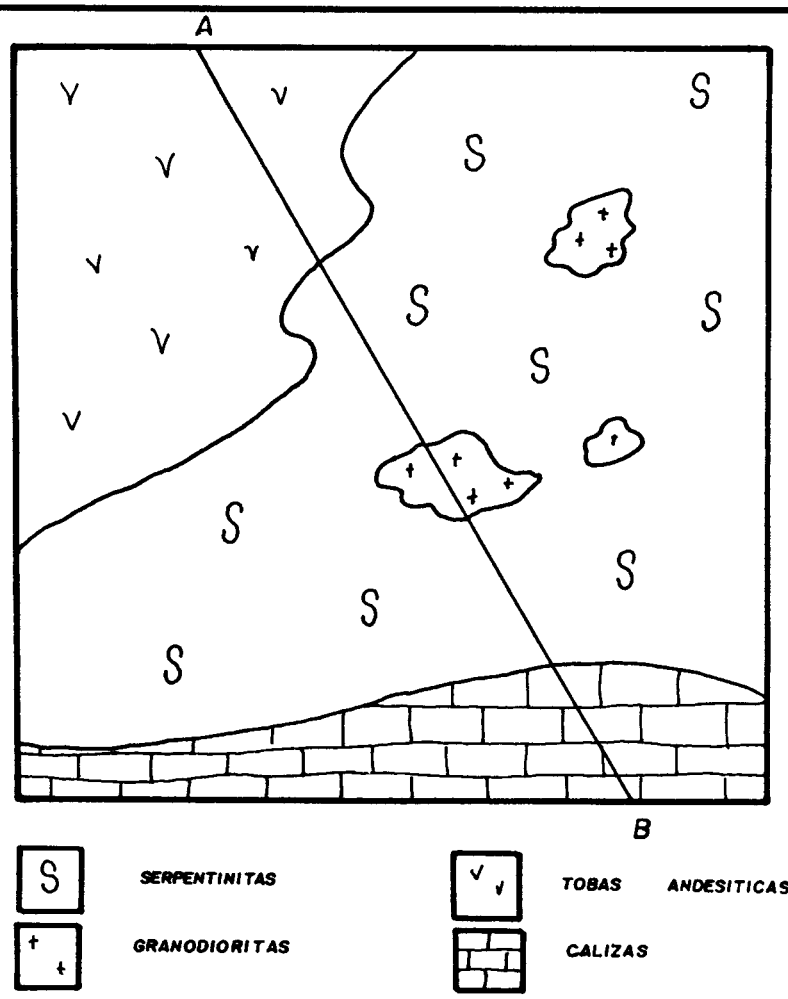

FIG. 2 ESQUEMA GEOLOGICO DEL SECTOR DE ESTUDIOS

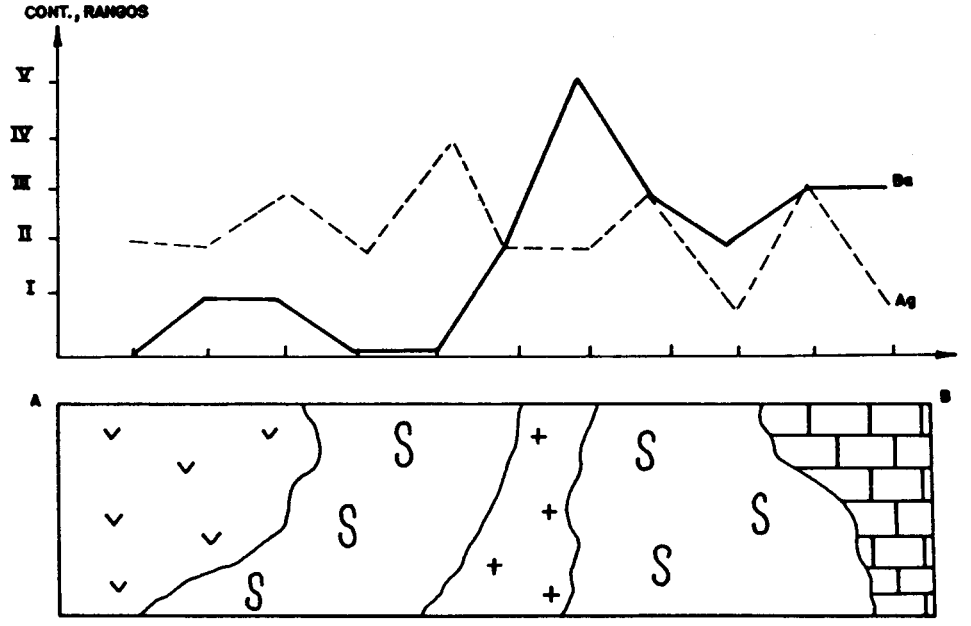

FIG. 3 SELECCION "GRAFICA" DE LOS ELEMENTOS INFORMATIVOS EN EL CORTE A-B.

BIBLIOGRAFIA

BONDARENKO, V.N. et al. Métodos estadísticos durante las búsquedas geoquímicas de los yacimientos minerales. Santiago de Cuba: 1985, pp. 103 - 108.

KAZHDAN, Alexei Borisievich et al. El modelaje matemático en la geología y la exploración de yacimientos minerales (en rusol. Moscú; Nedra, 1979, pp. 50 - 51.

LEAL OROPESA, Regino et al. Informe sobre la búsqueda orientativa de polimetálicos y oro San Marín (inédito). C.N.F.G., 1982, passim.

VALLS ALVAREZ, Ricardo Alberto. Los métodos geoquímicos en la búsqueda de talco en el macizo metamórfico del Escambray. Revista Tecnológica, Vol. XIX, No. 1, pp. 9 - 16, Enero - Marzo, 1989.

Modelaje geomatemático de la manifestación cuprífera "La Arena". Serie Geológica, No. 1, pp. 104 -118, 1987.

VOITKIEVICH, Georgii Vitoldovich. Pequeño prontuario geoquímico (en ruso). Moscú: Nedra, 1970, pp. 50-51.

YAMANE, Taro. Elementary Sampling Theory. La Habana: Pueblo y Educación, 1980, pp. 103 - 108.